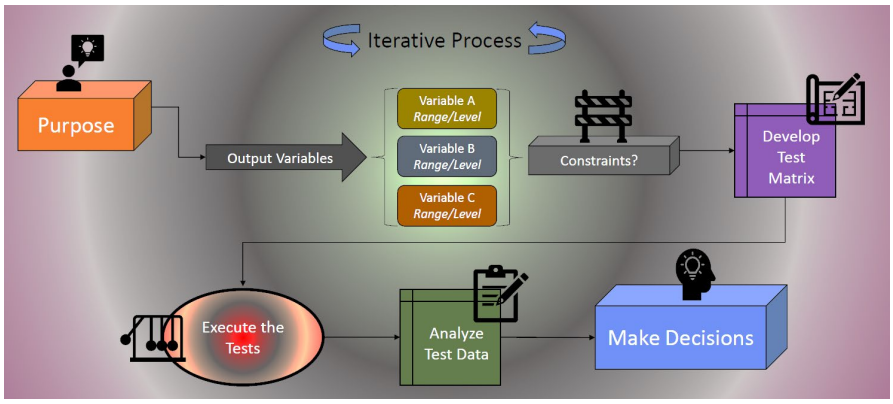

FIGURE 11-2 DESIGN PROCESS AND MEASURES

11.1.8 Execute / Collect the data

"*No plan survives first contact with the enemy.*" - Moltke the Elder. This could be the easiest step; the plan is in place and just needs to be followed. But what if things go wrong? It is important to plan for contingencies. What if a test does not go the way it was expected to? Is there enough time or resources to repeat the experiment? What if a resource believed to be available and thus planned for is not, what is the back-up plan? What if there is an unexpected "visitor", would the presence of a foreign element add to or prevent the experiment? It is also imperative to collect as much data as possible/practical. Some elements may not have been part of the original plan, but if recording an additional bit of data is relatively easy and inexpensive, then do so. That data might be useful later and can prevent the unfortunate position of having to repeat an experiment because something was not recorded and later discovered to be important.

11.1.9 Analyze the Data

Once the data is in hand, it is ready to be processed. From this point, the data needs to be checked and cleaned for any anomalies, and basic summary statistics conducted. Organizing it into tables, or even creating something as simple as a histogram might provide insight on the makeup of the data and the relationships between the variables. Building graphs and creating other visuals can help see what the data is saying. General exploring can be useful, but the focus should be on following the analysis plan that was created and determining the expected results/relationships are supported.

The purpose for analysis is to sift through the data and take it from its raw data form to information to knowledge and hopefully gain some insights along the way. Analysis is about finding the significance or implications of the experimental findings. It is about determining the relationship between variables, understanding the amount of impact they have on each other and potentially recommending areas for future research. It is important to relate results to the aims of the experiment through summarization and explanation and to lead cross-initiative analysis of the generated data. Oversight of specific initiative analyses for internal consistency and connection can contribute to overall program goals. Thus, it is important to maintain documentation, previous event results, and associated relevant data.

When a study is conducted at the end of the experiment, a bunch of numbers remain: the input values (or settings) and the output values (or results). The challenge is to extract from the data a meaningful summary of the behavior observed and a meaningful conclusion regarding the influence of the experimental treatment (independent variable) on participant behavior. Statistics can provide an objective approach to performing this process.

This document will not discuss the various ways statistics can be used or the statistical tests that can be used to process data. At this time, it is simply stated that there is more to data that just the mean and the variance. One key pitfall that must be avoided is thinking that correlation equals causation. Correlation is when two sets of data seem be closely associated with each other. An example of the pitfall would be the link between ice cream sales and sunburn. A quick look at that data would show that the values are highly correlated. Every

month that has high ice cream sales has many sunburn cases and months with low ice cream sales do not. The correlation says they must be linked; therefore, ice cream causes sunburn. This is of course not true, it just so happens that during summer months when it is hot, people tend to like ice cream and going outside. Summer/hot weather is the cause for both, and there is no real connection between ice cream and sunburn. This type of issue is why it is critical to have SME availability during the analysis phase to ensure that the interpretation of the results is appropriate.

11.1.10 Interpret Results

One of the most important things to look at when interpreting the results are the objectives; and it is also important to ask the questions again that needed answers. Now that the data is available, can the questions be answered? At the start of the experiment there were one or more hypotheses proposed (see section 11.2.4). Does the data support them? Going into the experiment assumptions were made regarding cause-and-effect relationships. Were the relationships observed?

