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
Authors:

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

Dr. Philip S. E. Farrell
Defence R&D Canada – Ottawa
Canadian Forces Experimentation Centre
National Defence Headquarters
Ottawa ON K1A 0K2
613-990-6732
farrell.pse@forces.gc.ca
philip.farrell@drdc-rddc.gc.ca

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 appropriate data being collected and does the analysis plan reflect the aims of the experiment? These questions are 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 no two experiments are rarely 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 terms of philosophies, participants in warfighting experimentation may believe they are part of a warfighting exercise that has the expectation of training, while the experimentation designers and analysts are not concerned about testing the concept – not the person. These competing expectations often get confused and they impact how survey questions are designed and answered. In addition to the services, recent military experimentation has included civilian government agencies, which 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 3 military services plus 2 government agencies 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 is sometimes voiced, 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 or volunteer 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 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 scenario was set in the country of Afghanistan with 24-hour days being placed into 8 hours of experimentation per day, however the experimental participants and analysts were geographically spread across six European and North American laboratories. 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 as part of an ongoing operation 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 between the two experiments are not the main focus of this writing, the background is significant to potential differences in the addressed challenges and for potential future warfighting experimentation.

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] Survey results are referred to primarily for purposes of this discussion and simplicity, . 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. Further complicating the analysis is that a sample size of 100 does not necessarily alleviate the sample size challenge. In many cases, even though the same questions are addressed to each participant, the different groups may have very different tasks to perform (while within the group their tasks are similar), thus an analysis of the population may not be worthwhile but a comparison within specific groups and between groups would be beneficial. The sample

size in this situation depends on the size of the group, which was a small as four people in some cases.

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 lies 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 less than 30? At sample sizes less than 30, it would appear that statistical methods such as linear regression, and ANOVA, may not be as useful depending upon the distribution of the data. 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, one needs to determine if nonparametric statistical tools are the methods of choice versus parametric statistical methods such as the t-test, should be used for their analysis. A brief explanation of the

difference between parametric and nonparametric is worthwhile at this point. Parametric tests are based on the premise that the data come from a probabilistic distribution, while non parametric methods are referred to as “distributionfree” tests, thus a probability distribution is not considered. [3] When deciding whether one should utilize parametric or non parametric methods, they should determine the answers to the following questions:

(1) *Do the data sets have \probability distributions?* See Figure 1:

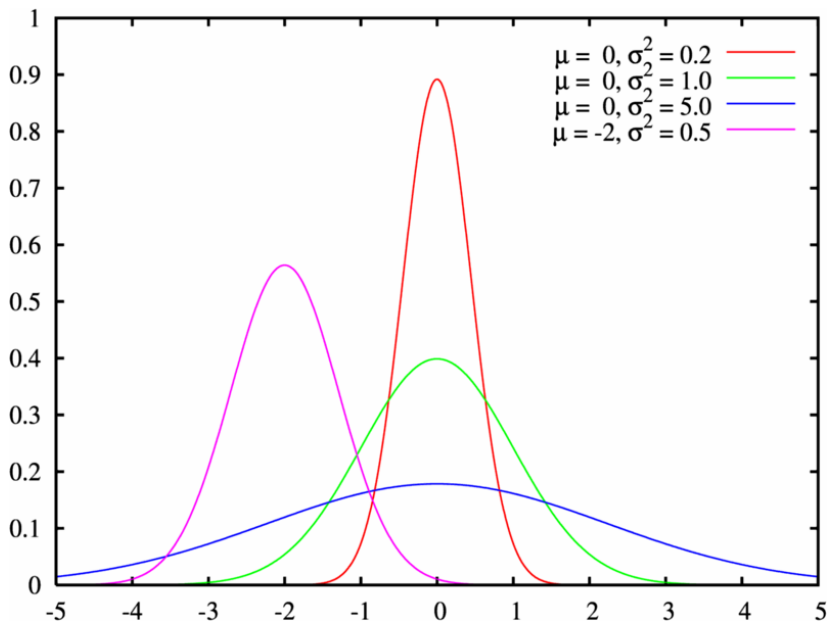

Figure 1 [12]

and

(2) *Can the data set be ranked in order of magnitude?*

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

EXAMPLE 1:

