12TH ICCRTS
"Adapting C2 to the 21st Century"
Title:
"Challenges in Data Collection and Analysis in Multi-National Experimentation"
Topics:
Organizational Issues
Cognitive and Social Issues
C2 Metrics and Assessment
Network-Centric Experimentation and Applications
Author:

Jeff Duncan
Evidence Based Research, USJFCOM/JI&E
1500 Breezeport Way, Suite 400
Suffolk, VA 23435
757-203-3359
duncan@ebrinc.com
duncand@je.jfcom.mil

Abstract:

Military Warfighting Experimentation is an event used to learn whether a function, method, process, machine, etc will work or better stated to learn "how it will work," in a simulated environment in order to make educated determinations for real world operations. In order to make these educated determinations, analyst must collect applicable data and analyze it in a manner/method which answers the questions or hypotheses being investigated. Is the data being collected the appropriate data and does the analysis plan reflect the aims of the experiment? This question is applicable in any experimentation endeavor. Multi-national experimentation is no exception. Some of the same challenges that face multi-national experimentation face other types of experimentation while some are uniquely multi-national.

We plan to focus upon our insights from experiments MNE4 (Multi-National Experiment 4) and UR 2015 (Urban Resolve 2015) as our basis of exploration realizing that not all findings presented are uniquely multi-national. Realizing that rarely are two experiments are the same the purpose of this paper is not to create firm and fast rules for data collection and analysis in multi-national experimentation but to leverage findings for future experiments such that we do not "reinvent the wheel". This should help advance and improve the overall community's experimentation results and products.

Outline:

I. Introduction discussing MNE4 and UR2015
   a. Type of experiment
   b. General background
II. Aspects of Multi-national experimentation and how they differ from others (laying out the groundwork for some of the challenges)
   a. Language and culture
   b. Differing viewpoints
      i. Concepts
      ii. Priorities of experimentation
III. Differences between MNE4 and UR2015 (brief overview of differences)
   a. Embedding of analyst in cells
   b. Solution oriented versus concept oriented
   c. Many surveys versus few
IV. Challenges
   a. Sample size
      i. Very small in some cases
      ii. Representative of population?
   b. Surveys
      i. Converting the qualitative into quantitative
      ii. Frequency of surveys
      iii. Language
      iv. Timeliness of completion and delivery
      v. Social network
V. Conclusions and Future Research

Abstract:

Challenges in multinational experimentation exist in many forms with varying significance to the analyst depending upon the idiosyncrasies of a given experiment. Two such challenges surface with sample sizes and the use of surveys to collect experimental data. Several methods of confronting these challenges are available to the analyst. This paper, while not exhaustive, examines several methods for dealing with small sample sizes and explores some of the challenges associated with survey administration.

Background:

Military Warfighting Experimentation is ongoing and while being similar to other experimentation, contains added aspects not normally seen outside the military environment. Performing experimentation within one country's armed services can create conflict between a minimum of three to five different services with competing needs, priorities, and philosophies. In addition to the services, recent additions of government agencies into the experimental community have increased the number of personalities and aspects to the experiment. While this may seem to be overwhelming, next consider the addition of not just one country to the community, but several. Let's do the arithmetic. Assuming an average of 4 military services plus interagency units per country, if we have 10 countries involved in the experiment, we may be dealing with 50 different entities with varying social connections, differing priorities and capabilities, unique cultures, and as someone once stated, in some instances, countries separated by a common language.

Another aspect to multi-national experimentation, in addition to the cultural and competing priorities of countries and organizations, is data collection and the required methods to analyze the collected data. The parametric statistical model requires some basic assumptions. Among those assumptions are that the observations are independent

and that the observations are drawn from a normally distributed population. [10][11] At times the sample sizes can be significantly small which affects the ability to conduct valid parametric statistical analysis and the population from which the participants are chosen is not a random process. Another aspect is the use of surveys and interviews. While the sample sizes can be of issue with surveys, the cultural issues coupled with the English as a second language challenge can amplify the human factor effects on subjective and qualitative analysis.