11.1.11 Output Reports

The final stage of an experiment/campaign is reporting the results in collaboration with stakeholders and Focus Area Leads. This reporting can come in many forms, from a simple bullet background paper that lists the very basics of the experiment and the results, all the way up to a formal paper that will be published in a technical journal or as a standalone book.

The real goal of a study is not the data, but the answers to the key questions. To truly do that, the results must be published in some manner in a report, a presentation, or some other way of disseminating the findings. The key stakeholders will want to know not just the conclusions, but the work involved. There may be a very specific format required (such as a report related to an acquisition process for submittal to the Requirements Oversight Committee (ROC)); or the format may be more open in structure. It is crucial that clarity and transparency be maintained. If the results of the experiment were ambiguous then the report should reflect that. If there is solid evidence of a finding, that should also be clearly communicated.

The format for the simpler reports varies considerably and depends heavily on the audience. The key to any good report is clarity. It needs, in the simplest way, to share the results of the experiment and what was learned. It might need some background on the problem and the basics of the experiments structure, but the real focus should be on the findings.

One common way to share experimental results is a scientific paper. Most scientific papers have the same structure in the writing and publishing process. In general, there are seven sections: Title, Abstract (a short summary of the paper), Introduction (provides background information and includes the hypothesis), Materials and Methods (the details about how the experiment was done), Results (the relevant data collected from the experiment), Discussion/Conclusion (explains the data and how it either supports or does not support the hypothesis), and Literature Cited (lists references relevant to the experiment).

As mentioned above, the most important is the results section. This is the heart the research paper. Here, statistical analyses of the collected data are presented using text, tables, and figures. Remember, statistical analyses do not prove anything, they only provide guidelines as to the reliability and validity of the results.

11.2 Appendix B - Data Collection and Analysis Plan

"If you fail to plan, you are planning to fail." - Benjamin Franklin

Properly designed and executed experiments enable the advancement of knowledge, understanding of systems/operational concepts within the operating environments and situations of relevance. These processes can be compared to that of building a house. There are three main phases to the construction of a house, design, building and finally, the move in. An experiment is similar as it is planned, executed, and then the data obtained is analyzed.

There are many steps that must take place prior to groundbreaking, such as consulting with the buyer to ensure requirements and expectations are established, followed by drafting of blueprints. Developing stakeholder relationships paired with detailed planning are just as essential to ensure the experiment design and

analytical strategy align with experiment objectives/sub-objectives. From this point, available resources and potential alternatives can be investigated. Just as a house is more likely to succeed if the preliminary steps are done, experiments entail an iterative process whose level of success is dependent upon many sub-elements and planning intricacies. The planning process is a continuous cycle that must be monitored throughout experiment design and execution.

As the supporting structure for experimentation, the DCAP provides analytical rigor specific to the experiment objective(s) and associated line(s) of inquiry (LOI). Developing a predetermined research method is a critical strategy to observe, collect, assess, and report on experiment data. Research design and methods are different, but closely related, because good research design ensures that the data obtained will more effectively answer the research question. Each experiment must have a DCAP that is tailored to the experiment that entails a holistic approach towards data collection and analysis procedures.

The DCAP comprises the main elements of the experiment, including the problem statement, objectives/sub-objectives, LOIs, CLAs, critical questions, measures of performance (MOPs) and measures of effectiveness (MOEs). Furthermore, it provides the analytical rigor required to collect and assess quantitative and qualitative data. Just like a house, the DCAP could be thought of as the blueprint, providing detail and clarity to see the individual parts/components, as well as enable participants to see the big picture. Each step in the experiment will add a piece of knowledge helping to fill the requirements, just as each brick laid helps establish the house. With each experiment, a body of knowledge advances, starting with the knowledge already in hand and expanding on it when possible and reinforcing where necessary. Developing a research method is a critical strategy to observe, collect, and report on data obtained during an experiment. Research design and methods are different, but closely related, because good research design ensures that the data obtained will more effectively answer the research question. For any type of experiment, the DCAP explains the end-to-end structure.

The study purpose is typically the first thing to be established, answering the question of why conduct the study. If it is associated with a capability or knowledge gap, that will be prominently featured as the problem statement. If the purpose for the experiment is more demonstration in nature, there might not be a problem statement, but instead have a key feature that is to be highlighted. Next, the hypothesis is developed, asking the question of what cause-and-effect relationship is trying to be established. This will in turn lead to the metrics and data that is required to answer the questions and establish the cause/effect relationships. Each of these elements will be discussed in detail below.

