Adaptive design of visual perception experiments

John D. O'Connor, Jonathon Hixson
US ARMY RDECOM CERDEC NVESD
Ft. Belvoir, VA 22060

James M. Thomas Jr.
E-OIR Technologies
Spotsylvania VA 22003

Matthew S. Peterson, Raja Parasuraman
George Mason University, Fairfax VA 22030

ABSTRACT

Meticulous experimental design may not always prevent confounds from affecting experimental data acquired during visual perception experiments. Although experimental controls reduce the potential effects of foreseen sources of interference, interaction, or noise, they are not always adequate for preventing the confounding effects of unforeseen forces. Visual perception experimentation is vulnerable to unforeseen confounds because of the nature of the associated cognitive processes involved in the decision task. Some confounds are beyond the control of experimentation, such as what a participant does immediately prior to experimental participation, or the participant's attitude or emotional state. Other confounds may occur through ignorance of practical control methods on the part of the experiment's designer. The authors conducted experiments related to experimental fatigue and initially achieved significant results that were, upon re-examination, attributable to a lack of adequate controls. Re-examination of the original results and the processes and events that led to them yielded a second experimental design with more experimental controls and significantly different results. The authors propose that designers of visual perception experiments can benefit from planning to use a test-fix-test or adaptive experimental design cycle, so that unforeseen confounds in the initial design can be remedied.

1. INTRODUCTION

The US Army RDECOM CERDEC Night Vision and Electronic Sensors Directorate's Modeling and Simulation Division (NVESD MSD) has conducted visual perception experiments since the 1970s. The goal of NVESD MSD's visual perception experiments is the improvement of human in the loop sensor performance models, such as Acquire and NVTherm IP, used in the sensor acquisition selection process and wargaming. These experiments involve classifying, recognizing, or identifying hundreds or thousands of image stimuli in rapid succession. In the mid 1990's, a standard visual perception test method for classification, recognition, and identification of military vehicles was developed. This standard methodology employed calibrated vehicle target

Infrared Imaging Systems: Design, Analysis, Modeling, and Testing XXI, edited by Gerald C. Holst, Keith A. Krapels,
Proc. of SPIE Vol. 7662, 76620H · © 2010 SPIE · CCC code: 0277-786X/10/$18 · doi: 10.1117/12.851032

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
2010		**00-00-2010 to 00-00-2010**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Adaptive design of visual perception experiments	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **US Army Research, Development and Engineering Command (RDECOM),Communications-Electronics Research, Development, & Engineering Center (CERDEC),Night Vision and Electronic Sensors Directorate (NVESD),Fort Belvoir,VA,22060**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Meticulous experimental design may not always prevent confounds from affecting experimental data acquired during visual perception experiments. Although experimental controls reduce the potential effects of foreseen sources of interference interaction, or noise, they are not always adequate for preventing the confounding effects of unforeseen forces. Visual perception experimentation is vulnerable to unforeseen confounds because of the nature of the associated cognitive processes involved in the decision task. Some confounds are beyond the control of experimentation, such as what a participant does immediately prior to experimental participation, or the participant?s attitude or emotional state. Other confounds may occur through ignorance of practical control methods on the part of the experiment?s designer. The authors conducted experiments related to experimental fatigue and initially achieved significant results that were, upon re-examination, attributable to a lack of adequate controls. Re-examination of the original results and the processes and events that led to them yielded a second experimental design with more experimental controls and significantly different results. The authors propose that designers of visual perception experiments can benefit from planning to use a test-fix-test or adaptive experimental design cycle, so that unforeseen confounds in the initial design can be remedied.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**Same as Report (SAR)**	**16**	

signature sets, calibrated displays, standardized test presentation software, controlled ambient lighting, participant target recognition and identification training and qualification, and many other experimental controls.

The goals of standardization were to reduce experimental variations due to human factors and to improve generalizability between experiments over time. To date, hundreds of experiments have been conducted using these methods, and the results of the standardized experiments are credited with vastly improving NVESD's sensor performance models. A more detailed description of this method follows in Section 2.

Participants in the standardized modeling tests often reported fatigue after several hundred images, but the effect of the perceived fatigue on human performance was not investigated in depth until 2008, when the authors conducted a first set of attentional fatigue experiments. The authors hypothesized that cognitive processes similar to those associated with vigilance decrement could reduce visual task attention performance over time, causing attentional fatigue, (Mackworth, 1948; Parasuraman, 1986). The first attentional fatigue experiment scored participants on their ability to classify and identify over 1,600 vehicle images in rapid succession without interruption. The participants' overall response times were also recorded, so that the effects of a perceived fatigue could be assessed for classification, identification and response time.

2. NVESD MSD STANDARD MODELING EXPERIMENTAL METHOD

2A. Target Sets

The foundations of current NVESD modeling experiments are standard *target sets*. Target sets are comprised of infrared and visible wavelength images of objects of military interest. For example, there are standardized vehicle, hand-held object, human activity, personnel, watercraft, and Improvised Explosive Device (IED) sets. The sets of objects have a measured overall discrimination difficulty known as the cycle criterion, (Johnson, 1958). The object images are selected such that the level of difficulty of discrimination between objects in a set varies significantly and approximates the level of difficulty (known as N_{50} for Johnson criteria and V50 for the Targeting Task Performance or TTP metric) sensor users are expected to face for a given military task (i.e., classification, recognition, or identification). Some pairs of objects will be very difficult to discriminate between, while others will be relatively easy. The standard NVESD MSD 12 target armored vehicle set (Figure 1) contains examples of easy discriminations (M113 vs. any other vehicle) and difficult ones, (T-72 vs. T-62 vs. T-55 and M109 vs. 2S3. It is important to note that the difficulty of vehicle discriminations can change with the wavelength being observed (O'Connor et. al, 2002).