The two experiments, Multi-National Experiment 4 (MNE4) and Urban Resolve 2015 (UR 2015) are the basis for much of the data and observations for this paper. Both experiments were distributed and involved coalition players, observers, analysts, and interagency participants. The differences were in the scenario; geographic, construct, and environment; and the physical locations of the analysts. MNE4's geographical location was the country of Afghanistan with 24 hour days being placed into 8 hours of experimentation per day, not being confined to a particular city or region with nearly all of the analysts being embedded with the experimental focus groups. The reasoning for this was to allow for the analyst to be able to observe the one-on-one conversations that did not occur over the IWS (Information Working Space) system or the distributed environment. UR 2015 was confined to the city of Baghdad, Iraq and experiment time was a minute to minute construct such that 10 days of 8 hour shifts daily resulted in 3-1/3 days of elapsed time. While the similarities and differences are not the main focus of this writing, the background is significant to potential differences in the addressed challenges and for potential future exploration.

Challenges:

Sample Size:

In order to accomplish a satisfactory statistical analysis, the sample size must be taken into consideration. Acceptable sample sizes for statistical analysis range from 15, 25, 30 or more, depending upon the source of reference. [1][2][3] Primarily for purposes of this discussion and simplicity, I am referring to survey results. During MNE4, sample sizes from surveys ranged from 1 to over 100 while UR 2015's sample sizes ranged from 4 to over 100.[4][5] When determining how to analyze the results, the analyst should treat the sample sizes differently to maintain analytic integrity? In addition, if the analyst is attempting to find a correlation between the players' backgrounds and the survey results, small sample sizes preclude use of ANOVA and other multi-variant tools from being utilized further complicating valid analysis. If a baseline is established prior to the experiment via LOE or other method, one can track the change or delta from the baseline. For example, if the process being evaluated is accomplished 3 times during the experiment, can we accurately say a statistical change has occurred? If we can say that a statistical change has occurred, has the sample size been large enough to validate? Herein lays a significant problem for the analyst.

Now is a good time to recall the Central Limit Theorem for Means: "For any population (with finite mean μ and standard deviation σ), the sampling distribution of the sample mean is approximately normal if the sample size n is sufficiently large." [2] What does "sufficiently large" indicate? "n" is the theoretical answer. The general rule of thumb from many statistics texts is that if $n > 30$, a normal approximation can be used. [2] [6] How does this affect statistical analysis of military experimentation when the

sample sizes are much less than 30? At sample sizes less than 30, it would appear that statistical methods such as linear regression, and ANOVA, will not prove useful or at least not provide valid results to the analyst. COBP for Experimentation states, “Most of the parametric statistics preferred for experimentation do not apply to sets of observations less than 30, though meaningful comparisons can be made between sets of 15, and non-parametric statistics can deal efficiently with as few as a handful of cases.” [1]

When dealing with these small sample sizes, in order to determine if nonparametric statistical tools are the methods of choice versus parametric statistical methods such as the t test, one should determine the answers to the following questions:

(1) *Do the data sets have a normal probability distribution?* See Figure 1:

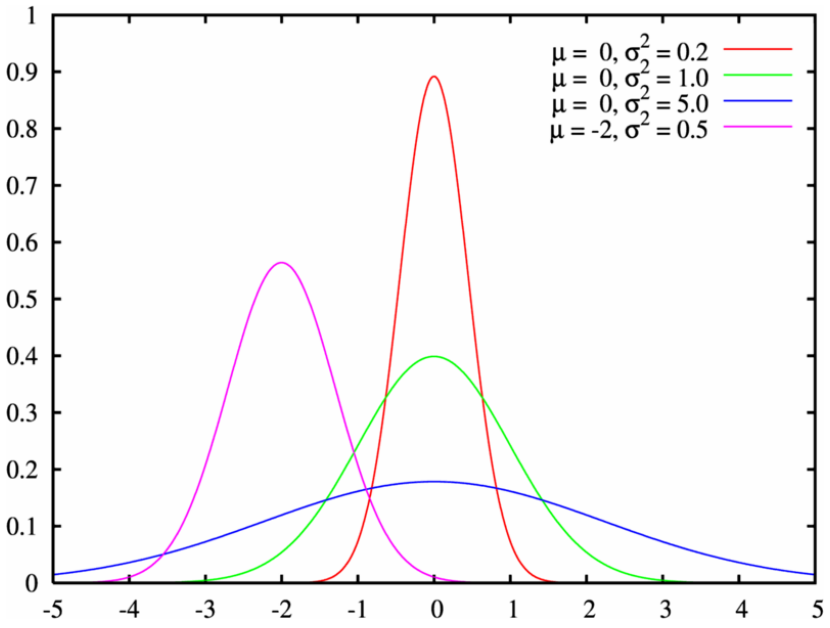

Figure 1 [12]

and

(2) *Is the data set not susceptible to measurement, but can be ranked in order of magnitude?*