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

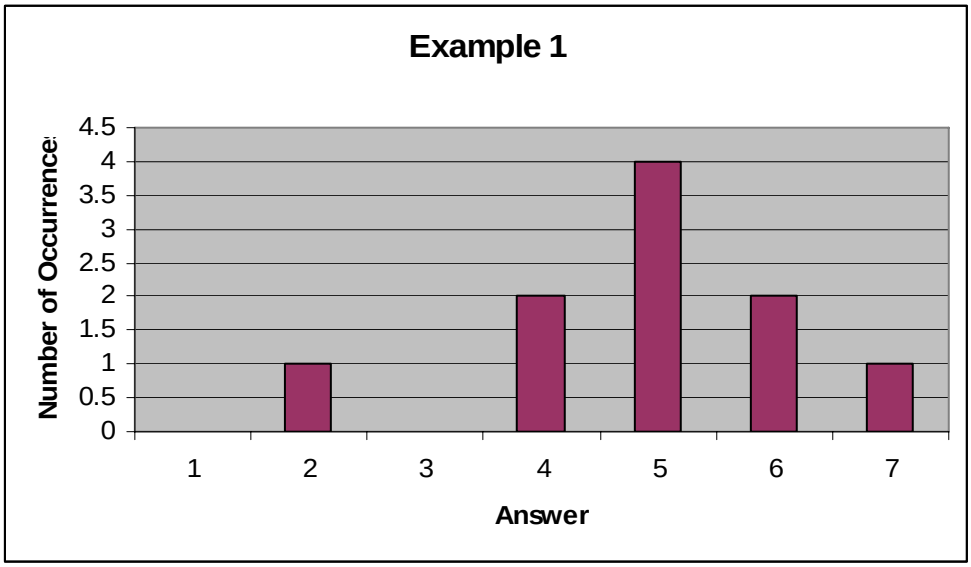

Figure 2

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:

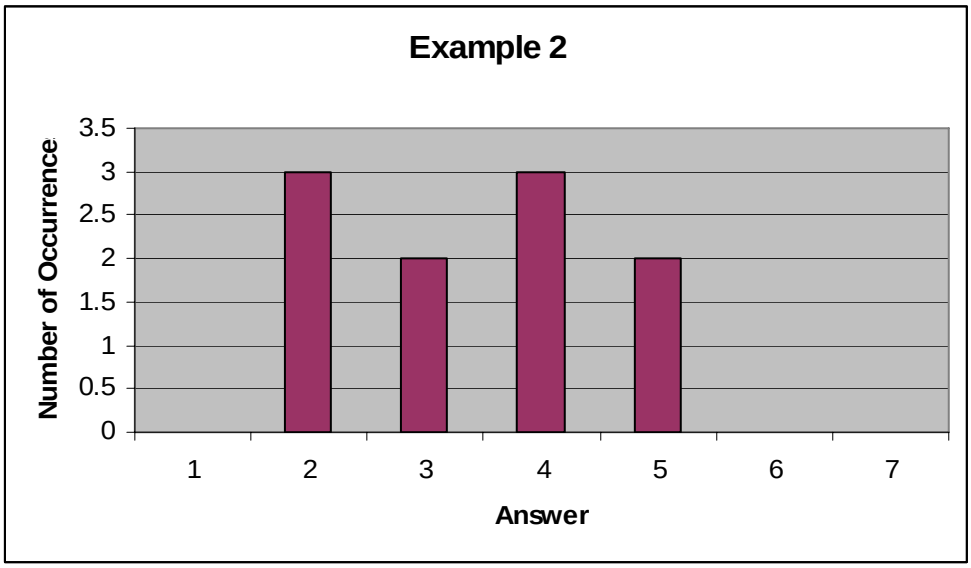

Figure 3

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 may not be the best statistical method for analysis and we should utilize 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 learning effects. 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.