At least 12 different objects viewed from at least eight different aspects are used to make up a target set. Four aspects will be cardinals where 0˚ = front, 90˚= left flank, 180˚= rear, and 270˚=right flank and the rest will be selected at oblique angles based on the object set's general aspect ratio. In the 12-target 8 aspect vehicle set, oblique aspects are collected such that the nearest cardinal aspect target surface areas are inversely proportionally represented in the oblique image aspects. For example, the degrees of a

right rear oblique for a vehicle with a 3:1 length to width ratio would be 180˚ + 90˚/ (3 + 1) = 202.5 Thus, ground vehicle images in the 12 target set are collected at aspects 0, 30, 90, 150, 180, 210, 270 and 330˚ (Figure 2), because each of these vehicles have an approximate length to width ratio of 2:1.

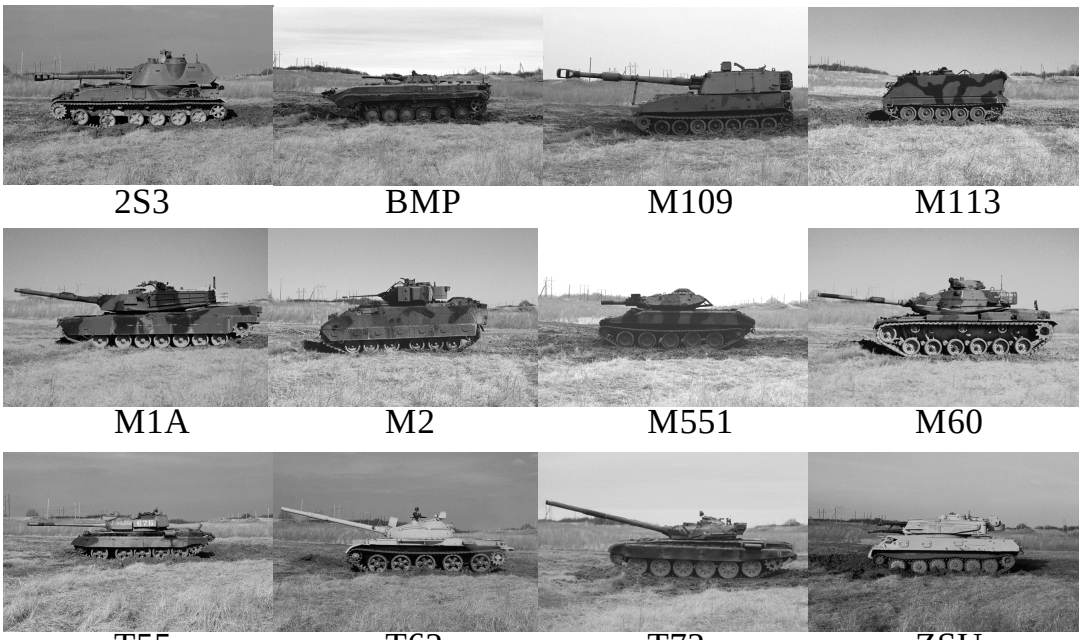

Figure 1. Standard 12 target vehicle set

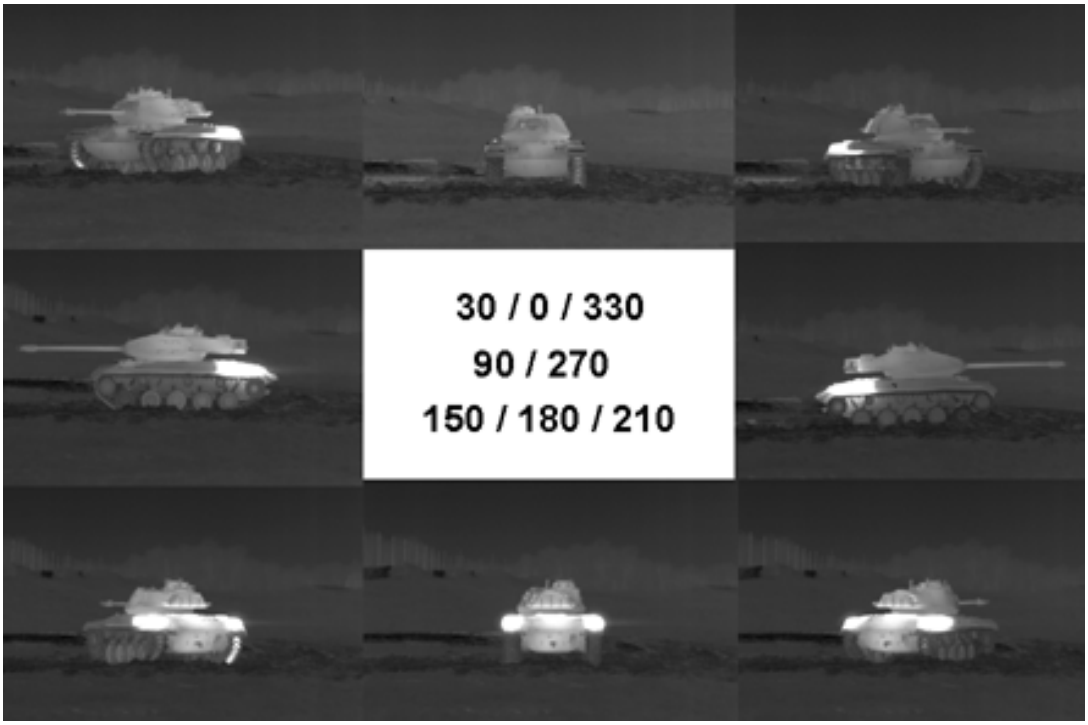

Figure 2. Standard ground vehicle aspects

Object images are collected according to a standard method. High resolution images of all objects are collected under the same conditions at the same location from the same very close range to maximize the number of object pixels within the image and to minimize the effects of atmospheric degradation. All images are radiometrically calibrated such that a temperature value may be assigned to each pixel.

Once collected, each object image must undergo a painstaking process of segmentation. Segmentation is conducted such that a target-only and silhouette mask (Figure 3) are created that define each pixel as either a background pixel or a target pixel. The key to segmentation accuracy and precision is to train personnel performing the segmentation task to be experts in identification of the specific objects to be segmented and then have them segment and re-segment each image, spending several hours on a single object image.

Once segmented, the squares of the average background temperature subtracted from the average target temperature value is added to target pixel variance. The square root of this sum yields the Root Sum of Squares (RSS) ΔT. The RSS ΔT is a standard measure of target to background temperature difference used, in combination with the target area to determine cycle criteria for target acquisition and sensor performance modeling. The NVTherm IP User's Manual defines RSS ΔT:

$$RSS\Delta T = \sqrt{\left(T_{Tgt} - T_{Bkg}\right)^2 + \sigma_{Tgt}{}^2}\,.$$

2b. the Perception Interface

The calibrated image set is then incorporated into the NVESD MSD perception test software interface. In some cases, sensor effects or some other treatment may be applied to the image stimuli. In other cases, the output of different sensors or sensor fields of view is the source of input stimuli for perception testing. In the latter case, the varying sensor outputs can be defined as varying image treatments.

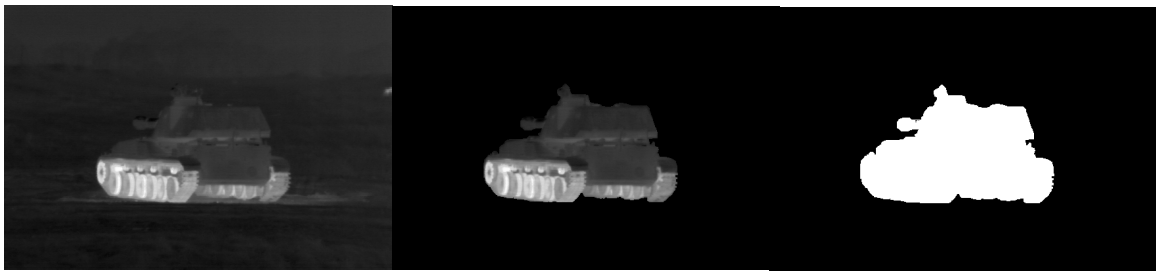

Figure 3. Segmentation stages: original, target only, and silhouette.

Through the perception test interface, the experimenter may ask participants to perform multiple target acquisition tasks including detection, recognition, classification and identification. All of these tasks may be performed through the use of forced-choice interfaces. For detection, the participant could mouse click on a perceived target or select a button for target present or target not present. For an 8 target ID test, the participant would choose the identity of the displayed object from 8 different buttons. Figure 4 illustrates representative search and identification display configurations.

Any number and combination of target acquisition tasks may be incorporated into the interface, such that a participant could be asked to detect, recognize, classify or identify an object within a given image or series of images. Thus, a participant could, for example, be asked to both classify and identify a given stimulus.

New target acquisition tasks may also be defined by the experimenter. Thus, if the experimenter wishes to define a type of recognition as tank vs. truck, or as tracked armored vs. wheeled armored vs. soft wheeled vehicle, he or she may do so.

Time limits may be placed on viewing and response times as desired by the experimenter. The time to view the image and perform the required task can be independent of each other. Time limits can range from 1/70 second to infinity. If the time to perform the task is longer than the viewing time, the image display window turns black at the end of the viewing period. For example, a participant may have 3 seconds to view an image and an infinite amount of time to choose a response.