If the answer to question (1) is “No,” or the answer to questions (2) is “Yes,” then the t test is not appropriate, thus nonparametric statistical tools may prove useful. [3]

EXAMPLE 1:

We have a sample size of 10 with the following sample distribution:

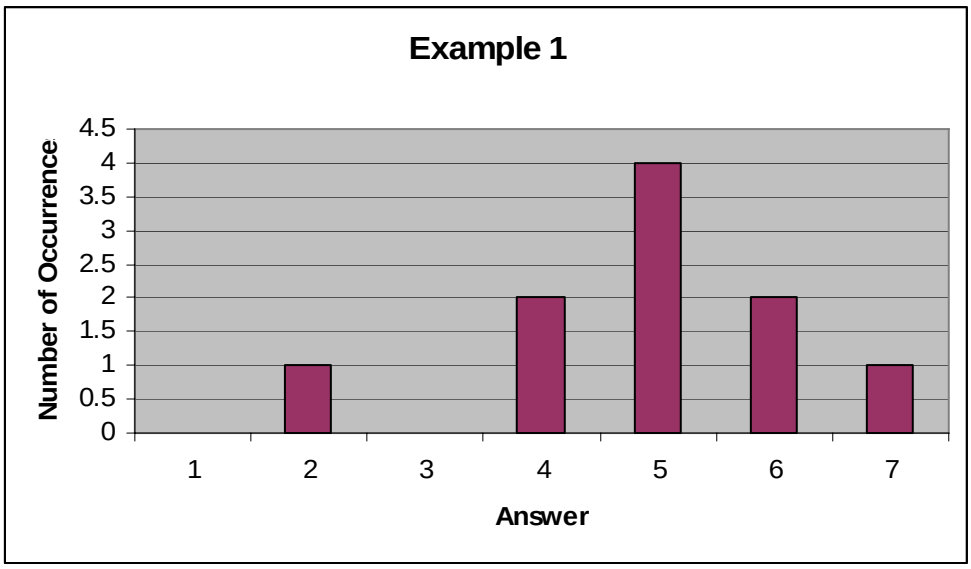

Figure 2

Example 1 has a sample size of less than 30. Thus we need to determine if a t test is applicable. The sample distribution seems to have a mound with two tails. Even though the distribution is not perfectly normal, it appears to be “normal enough,” thus a t test as well as other parametric statistical methods would be appropriate.

EXAMPLE 2:

Sample size = 10 with the following sample distribution:

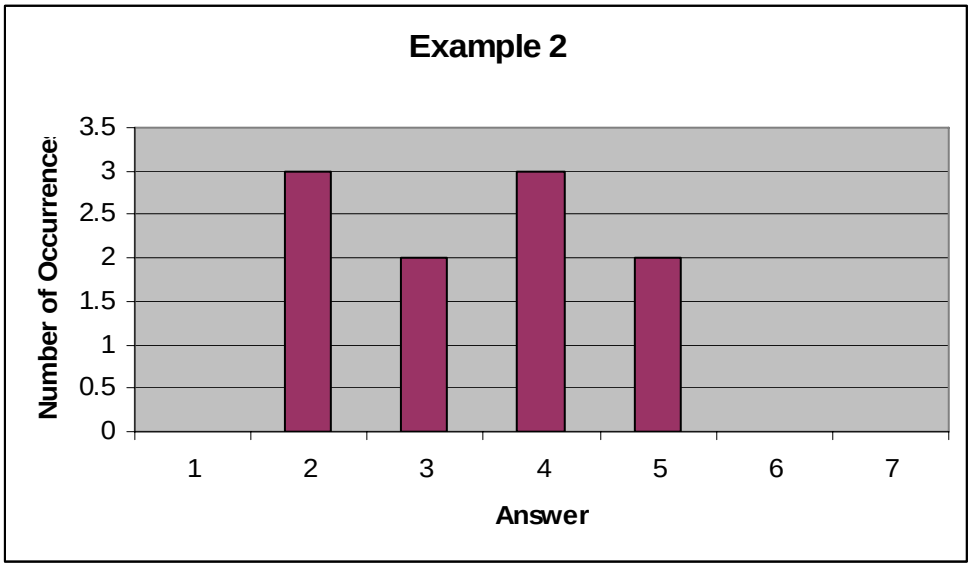

Figure 3

Example 2, just as in Example 1, has a sample size of less than 30. When we view the sample distribution, it is flat, no significant mound is present, thus the t test is inappropriate and we need to resort to nonparametric statistical methods.