Possible Solutions For Dealing with Small Sample Sizes:

Vector Method:

The Vector Method can be applied to any sample size, large or small. This method treats each response as an element in a vector (rather than datum from a distribution) and then compares the resultant measured vector to a reference vector. Expressing the data as a vector means that the method is scalable and can cope with data

sets of any sample size. This method is explained and applied to MNE3 and MNE 4 Common Intent analyses and other complex experiment analyses. [13] [14]

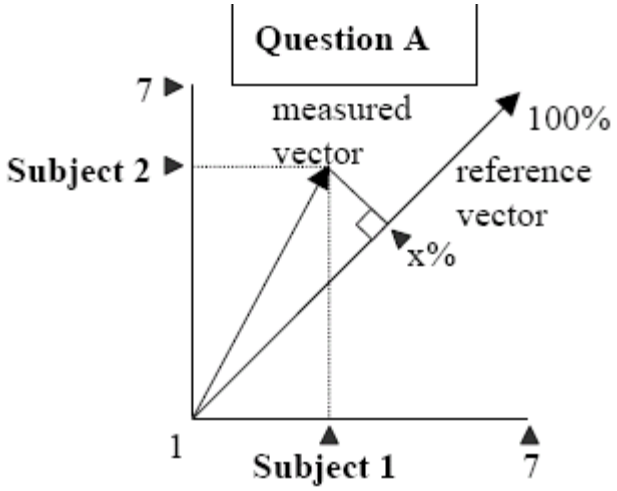

Figure 4 [14]

Figure 4 shows the response to Question A from two Subjects: Subject 1 provides a response of 3.5 and Subject 2 responds with 5. The responses are combined using vector algebra. The measured vector is compared to a reference vector that the analyst chooses. In this case, the reference vector represents the highest value the subjects may select. The measured vector's magnitude is 56% of the reference vector magnitude (8.5 units) and has an angle of 13° from the reference vector. The magnitudes of both vectors as well as the angle between the two vectors can be compared to provide a measure of similarity. The Euclidean product (the projection of the measured vector onto the reference vector) is a calculation that combines both magnitude and angle in order to provide a measure of similarity (54% of the reference vector magnitude, in this case).

Common Intent was tracked over the course of MNE 4's 3-week experiment for the operational level headquarters [14]. The Coalition Task Force (CTF) headquarters had

seven staffs: Command Group (CG), Effects-Based Planning (EBP), Effects-Based Execution (EBE), Effects-Based Assessment (EBA), Knowledge Support, which was further divided into the Knowledge Based Development (KBD) and Knowledge Management (KM) staffs, and finally the Multinational Interagency Group (MNIG). The analysis challenge was that, on any particular day, there was no guarantee that the number of people in a staff would be the same or that there would be the same people answering the survey question. Another challenge was that each of the staffs were different in size. The Vector Method makes it possible to compare data sets of different sample sizes.

To illustrate the Vector Method, Table 1 presents the results for only 1 out of 21 questions on Common Intent - *To what extent do you believe CG's actions are consistent with the Commander's intent? (Not at all) 1 2 3 4 5 6 7 (Completely).* Note that the sample size ranges from 4 to 34. Projections of measured vectors (one for each staff per day) onto their respective reference vectors are calculated using the Euclidean Product and expressed as a percentage of the reference vector magnitude. In this form, the analyst may compare the projection percentages across days for a single staff, and across staffs for a single day.

Table 1. Projection of measured vector onto reference vector for different sample sizes

		CG	EBP	EBE	EBA	KBD	KM	MNIG
28 Feb	**Sample**	**4**	**12**	**19**	**32**	**22**	**18**	**9**
	Projection (%)	88	67	58	64	67	57	56
2 Mar	**Sample**	**4**	**12**	**22**	**34**	**21**	**18**	**10**
	Projection (%)	88	69	61	60	65	53	53
7 Mar	**Sample**	**4**	**11**	**21**	**34**	**21**	**18**	**11**
	Projection (%)	96	67	70	70	66	59	65
9 Mar	**Sample**	**4**	**12**	**21**	**33**	**22**	**18**	**11**

	Projection (%)	96	69	74	69	58	66	64
14 Mar	**Sample**	**4**	**10**	**20**	**34**	**21**	**18**	**8**
	Projection (%)	96	78	77	75	66	64	60
16 Mar	**Sample**	**4**	**10**	**20**	**34**	**20**	**18**	**7**
	Projection (%)	96	80	78	72	68	69	69

The measured vector of one staff on a particular day can be compared directly to the measured vector of the same staff on another day only if the two vectors have the same number of elements so that the Euclidean Product can be calculated. This means that the same subjects must have answered the question on both days, and the order of responses appearing as elements in each vector must be preserved (recall from Figure 4 that the response for a specific subject is a vector element, which represents a value on an orthogonal dimension within its vector space). For example, the CG staff had the same 4 respondents, and therefore, their responses form 4-dimensional vectors. Table 2 lists these vectors, where the first element represents the response from Subject 1, and the second element from Subject 2, etc.