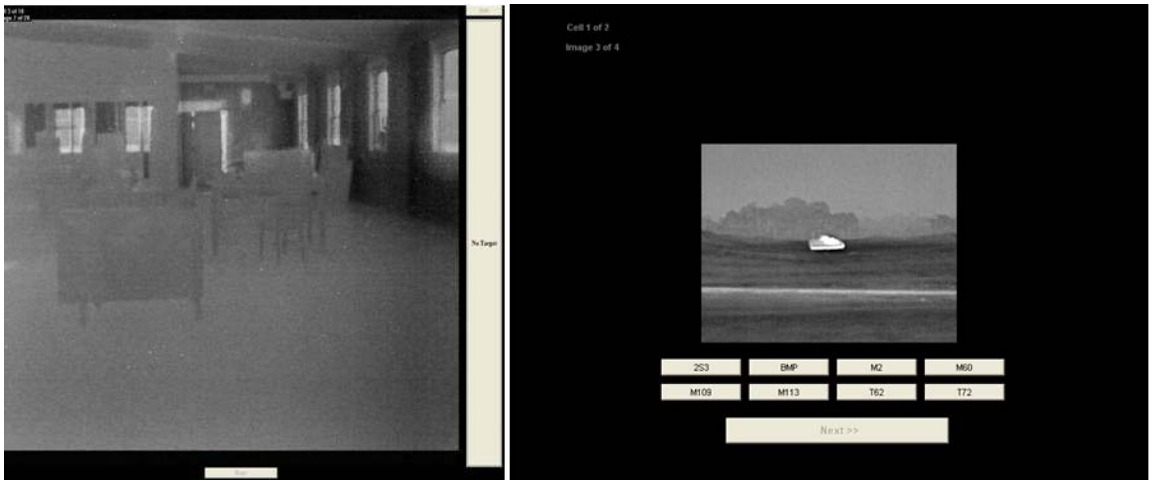

Figure 4. Search (left) and identification (right) graphic user interface examples.

Image stimuli may be presented in a specified order, a completely randomized order, or in image cells containing images with similar properties. Cells may be presented in a specific or random order, but images within cells are randomized. These differing methods of image stimulus presentation are used for specific tasks. For example, a short training pre-test may contain a specified order of image stimuli of increasing difficulty, allowing the trainee to gradually learn a specific skill. Search stimuli may be presented in random order, or in randomized cells of images when different sensor or image treatments are utilized. In the case of target recognition and identification, where different

types of sensor or image treatments are commonly used, each image cell contains images with the same treatment type, and the image stimuli within the cells are presented in random order.

2c. Calibrated perception laboratory

The presentation of calibrated stimuli requires the use of calibrated displays in a controlled environment. The NVESD MSD perception testing facility employs 10 participant testing stations, each with one photometrically calibrated 10-bit grey scale 2048x2560 BARCO display, and one photo metrically calibrated 20" color 1600 x 1200 LCD panel. The luminance of each display is standardized immediately prior to experimentation to ensure homogeneity. A black target will be adjusted to between 0.1 and 0.2 Cd/m^2 for either monitor. The ten bit monitors at 2048 gray shade (mid-level) and flat panel 8-bit displays at 128 grade shades are adjusted to a photometer reading of 20 Cd/m^2 (+/- 2).Additionally, each station is equipped with joy sticks, head phones and pen tablets to support specialized experiments.

The NVESD MSD perception lab was constructed such that light levels can be controlled from .01 Cd/m^2 to over 100 Cd/m^2. For most experiments, near total darkness is preferred, as it is desired that participants are dark-adapted. The dark-adaptation periods usually extend for at least fifteen minutes before experimentation begins.

Figure 5. Participant stations in the NVESD MSD perception facility.

The spatial arrangement and construction of the testing stations is such that the average viewing distance is 450mm, but may be varied between 250mm and 650mm. Head motion may be unconstrained or fixed through the use of chin rests. Partitions block participants' view of other participants' displays. Participant testing stations are arranged in two crescent-shaped rows facing a wall –mounted flat panel display used by the test administrator to provide examples and instructions, and to lead training sessions.

2d. Participant training and qualification

Participants typically attend NVESD MSD testing sessions over a five day period. The first one or two days are used for training participants to perform the types of tasks expected from them during experimentation. This most often involves training with the Recognition of Combat Vehicles (ROC-V) software to identify the thermal and visible signatures of at least 12 vehicles, objects or people. Participants may also be trained in the use of specific testing interfaces for tasks such as search and detection, recognition or classification. All participants take qualifying examinations and must demonstrate mastery for any given skill before participation in a given experiment. For example, participants trained in vehicle identification must be able to identify the trained set of vehicles 96% correctly before beginning a combat vehicle identification test.

Training participants in task proficiency is critical to reducing experimental error. Response variation from individual to individual is greatly reduced, and the variation in data is more likely to be due to the experimental treatments rather than differences between participants. Further, tasks such as thermal recognition and identification would be exceedingly difficult for untrained or inexperienced participants, such that their test data would be of questionable value.

2e. Testing

Participants begin testing upon completion of all training and relevant qualification tests. Participants must re-qualify daily and may re-train as necessary to reach desired task proficiency levels. Participants may not collaborate and are monitored by a test administrator. The test administrator answers questions, leads training and re-training sessions, administers qualifying tests and experiments, records participant progress and records the details of any notable or anomalous occurrences. Participants proceed at their own pace and generally complete between six and eight experiments per day. Regular breaks are enforced, and participants may take breaks as necessary. All relevant human use and experimental protocols and regulations are observed.

3. the Attentional Fatigue (ATTFAT 1) experimental methods

The first attentional fatigue experiment, or ATTFAT 1, was designed based on some, but not all, of the experimental standards of the NVESD MSD experimental method. The ATTFAT 1 experiment was a repeated measures design. Thirteen blocks were presented 7 times for a total of 91 blocks. Each block was a unique combination of 18 images, so that 1,638 stimuli were presented to the participant in one sitting. Each block contained images from the same infrared (IR) imager taken from a continuum of ranges, and thus representing a continuum of discrimination difficulty. Each cell was made up of images from 3 aspects of 6 vehicles (Figure 6), which is a significant departure from the NVESD MSD standard 12 target 8 aspect vehicle set.