11.2.1 Problem Statement

"If I had an hour to solve a problem, I'd spend 55 minutes thinking about the problem and five minutes thinking about solutions." - Albert Einstein

One of the most important parts of a journey is knowing the destination. One can have great snacks, a car full of gas, and the best playlist, but without knowing the destination it is hard to map the route. It is the same for an experiment campaign, starting with the end goals in mind is critical to overall success. While it would be wise to not take Einstein too literally and spend 95% of the time on this step, it does merit significant time, thought, and energy to ensure the problem is fully understood before undertaking the design. A big part of that understanding is background research, often called a literature review. The experiment should **add** to the body of knowledge. If something has already been done, it is not adding, only repeating (some things bear repeating to provide verification that the previous endeavor was accurate). In general, however, experiments should generate new information. This background research is the foundation the house will be built on.

A clearly articulated problem statement should address three areas: The capability gap, the key stakeholders, and the needed capability. Important sources that can be used for researching these are Integrated Prioritized Capability Lists (ICPLs), Integrated Priority Lists (IPLs), Navy Lessons Learned Information System (NLLIS) and Joint Lessons Learned Information System (JLLIS). ICPLs and IPLs capture the major warfighter gaps and provide details on why they are important and who the stakeholders are. NLLIS and JLLIS provide tactical and operational lessons learned for experiment planners and fleet operators. These lessons can

provide the framework for development of doctrine, TTP, concepts of operations, or for improving naval and joint operations of current combat systems, including systems approaching initial operational capability.

It is important to periodically review the campaign plan from top to bottom to verify that everything is still applicable, as well as to incorporate any new information. The results of one experiment might indicate that the original goal is either unreachable or too easy. An experiment might have gone even better than expected and provided data that makes the next planned experiment unnecessary. When this happens, the experiment can be restructured to look at another factor. In some cases, data might have been lost and the experiment needs to be repeated. It is also possible that the original problem has changed, and what was once a capability gap is no longer a problem. Perhaps a new gap has been found, or an unexpected result might indicate that the tactic/technology can solve a need that was not part of the original scope. The big picture might be completely different than it was at the start of the project. For this reason, periodic review of the problem statement can ensure that it is still applicable to the problem at hand.

11.2.2 Research

"Those who do not learn history are doomed to repeat it." - George Santayana

Research is a key element to any experiment. Literary research starts with a general idea and a need to know more, much like an informal discovery experiment. It is often not hypothesis-central, instead the focus is to see what happens in the environment in which this idea is tested. In most cases, the intent of an experiment is to add to the existing body of knowledge, and/or to make a comparison of the new system against the current standard. The answers to questions examining the current standard are critical and will be used not only in the pre-experiment development phase, but also in the post experiment analysis phase. If it is known how the old system was tested, the new can be tested in the same way to do a side-by-side comparison. Without knowing the answers to these questions, it is difficult to demonstrate that the new system is better.

11.2.3 Objectives

Many activities are done "just for the fun of it" such as playing a game of tag, jumping out of an airplane, or climbing a mountain. And that may be true for a small number of experiments, they are done just to see what happens. It may be that A, B & C have never been done at the same time, but now doing that may make a difference. However, a good experiment, especially for an experiment campaign, the experiment needs to have an objective or goal in mind. Also, at each step within, the experiment needs to have sub-objectives that lead back to the main objective.

The problem statement and the objectives are not the same things, but they are joined at the hip. Being able to clearly state the problem helps to define the objectives. One of the easiest ways to flush out an initial list of objectives is to look at the problem statement and ask critical questions such as: "*Why is that important?*" This chain of questions and answers helps flush out the objectives and leads to clues as to how they can be measured.

Once the main objective is established and drafted, it can be broken down into parts. Building up the list of objectives and sub-objectives is a matter of asking the how, what, and why. And then ask again until arriving at a specific task or identifying a metric (see metrics in Section 11.2.5). For example, if the overall objective is: "*I want to sell my widget/idea*" or "*I need to impress the decision maker*", how can that be accomplished? If the widget or idea is proven to work, what makes it better than the other options? How is it unique? How can that be demonstrated? Keep asking questions because the answers to these questions will create a list of potential objectives. Not every combination of question and answer will need to be a documented objective but going through the process will help to make sure that a complete picture is created. In theory, completing all the sub-objectives will in turn complete the objective above it. If this is not the case, then consider what else needs to be done to achieve that.