Table 2. CG measured vectors

Date	Name	Vector
28 Feb	v1	(7, 5, 6, 7)
2 Mar	v2	(7, 6, 5, 7)
7 Mar	v3	(7, 6, 7, 7)
9 Mar	v4	(7, 7, 7, 6)
14 Mar	v5	(7, 6, 7, 7)
16 Mar	v6	(7, 7, 7, 6)
Reference Vector	r	(7, 7, 7, 7)

Although the following magnitude and angle are calculated for all 4 subjects, Figure 5 plots the vectors for Subjects 2, 3, and 4 only (N.B. subject 1 answered “7” for all days and was omitted from the drawing). Figure 5 shows five vectors. However, v3 and v5 are coincident, as well as v4 and v6. The Reference Vector (r) is shown as a dashed arrow, and the measured vectors cluster around r. Although v1 and v2 have identical

elements (and their magnitudes are the same) they appear in a different order, and therefore the vectors pointing in different directions. Thus, magnitude alone is not sufficient to determine if two vectors are similar since two vectors may have the same magnitude but point in opposite directions.

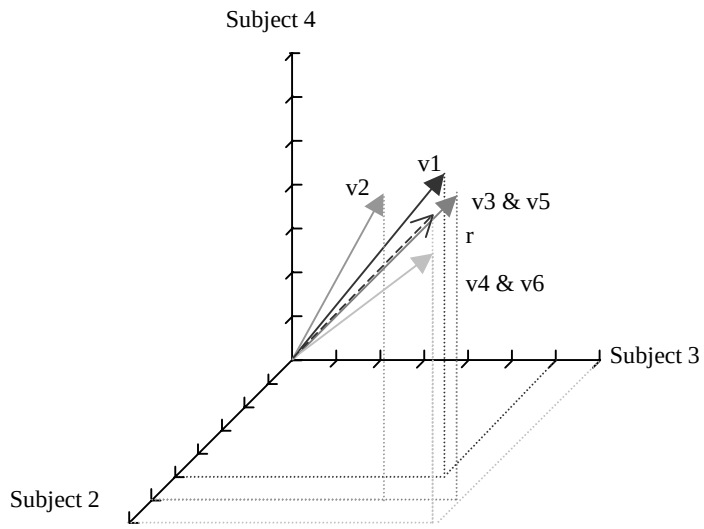

Figure 5. Measured vectors for 6 days representing responses from CG Subjects 2, 3, and 4 (3 subjects for graphing purposes only), plotted with their Reference Vector.

The magnitude ratio between v1 and v2 is 1, v1 and v3 is 0.93, v1 and r is 0.90, and v3 and r is 0.97, from which the following 'similarity' relationships are derived:

$$||v1|| = ||v2|| < ||v3|| = ||v4|| = ||v5|| = ||v6|| < ||r||$$

Table 3 is a symmetric matrix that lists the angle between vector pairs. The vector pairs, (v3, v5) and (v4, v6), are coincident and the Table shows an angle = 0 degrees for them. v1 and v3 are separated by 4.7 degrees, while v1 and v4 yield the largest difference of this data set of 10 degrees (magnitude ratio = 0.93 for both pairs). Thus, v1 is closer

(more similar) to v3 than it is to v4. Cluster analysis techniques can be applied to systemically determine how the vectors form similar clusters.