Participants were tasked with recognizing each stimulus as a tank, armored personnel carrier (APC) or self-propelled artillery piece (SPA) and to further identify

each vehicle. All participants had been trained to a 96% competency for recognizing and identifying these vehicles. Participant responses were recorded and scored, and the total response time for both tasks was also recorded.

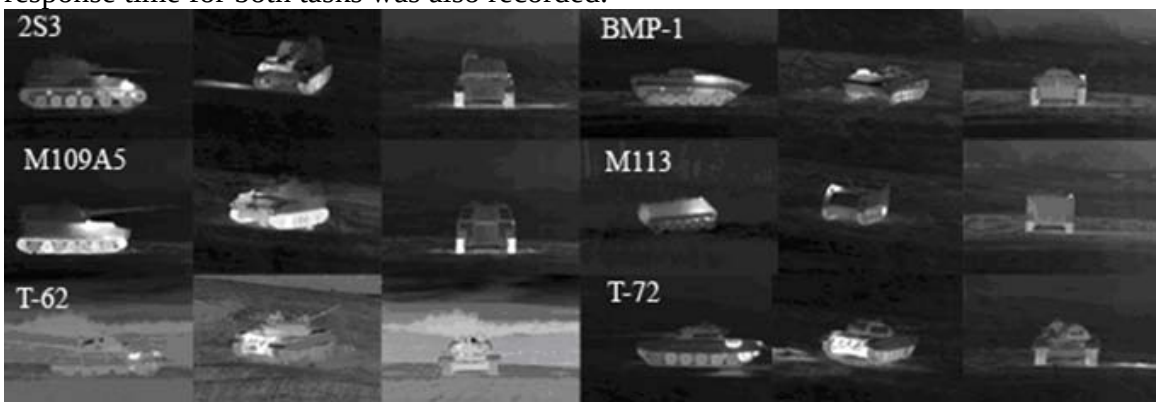

Figure 6. Six target three aspect target set

4. Attentional Fatigue (ATTFAT 1) Results

Initial regression analysis of 20 sets of participant data (Figure 7) for identification and recognition indicated a significant decrement to task performance over time might exist, ($F=15.7$, and $\alpha=.01$ for identification, $F=18.2$, and $\alpha=.01$ for recognition). Further examination of the individual data sets indicated that participants 9 and 10 were potential outliers. After the early stages of the experiment, these participants' scores approached chance. Examination of the data output for participants 9 and 10 revealed that later in the experiment they selected the same recognition class and vehicle identity for nearly all of their stimuli in all but the easiest cells (Figures 8 & 9). Further discussion with the lab administrator uncovered that these individuals reported feeling ill. Response times for participants 9 and 10 (Figures 12 and 13) indicate that, in the latter portions of the test, they were "clicking through" the stimuli as quickly as possible, biasing the experimental results. When the data for participants 9 and 10 was removed from the analysis (Figure 10), no significant performance decrement was observed ($F= 7.33$ and $\alpha=.796$ for identification and $F= 9.43$ and $\alpha=.466$ for recognition). Assuming the task performance of participants 9 and 10 was impaired such that they merited exclusion from analysis, no significant recognition or identification decrement was observed in the ATTFAT 1 experiment.

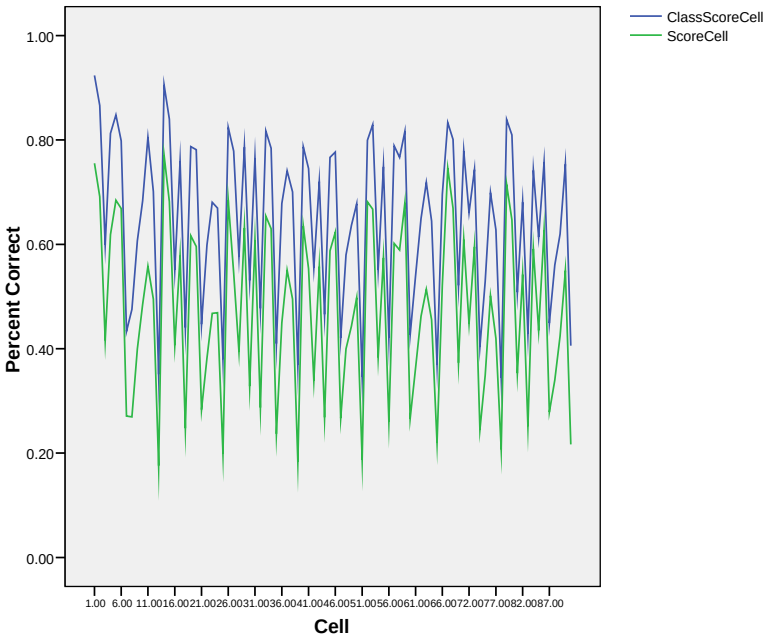

Figure 7. Mean classification (Blue) and ID (Green) scores versus cell number, n=20