On a multiday road trip, each day might have a destination as a sub-objective and the snacks might be the topic for a separate set of objectives. Example objectives might be healthy and limited snacks, so dinner is not ruined. Just one change could impact multiple objectives. Lots of snacks are needed to replace lunch to avoid a stop and make better time. Perhaps an accident heard on the radio causes a change in the overall route.

In that case, many of the sub-objectives might also need to be changed. The traveler may receive news that a friend needs help, in which case the main objective is changed and head out in a completely different direction. Similarly, changes might need to be made to an experiment campaign plan as it is being executed. New knowledge leads to adjustments; a new gap might be discovered, or the gap originally targeted might already be filled.

11.2.4 Hypothesis

A hypothesis is a statement or claim that has yet to be supported with data. It proposes a cause-and-effect relationship between two elements of concern. A well-crafted hypothesis helps to focus an experiment and points it in the right direction for what and where to investigate. It typically has two parts, the independent variable in the "if" half and the dependent variable in the "then" half. Examples could be as simple as using a new fuel additive to extend the range the fleet can cover between refills, an approach of changing factor A to improve factor B. It is important for the planners to document their hypothesis before designing the experiment to ensure that the correct data can be collected.

A hypothesis test is the process of determining if there is enough evidence support the proposed cause/effect relationship. However, statistics can be a little tricky in a way, statistical evidence can never prove that something is true. Instead, evidence is used to establish that the likelihood an opposite statement (also called the alternative hypothesis) is so small that the alternative must be false and in turn the original statement can be "considered" true. In a court of law, the public wants to know if someone is innocent or not, however a defendant cannot be found innocent. Instead, the opposite question is asked to prove/disprove the question of guilt. If there is not enough evidence, they are not found innocent but found **not** guilty. The assumption is first that a person is innocent and may prove they are not. In a similar way, the data will not prove the original hypothesis, it proves that the alternative is not true.

Consider testing a new sensor to establish that the new sensor is better at identifying targets then the legacy system. The hypothesis could be as simple as "the new sensor is better." But "better" is very vague and is hard to define. Is the new system more cost effective? Is it more compact? Does it have a higher ID rage? Is that under specific conditions? A higher quality hypothesis might be, "By incorporating the new sensor package, target identification processing time can be increased 50% by eliminating the need for second looks." This version states not only the "if" (incorporating the new sensor package) and "then" (increase target identification processing time 50%), but also adds a possible reason for the impact. Now there are two things to test for: the processing time, and the number of times a second look was required.

Consider an example where the primary hypothesis is that a new sensor has a higher detection rate than the legacy system. The alternative hypothesis is that the two sensors are not different, but in fact produce similar results. If the experiment provided the following data:

	Day light conditions	Low/No light conditions
Legacy	75%	55%
New system	80%	75%

These are different, but are they different enough to reject the alternative hypothesis that the two systems are equal in capability? In effect, is there enough evidence to convict? In this case (unless the values represent a very large sample set) for daylight hours there is not enough evidence to dismiss/reject the alternative hypothesis, and the two systems might in fact be the same. As a result, there is no data to reject the alternative and thus cannot say anything about the original hypothesis. For the low/no light, the likelihood of getting values this different while they are the same is very low. Thus, the alternative hypothesis can be rejected while supporting the original hypothesis that the new sensor does have a higher detection rate under low/no light conditions.

11.2.5 Metrics

One of the main questions that the DCAP needs to answer is what is going to be measured or collected. There are many ways to describe a metric, but the most important way is simply good versus bad. A good metric will help answer the most important question, "**So what**?" What was learned? Does it make a difference? Was an important point discovered? Depending on the topic being studied, different questions will be important; and in turn will influence the metrics that need to be collected. It is important to not focus on one category of data as being superior to another. Like most things dealing with analysis, what is important depends on many factors. The best advice regarding metrics is to think about the "so what" and do not ignore data that can be harvested. If in doubt, it is best to over-collect and sort it out/analyze it later.