Table 3. Angle between vector pairs (in degrees).

	v1	v2	v3	v4	v5	v6	r
v1	0	6.4	4.7	10.0	4.7	10.0	7.6
v2	6.4	0	7.8	10.0	7.8	10.0	7.6
v3	4.7	7.8	0	6.0	0.0	6.0	3.7
v4	10.0	10.0	6.0	0	6.0	0.0	3.7
v5	4.7	7.8	0.0	6.0	0	6.0	3.7
v6	10.0	10.0	6.0	0.0	6.0	0	3.7
r	7.6	7.6	3.7	3.7	3.7	3.7	0

This analysis is scalable and can be applied to all other staffs, except for one essential difference. Not all staff members responded to the question for all the days, like the CG staff did. For example, only 7 MNIG staff members posted answers for all six days. Thus, six 7-element vectors share a common vector space and can be compared to each other. In essence, missing responses reduce the number of orthogonal dimensions in the vector space.

Comparing a vector from one staff with a vector from another staff is non-trivial, because these two vectors come from two orthogonal spaces (i.e., the participants between staffs are not the same). There must be common aspects between Subject 1 in staff A and Subject 1 in staff B, etc. Otherwise Subject 1 in staff A and Subject 1 in staff B must be treated as orthogonal dimensions. The vector method provides an indirect comparison by first comparing a staff's vector to its reference vector, and reporting the results as a percentage of the reference vector – effectively normalizing the results and making it easier for the analyst to compare percentages and determine their similarity.

The Vector Method substitutes the notions of multi-dimensional vector spaces and similarity for the notions of variance and significance. The method does not claim to derive significance from the data set in the same manner that statistical methods can do, but rather it provides a means of combining and comparing data from different groups of any sample size.

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		
(baseline) 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.

a. $\alpha = .025$ one-tailed; $\alpha = .05$ two-tailed

n_2 \ n_1	3		4		5		6		7		8		9		10	
	T_L	T_U	T_L	T_U	T_L	T_U	T_L	T_U	T_L	T_U	T_L	T_U	T_L	T_U	T_L	T_U
3	5	16	6	18	6	21	7	23	7	26	8	28	8	31	9	33
4	6	18	11	25	12	28	12	32	13	35	14	38	15	41	16	44
5	6	21	12	28	18	37	19	41	20	45	21	49	22	53	24	56
6	7	23	12	32	19	41	26	52	28	56	29	61	31	65	32	70
7	7	26	13	35	20	45	28	56	37	68	39	73	41	78	43	83
8	8	28	14	38	21	49	29	61	39	73	49	87	51	93	54	98
9	8	31	15	41	22	53	31	65	41	78	51	93	63	108	66	114
10	9	33	16	44	24	56	32	70	43	83	54	98	66	114	79	131

Table 6 [3]

n1 and n2 are both “7”, thus the range according to Table 6 is 37 to 68. Our result of the Wilcoxon rank-sum test is 48.5 and 49.5 for the two comparisons. These values fall within the range from Table 6, thus there was no statistical significance between the trial runs.

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 they 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
Trial 1	6	16	685
Trial 2	13	27	811
Trial 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 paragraph is derived from 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 electronic surveys for larger target groups.

MNE4 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 face-to-face interviews 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.

The UR2015 experience with face-to-face interviews 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, while UR2015 spent 2 weeks performing 8-hour segments daily resulting in 3-1/3 days of elapsed time. The analyst needs to be cognizant of the collaboration the may occur between experimental hours and take this into consideration when designing questions and performing the analysis. Also, experimental control needs to take positive steps to control and minimize this interaction. This can be easily accomplished by explaining to participants how data can be contaminated when information is shared outside of the experimental context.

Both experimental timeline designs 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? 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 MNE4 analyst indicated that his timely completion rate was high due to the focus area leader's persistence with his fellow players. This is one solution to avoid contamination of data. In conjunction with persistence, a full explanation of the consequences of the players' hard work being contaminated by "less than perfect" survey completion may amplify the need for conscientious and timely survey completion. Another potential solution is to either remove incomplete electronic surveys upon reaching an overdue time or the analyst removing the data points should he/she determine the data points are contaminated.

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 experimental design plan is developed in conjunction with the data analysis and 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 this paper argues that the 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 issues of sample size and data collection method needs to be considered for each experiment. 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.

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. Therefore, Future multinational experimentation research can be considered in the area of "bootstrapping" or re-samplings of large samples sizes.

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, Christine H. Mills, Gabriel Rouquie, Mike Wahl, Charles T. Wall, and Candice M. Pink

We 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 Final Report (Draft) 03 Jan 07

[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