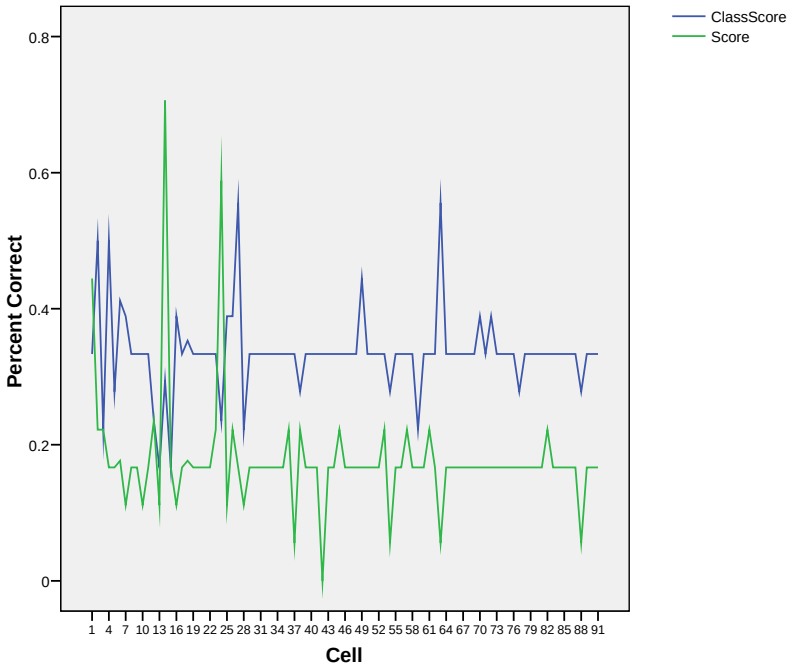

Figure 8. Participant 9 classification and ID scores

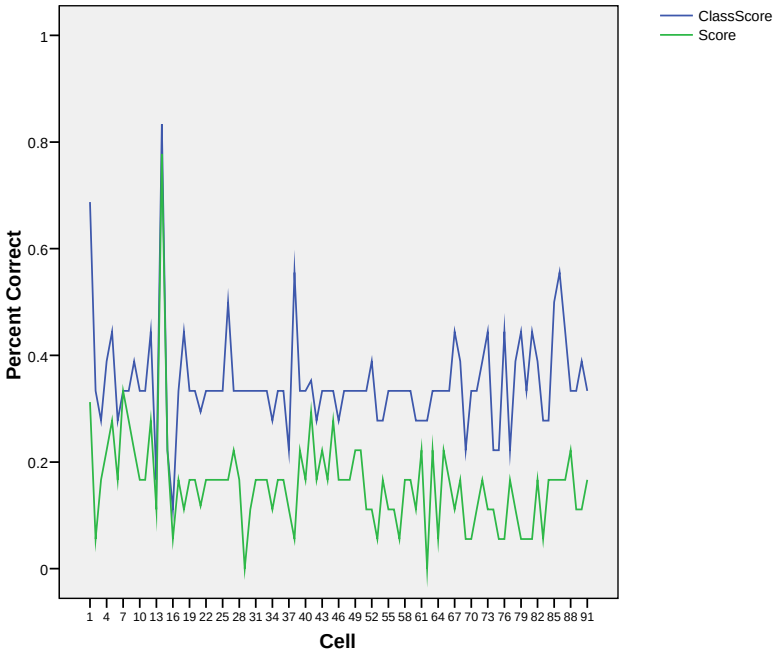

Figure 9. Participant 10 classification and ID scores

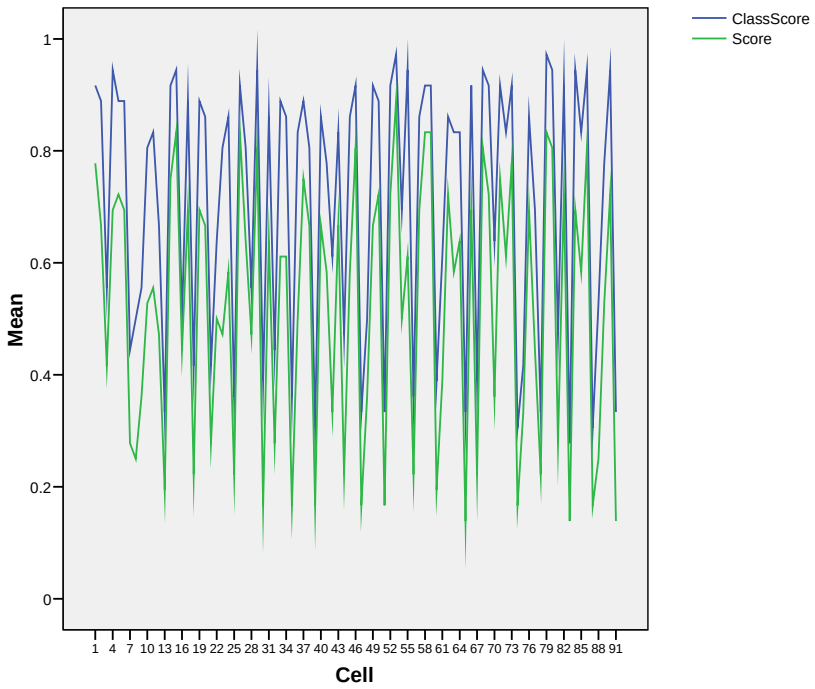

Figure 10. Mean classification and ID scores excluding participants 9 and 10, n=18

Another potential confound was noted, in addition to the questionable performance of participants 9 and 10. Many participants' performance on the first few cells appeared to improve when compared to equivalent cells later in the experiment. Some participants reported that they could "game" the interface- that the limited number of image-aspect combinations (18) enabled them to track which answers they had made within a cell, and this influenced later choices. It is also possible that a learning or automaticity effect occurred, affecting temporal performance.