Good metrics have three qualities: Valid, Reliable, and Credible. *Valid* metrics means that the measurements being taken are true indicators of the situation. Looking at the color of a strip of bacon would be a valid way to measure the thoroughness of cooking; however, in the case of a roast beef, the color only reveals the surface. The temperature at the core of a roast determines if it is cooked to completion. In this situation, one would ask "What is a valid metric in relation to the subject being measured and what is being determined?" *Reliable* metrics are consistent metrics. In this case, one would conduct an experiment under the same circumstance as a previous experiment and then see if the same result be recorded. With the thermometer and roast, a reliable thermometer inserted into different areas of the same roast should have identical readings. This can be used for metrics that are qualitative in nature (see Quantitative versus Qualitative below) or are opinion-based. *Credible* metrics are synonymous with what can be believed or trusted. Asking a colorblind person a question that is dependent on reading a digital number makes sense, but asking that person if a certain shirt clashes with pants would yield a non-credible opinion.

In general, metrics can be sorted three different ways: Quantitative versus Qualitative, MOE versus MOP, and level of measurement. In many cases, the way a question is asked can impact the way it is measured and the type of data that is to be collected. So, think about the question and be sure the "so what" can be answered. To help explain the differences in the data types consider an example of testing a new pistol below and look at the types of measurements that might be made and how they fall into each category.

11.2.5.0 Qualitative vs. Quantitative

Some things can be counted or measured and given numerical values, others are hard to determine numerically. Those metrics that can be counted are considered "quantitative," and while they might not always be the easiest to count and measure, the value that is recorded is not subjective nor a matter of opinion. How many times did a thing happen? How far can it go? How fast is it? Questions that fall into the "qualitative" category are subjective in nature and are opinion-based, such as "Do you like it?" and "How much do you like it?" It is important to rank these qualitative options and list advantages and disadvantages.

For the pistol example, one might ask the following: *How much does it weigh? How many rounds can it hold?* These are measurable or countable and as such are quantitative. Next, one might ask the following: *Does it feel comfortable in the hand? Does it look good with this outdoor gear?* Those considerations are a matter of opinion. Depending on the hand-size and the shape of available grips, a comfort and a preference scale can be used but the results are very subjective. Some questions can be both qualitative and quantitative depending on how the topic is addressed: *How much recoil is there? Is it too much? Does it impact accuracy?* In the hands of a trained shooter the recoil might not be significant, but for a novice shooter the same recoil could be significant. The recoil can be counted by setting it on a stand and firing the weapon, but to measure the impact would be less accurate perhaps? It can also be tempting to try to turn a qualitative observation into a quantitative one, by rating the comfort level 1 to 5. The answer will be a number, but treating those subjective opinions like true quantitative metrics can be dangerous.

11.2.5.1 Measures of Performance vs Measures of Effect

MOPs and MOEs are often mistakenly thought to be the same, but they are as different as the questions they answer. MOP's answer the question "what?", whereas MOE's answer the question "why?" MOPs are all about **performance**, they often point inward and are focused on actions and what was done. They answer

questions along the lines of: "What did you do? What can you do?" With the pistol example, one might ask about its effective range or misfire rate. Did the warhead explode? MOE's are about **effect**, or impact. They are typically focused on the second; and even the third order of effects from actions. For MOEs the questions are: *"Why did you do it? What was the impact? Did you see the behavior you wanted? Was the building targeted sufficiently damaged?"* For the pistol example, an MOP might be as follows: *"How many rounds per minute can be fired? What is the number of rounds capacity? What is the muzzle velocity?"* The MOEs would be stopping power, ability to penetrate soft skin vehicles, or comfort level in-hand.

Some people look at MOPs as answering the question, "Are we doing things right?" Likewise, they are looking at MOE's as answering the question, "Are we doing the right things?" Below is an example involving an experiment to study the impact of going to the gym, losing weight, and making new friends. In this example, free gym memberships are sent to collect metrics (MOP's and MOE's).

The MOP's and MOE's for this example can be quite simple. An MOP might be related to the tasks of going to the gym and completing a workout. The potential MOEs are the impact targeted with this new behavior, body shaping, weight loss, etc. The following questions are asked: "Did you go to the gym three times a week? How many hours of cardio did you complete in the week? Did you interact with other people while there?" These generate the MOP's. Some of the answers are very easy to count with a yes or no answer. Some answers are harder to measure, such as greeting the employee checking IDs only, or making eye contact without verbal communication.