MNE4 observed procedures repeated several times over the course of the experiment. In this case improvement in effort was noted and was expected to occur as the players became accustomed to the CONOPS. UR 2015 performed the same process during 3 different capabilities. While players were instructed to treat each iteration as if the previous iterations never occurred, it could be suspected that some of the improvement was due to educational aspects. Is the analyst trying to validate the learning curve or is he validating the process? In the case of MNE4, the change in performance was the focus while UR 2015 was attempting to evaluate solutions to the urban warfare problem under changing conditions. What variables changed and do we know all the variables that changed from one sampling to the subsequent samplings? During UR2015

the cognitive factor of the players had to be considered as a portion of the change in performance with the addition of the tools and concepts.

Solutions:

Vector Algebra:

Vector algebra coupled with cluster analysis is a method for coping with small sample sizes. Farrell, 2005, [13] details this method as used for MNE3 analysis. In addition, the Vector Method was used by Farrell to analyze data from MNE4. [14] Basically, this method treats each response as an element in a vector and then compares the resultant measured vector to a reference vector. The reference vector represents the highest value the test subjects could select. The resultant vector is the summation of the response vectors. A comparison is then made to determine the similarity of the two vectors. See Figure 4:

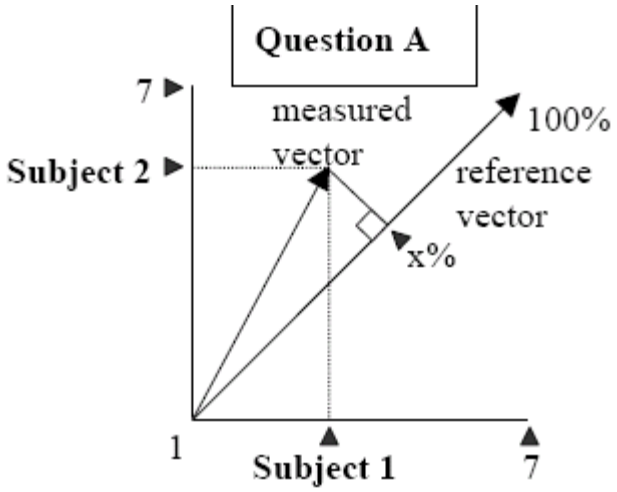

Figure 4 [14]

The analysis discussed in Farrell, 2006, [14] was used to determine if an improvement was seen for Common Intent over the course of the 3-week experiment. In addition to utilizing this method for small sample sizes, it can be used for large sample sizes and

helps to reduce a large data set to much fewer vectors, resultant vectors, aiding in a cleaner visual depiction.

<may need better example of Dr. Farrell's work>

Wilcoxon rank-sum Test:

Another method to analyze two sets of data samples is the Wilcoxon Rank Sum Test. This method is suggested for use when the sample size is relatively small and it cannot be determined if the sample sets are normal. The null hypothesis for this test is H_0: The two population probability distributions are identical. The sample sets, for this explanation, set 1 and set 2, are combined in numerical order and ranked. If the samples sizes are n_1 and n_2, respectively, with $n = n_1 + n_2$, the ranking will be from 1 to n. T_1 and T_2 represent the sum of the ranks for the two sample sets such that $T_1 + T_2 = n(n-1)/2$. Determination to reject the null hypothesis can be performed in two manners. One manner is to compare the T values with a Critical Values Table for Wilcoxon Rank Sum Tests or to compare the p-value's derived from statistical software such as SAS. [3] The following link provides a tool to calculate this statistic: [16]

http://www.fon.hum.uva.nl/Service/Statistics/Wilcoxon_Test.html

UR 2015 utilized this method to determine if a statistical improvement was realized with the addition of a C2 tool, JCPOF. The following example was used to determine if an improvement to the operational communication occurred after the inclusion of the JCPOF tool. Trial 1 was the baseline creation utilizing current capabilities. Trial 2 and 3 were testing of the JCPOF tool. A 7 point Likert scale [18] was used where 1 represented strong disagreement and 7 represented strong agreement

with relation to the ease of use of the tool. The following table represents the survey answers to the question of "Understanding:"