Response time data was analyzed by ANOVA and linear regression, but without participants 9 and 10, as the authors believed there to be evidence to exclude them. Tables 1 and 2 indicate the results of this analysis. A significant but somewhat erratic decline in response time occurs throughout the course of the experiment (Figure 11). Analysis further indicates reaction times when identification is correct are faster than those when identification is incorrect (Figure 14). It is notable that in the first three cells (54 images), mean reaction time was significantly longer, and ID performance was lower than subsequent equivalent treatments (Figures 10 and 11). An automaticity effect could account for these differences, such that participants "warmed up" after the first few cells. Note that subjects had already undergone two days of training, so significant decreases in speed and increases in accuracy are unlikely to be due to learning effect.

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	6256.939	1	6256.939	415.508	.000[a]
	Residual	201272.1	13366	15.059		
	Total	207529.0	13367			

a. Predictors: (Constant), Cell

b. Dependent Variable: ResponseTime

Table 1. Response time vs. Cell ANOVA

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	3.996	.066		60.889	.000
	Cell	-.026	.001	-.174	-20.384	.000

a. Dependent Variable: ResponseTime

Table 2. linear regression of response time vs. cell

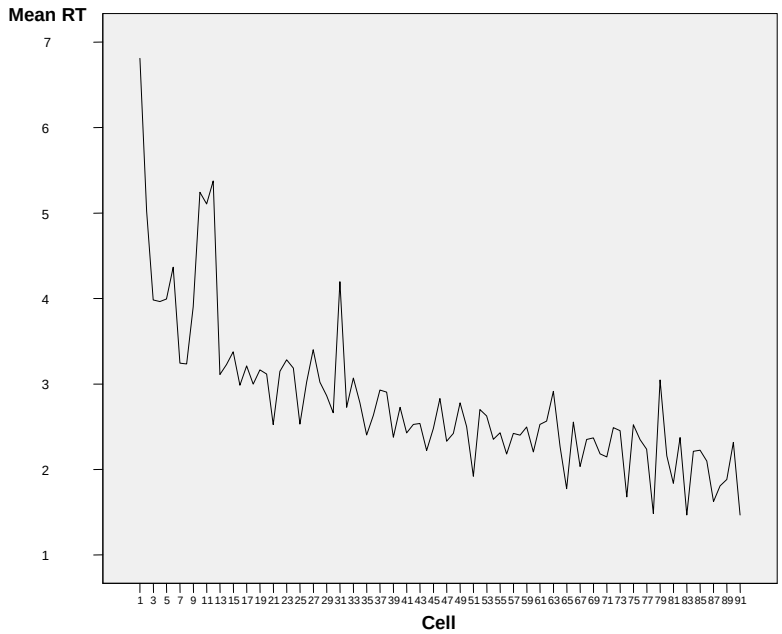

Figure 11. Mean participant population Response Time (RT) in seconds vs. cell

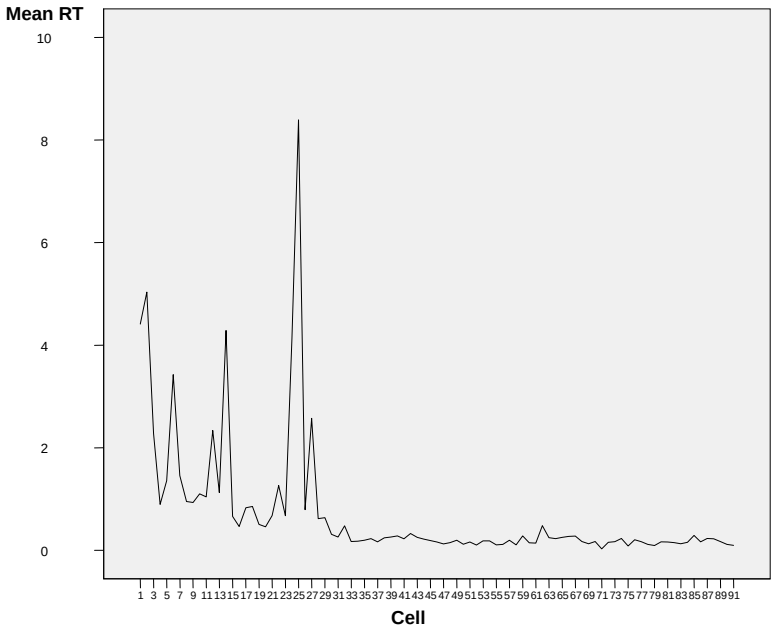

Figure 12. Mean RT in seconds for participant 9

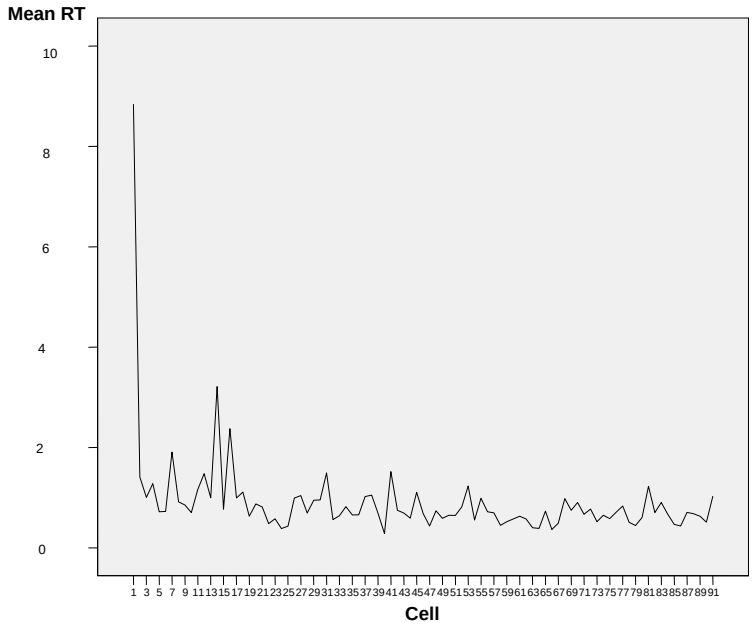

Figure 13. Mean RT in seconds for participant 10

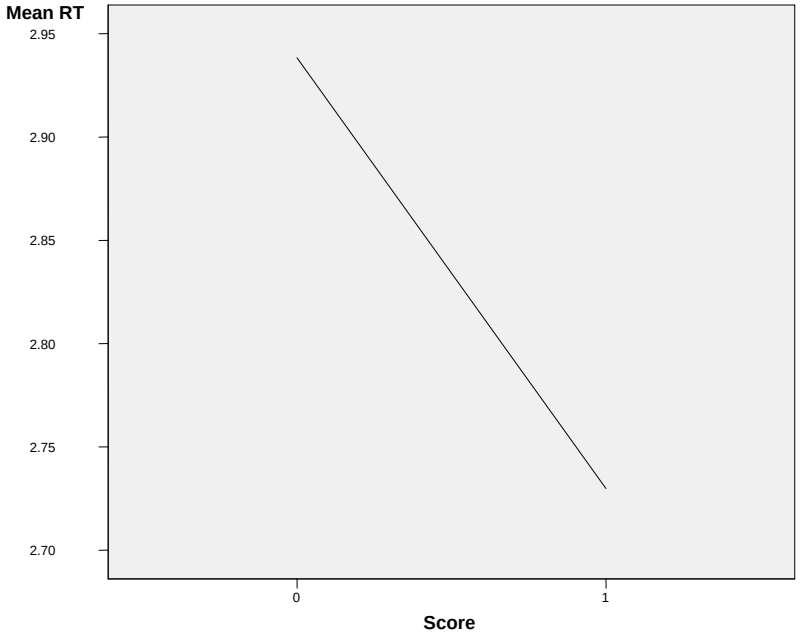

Figure 14. Mean response time vs. ID score (1 = correct, 0 = incorrect) for all observers

5. Conclusion

In the absence of confounding issues, ATTFAT 1 results do not indicate a significant decrement to human performance from the viewing of long sequences of image stimuli when the participants are asked to perform a target acquisition task for each stimulus. Potential shortcomings in the experimental design, however, bring into question any conclusions based on the recognition and identification task performed by these participants. Response times decreased significantly over the course of trials without a significant decrement to identification and recognition performance, indicating a potential automaticity effect that quickened response time.