Understanding		
Trial 1	Trial 2	Trial 3
6	7	7
6	6	6
3	3	3
7	5	2
6	6	6
5	6	7
6	2	2

Table 1

The next tables represent the sorting process:

Trial #	**Score**	**Rank**
Trial 2	2	1
Trial 1	3	2.5
Trial 2	3	2.5
Trial 1	5	4.5
Trial 2	5	4.5
Trial 1	6	9
Trial 1	6	9
Trial 1	6	9
Trial 1	6	9
Trial 2	6	9
Trial 2	6	9
Trial 2	6	9
Trial 1	7	13.5
Trial 2	7	13.5

Table 2

Trial #	**Score**	**Rank**
Trial 3	2	1.5
Trial 3	2	1.5
Trial 1	3	3.5
Trial 3	3	3.5
Trial 1	5	5
Trial 1	6	8.5
Trial 1	6	8.5
Trial 1	6	8.5
Trial 1	6	8.5
Trial 3	6	8.5
Trial 3	6	8.5
Trial 1	7	13
Trial 3	7	13
Trial 3	7	13

Table 3

The next step is to add the sum of the ranks for each trial sample set:

Trial #	Score	Rank
Trial 1	3	2.5
Trial 1	5	4.5
Trial 1	6	9
Trial 1	6	9
Trial 1	6	9
Trial 1	6	9
Trial 1	7	13.5
TOTAL		56.5
Trial #	**Score**	**Rank**
Trial 2	2	1
Trial 2	3	2.5
Trial 2	5	4.5
Trial 2	6	9
Trial 2	6	9
Trial 2	6	9
Trial 2	7	13.5
TOTAL		48.5

Table 4

Trial #	Score	Rank
Trial 1	3	3.5
Trial 1	5	5
Trial 1	6	8.5
Trial 1	6	8.5
Trial 1	6	8.5
Trial 1	6	8.5
Trial 1	7	13
TOTAL		55.5
Trial #	**Score**	**Rank**
Trial 3	2	1.5
Trial 3	2	1.5
Trial 3	3	3.5
Trial 3	6	8.5
Trial 3	6	8.5
Trial 3	7	13
Trial 3	7	13
TOTAL		49.5

Table 5

Table 6 shows the use of a Wilcoxon table to determine is the differences in rank sum are statistically significant.

<insert graph>

<consideration given to adding a web tool and explanation>

Surveys:

General Discussion:

Surveys are a popular method of data collection because they provide data in an easy to view format [1] and allow for easy manipulation of the response data. In addition, they allow for ease in collection when working from a very large sampling of personnel when individual interviews would be labor intensive and time consuming. Ideally surveys contribute to the cognitive aspect of the experimental data for the data collection plan. In addition to gaining data concerning the cognitive aspect, surveys can be used when no other method exists to collect the needed data. A few of the questions that arise with surveys include: Where is the cut-off between utilizing the survey tool versus interviews? How frequent should the same survey be passed to the players? What is the maximum amount of survey questions a player can receive daily and how does an analyst treat survey results that were not completed at days end or other than the designated time for survey completion? Is the wording correct and understandable to the players for whom the language being used is not their first language? And even though the survey language is the player's first language, is it in the form he would expect?

MNE5 and UR 2015 were very different in regards to survey utilization. Below is a comparison of survey statistics between the two experiments:

UR 2015 Survey Summary

	Number of Unique Surveys	Number of Surveys Pushed	Number of Total Persons Surveyed
HITL 1	6	16	685
HITL 2	13	27	811
HITL 3	14	29	898

TOTAL	33	72	2394

Figure 1. UR2015 approximate survey data. [4]

UR 2015 had approximately 21 surveys for which the sample size was less than 10.

MNE 4 Survey Summary

	Number of Unique Surveys	Number of Surveys Pushed	Number of Total Persons Surveyed
TOTAL	88	141	14,400

Figure 2. MNE4 approximate survey data. [5]

MNE4 had approximately 25 surveys for which the sample size was less than 10.

As you can see, MNE5 was very survey intensive while UR 2015 use of surveys was moderate with respect to MNE5.

Challenges that can arise in relation to survey results include timeliness of the survey, workload of the participant, frequency of surveys, and promptness of responses from the participants.

This following includes the results of interviews with analysts from MNE4 and UR2015. Planning efforts for survey distribution for MNE4 attempted to combine timeliness of surveys without overloading the participants. Much work and effort was put into this initiative with mixed results. One can see from Figure 2, the survey load was significant and the workloads for surveys were taxing. The effect this had on participants was that some participants did not complete surveys until the following morning when experiment time resumed or that some participants answered questions without significant thought, thus survey results potentially were skewed. In focus areas where immediate survey completion was evident the player lead played a significant leadership role in this accomplishment.

Survey versus Interview:

In general, MNE5 did not use formal interviews but interviews were conducted on an ad hoc basis to gain further insights and clarification into actions taken by the players. UR2015 strived to use face-to-face interviews when the target audience was less than 10 and use interviews for larger target groups.

MNE5 operated in a distributed environment such that the possibility of face-to-face interviews was diminished. While analysts were embedded with the majority of the players in their focus groups, the distributed players could not be interviewed in the same fashion. The interview questions for UR2015 were similar to the survey questions such that they utilized a 7 point Likert scale. [18] The primary reasons for use of the interviews was to reduce survey loads, the small samples sizes required all participants' responses, and it was felt that much could be gained from the face-to-face experience such as the interviewee further eliciting his/her response versus a short written answer. The unexpected pitfalls to this theory arose when some of the questions presented during the same interview were similar and the interviewee's response was, "Same as the last answer." Some of the observers/interviewers were reluctant to press the player for further answers. This seemed to counteract the gains that were expected by use of interview versus electronic survey questions. The advantage of the electronic survey tool was that the player was forced to make a decision for each individual question and not take the "easy" route.

This confounds answering the question of which is better: electronic surveys or face-to-face interviews. The answer will most likely be different depending upon the type of experiment, characteristics of the players, nature of the questions, and training of

the observers/interviewers. All these issues should be addressed and considered when making this decision.

Social Networks:

To examine the effects of social networks on multi-national experimentation, it makes sense to review the four domains of warfare: physical, informational, cognitive, and social. Power to the Edge states, "C2 processes and the interactions between and among individuals and entities that fundamentally define organization and doctrine exist in the social domain." [7] It follows that military experimentation is going to encounter similar interactions as actual warfare will encounter although some of the encounters may be characterized differently. Actual warfare, in the social domain, will include relationships among the combatants, historically ingrained processes and practices, levels of trust among the combatants, potential cultural influences, and personal agendas. All these can exist in experimentation as well; however experimentation adds other aspects to the social domain. Many experiments are executed during daylight working hours which allows for "after-hours" conversations and interactions as well as extended cognitive dwelling time to assess the previous day's experimental scenario. Amplifying the effects of social networks is that the players are not normally selected on a random basis. Some of the players already have a social or working relationship and trust each other or understand the other's nuances.

The art of warfare is a 24/7 proposition in today's world. While the American Revolutionary War may have been fought predominantly during daylight hours, after hours maneuvers certainly existed such as Washington's crossing and the adventures of the *Turtle*. [8][9] MNE4 operated during daylight hours with overnight happenings being

divulged during the following day's morning brief. UR2015 spent 2 weeks performing 8 hour segments daily resulting in 3-1/3 days of elapsed time.

Both experiment scenarios have advantages as well as disadvantages. Which one is best applicable depends upon the goals and end states desired from the experiment. How does this affect survey/interview questions? Many surveys are designed to collect data regarding a specific occurrence or action thus timeliness is paramount. For instance, if a workload survey is given to players on a specific day and the survey is not completed until the close of experimentation the following day, the results could be contaminated. With a sample size of 100, only one set of data may not significantly affect the statistical findings, however in the case of a small sample size or numerous delayed responses, the statistical analysis could be flawed.

How can this be countered? Neither experiment had a statistical method to compensate for the "pub" factor and one analyst suggested that this aspect of multinational experimentation was a complete experiment in and of itself. Both experiments allotted time at the conclusion of each experimental day for survey completion. Some participants were conscientious and dutifully completed their surveys while others were not. MNE4 players were permitted to complete surveys the following morning while UR2015 removed the surveys from possible completion prior to the following day's resumption of experiment play. One analyst from MNE4 indicated that his timely completion rate was significantly high due to the focus area leader's persistence with his fellow players.