6. Discussion

The lessons learned from ATTFAT 1 led to the design and implementation of a second attentional fatigue experiment, dubbed ATTFAT 2, which is ongoing at the time of this publication. The authors believed that changes in experimental design could yield more dependable results. The first design change was to increase the number of vehicles and aspects used as stimuli. For ATTFAT 1, 6 vehicles were viewed from 3 aspects, whereas ATTFAT 2 employed 12 vehicles from 8 aspects. This change was implemented to counter any potential learning or gaming effects.

The second change was the inclusion of “warm-up” and “cool down” cells. These cells were identical in composition (though randomized in image order) to other cells in the test, but are not analyzed. This change was implemented to counter any potential effects related to task initiation and completion that would affect participant performance.

The process of adaptive design, by which the initial participants are used to gauge the test design effectiveness, is believed to have yielded a better scientific instrument to measure the effects of extended exposure to test stimuli. While pre-testing with colleagues may expose some weaknesses, some confounds may not be evident until enough data is collected from the intended participant population to provide a clear picture. Feedback from participants was critical to the identification of the potential confounds described herein.

Meticulous design can control for known confounds, but in experimental cases where human participants are involved, some confounds may be difficult to foresee until after the subject population is sampled. It may therefore be beneficial for researchers to plan to employ an adaptive design process whereby time and resources are reserved to allow for multiple test design iterations.

7. References

Johnson, John (1958) Analysis of Image Forming Systems. *Proc. of Image Intensifier Symposium,* pp. 249-273

Mackworth, N.H. (1948) The breakdown of vigilance during prolonged visual search. *Quarterly Journal of Experimental Psychology, 1,* 5-61

O'Connor, John (2003) Fifty percent probability of identification comparison (N50) for targets in the visible and infrared spectral bands. *Optical Engineering, 42,* 3047

Parasuraman, Raja (1986) Vigilance, monitoring and search. In K. Bauff, L Kaufman and J. Thomas (Eds.), *Handbook of Perception and Human Performance, Vol. 2, Cognitive Processes and Performance(pp. 43.1, 43.39)* New York: Wiley