Conclusions and Future Research:

Multinational experimentation has many challenges and no two experiments experience the challenges in the same manner. Small sample sizes can be analyzed with both parametric and non-parametric methods depending upon the distribution of the data. It should be a given that the analysis plan is developed in conjunction with the data collection plan, thus the analytical methods, while not needing to be completely determined, must be thoroughly considered with flexibility as part of the plan as the analyst will not know the exact distribution of the data prior to the experiment.

Surveys can serve an important purpose in experimentation but I suggest use of surveys be judicious. Overburdening the player can potentially result in contaminated data if the survey taker simply offers random answers in order to complete the survey. Also, positive leadership and leading by example in the focus groups can increase timely completion of surveys. Consideration should be given to removing surveys from the queue upon completion prior to the following days experimentation play. In addition to the management aspects of multinational experimentation, surveys must be worded properly such that the player has no questions as to what is being asked. In addition, the analyst must know exactly what he/she is attempting to discover from the survey question. These two aspects are necessary or else time and energy has been wasted. It is also suggested that a subject matter expert review the questions and that questions be sent to only those who can intelligently answer the questions.

No two experiments are the same, thus the answers to the topics above do not have firm answers. Analyst need to take into consideration several aspects of the experiment such as the environment – distributed or not distributed, sample size of the

data set, target audience of the survey/interview, can the data be collected by observation, and many others.

Future multinational experimentation research can be considered in the area of "bootstrapping" or re-samplings of large samples sizes. Many surveys are sent to the entire player audience and thus encompass the entire population, however this population is of the experiment players and the players are not normally randomly chosen to participate in the experiment. Most likely the players were chosen due to their expertise in their given field and do not represent the population from their country's military population.

Acknowledgements:

The following contributed significantly by conveying their expertise in multinational experimentation analysis, data collection, and survey management: Dr. Brooke B. Schaab, Dr. Elizabeth Bowman, Dr. Phillip Farrell, Christine H. Mills, Gabriel Rouquie, Mike Wahl, and Charles T. Wall.

I wish to extend my appreciation to Dr. Michael Cochrane for his guidance and direction in the use of statistical methods.

References:

[1] Alberts, David S., Hayes, Richard E., Code of Best Practice for Experimentation, July 2002, reprint 2003.

[2] Hildebrand, David K., Ott, R. Lyman, Statistical Thinking for Managers, Duxbury, 1998.

[3] Mendenhall, William, Sincich, Terry, Statistics for Engineering and the Sciences, Prentice Hall, fourth edition, 1995.

[4] UR 2015 Survey data collection file, USJFCOM/JI&E, Suffolk, VA.

[5] MNE4 Initial Impressions 18 Apr 06 v6.ppt.

[6] Kiemele, Mark J., Schmidt, Stephen R., Berdine, Ronald J., Basic Statistics – Tools for Continuous Improvement, Air Academy Press, fourth edition, 2000.

[7] Alberts, David S., Hayes, Richard E., Power to the Edge: command and Control in the Information Age, June 2003, reprint June 2004.

[8] Gidwitz, Tom, © 2005 by the Archaeological Institute of America www.archaeology.org/0505/abstracts/warsub.html

[9] http://en.wikipedia.org/wiki/Washington's_crossing_of_the_Delaware

[10] Gardner, Paul L, Scales and Statistics, Review of Education Research, Vol. 45, No. 1 (Winter, 1975), pp. 43-57

[11] Siegel, Sidney, Nonparametric Statistics, The American Statistician, Vol. 11, No. 3. (June, 1957), pp. 13-19

[12] http://en.wikipedia.org/wiki/Image:Normal_distribution_pdf.png

[13] Farrell, Philip S. E., Calculating Effectiveness with Bi-Polar Scales and Vector Algebra, Defence R&D Canada – Toronto, June 2005.

[14] Farrell, Philip S. E., Common Intent and Information Processing Frameworks applied to Effects Based Approaches to Operations, ICCRTS, September 2006.

[15] UR2015 Analytical Report, Appendix C, JCPOF Analytical Report, USJFCOM/JI&E, 2006.

[16] http://www.fon.hum.uva.nl/Service/Statistics/Wilcoxon_Test.html

[17] http://fsweb.berry.edu/academic/education/vbissonnette/tables/wilcox_r.pdf

[18] http://en.wikipedia.org/wiki/Likert_scale