In theory, it is good to define what is required to count a MOP as successful, but not always best to let the subject being measured know this in advance. The question "How many times did you go to the gym this week?" is better than "Did you go three times?" The second question will only receive a yes/no answer and provide limited opportunity for detailed analysis. For instance, a person attending three times a week could look equal to a person attending six times a week. The first question will get better data, and if three is a key number, the detailed data can be used to split results into two groups and generate the data received from question two. Additionally, question two could create a bias in the study. The way the question is asked could impact the subject's behavior, and in turn, the study. If the subject hears a question about three times a week, the subject may see that as a goal. The subject would then stop at three or not bother with a second trip. This is an example of why one should carefully consider how a question is asked.

The MOEs can be straight forward: "Are you losing the weight that you set out to lose?" "Do you feel more self-confident about your looks?" "Are you making new friends that will encourage healthy activities?" As with the MOPs, some are easy to measure, like the weight. Step on a scale every day for a month and record the value, what does the needle say? Making new friends can be a little harder to measure. The subject may say, "I see the same people there every time I go, so I recognize them but never talk to or see them outside the gym" versus "I talk to them all the time both inside and outside the gym, we are practically best friends now." But where on that scale is a "Yes" and where is a "No?" It is critical to understand the difference between the two and how they work together to provide a complete picture. There are four combinations of ways that MOPs and MOE's can interact listed in the table below. When looking at the combinations, it is good to remember that the purpose for an experiment is to gain knowledge about something not previously explored in this setting.

MOEs / MOPs	Accomplished	Failed
Accomplished	Good (#1)	It depends (#3)
Failed	It depends (#2)	Bad (#4)

Combination #1 looks like it is all good news, a big win for the team. But remember that the purpose for the experiment is to gain knowledge. Always ask "What can we learn from this?" and "Why does that matter?" First the MOPs indicate what can be done and concerning the MOE's, #1 provides evidence that the assumptions on the association between dependent and independent factors seems to be true. By doing the things set out to do, it is possible to achieve the intended effects. But besides the fact that those things are

possible, what was learned? Having the right metrics enables the analysis and gaining understanding of the "so what." How did the things done impact the effects targeted? How are they linked?

As good as #1 looks, #4 looks bad, but does not have to be considered a complete loss. Remember the purpose for an experiment and think about what was learned. If tasks were not completed, why? If the desired results were not produced, why? The cause-and-effect relationship seems to be intact, but is it? Is the model of the situation sound? What can be learned from the data? Even a "failed" experiment can provide valuable information that will make the system/processes better, if only one can learn from them. Like most things in life, if one can learn from a situation, then there is value. Focus on the purpose for the experiment and learn something.

For cases #2 and #3, because the two types of metrics had different results, it can provide a significant learning experience. In case #2, everything was done that the team set out to do and MOPs are green, but the expected impact was not accomplished. Does this mean that the tasks and objectives are not connected as previously thought? For case #3, even without completing all tasks, the intended effect occurred anyway. Was there something else that caused the result? Perhaps not every task needed to be completed to achieve the results.

In the gym example, case #1 is easy to describe. The subject went to the gym all the time, has lost the weight, and made the friends. Case #4 is equally easy. This subject signed up for the membership but has not set foot back in the gym, so they gained more weight and have lost the one friend they thought they had. In Case #2, the subject went to the gym and talked to people but is not losing any weight nor making any new friends. Why is that? Are they not working hard enough on the equipment? Are they not coming across as friendly? In case #3, the subject missed several workouts but is losing weight and made several friends. Are they doing something else that is not being measured in the MOPs, therefore causing the weight loss? In Case #5, this subject found that because people are hot and sweaty while working out, they are not associating with each other. This subject had more impact making friends by talking to people at the juice bar in the lobby.

11.2.5.2 Level of Measurement

Not all metrics are equal in the eyes of analysis and the way a question is asked will impact the data collected and in turn the type and depth of analysis that can be completed afterward. Data can be sorted in to four tiers: Nominal, Ordinal, Interval and Ratio. A different meaning can be extracted with each tier. Nominal data is typically categorical in nature, but in its basic form there is no sequence that each category should be placed in, such as blood type. Thus, minimal meaning can be extracted. For ordinal data there is a logical sequence, but relative size is meaningless, such as rating something poor/ fair/good/excellent, or age groups: infant, toddler, child, teenager…. The sequence is obvious, but how much better is good than fair? Interval data provides the answer to the range of difference between data points, such as temperature: 30 ºF, 60 ºF degrees and 90 ºF degrees are all exactly 30 ºF degrees different. So, there is true meaning in not just their sequence but also in the intervals between the points. However, 60 ºF is not twice as hot as 30 ºF. Interval data is stronger than ordinal, but it lacks a true zero and some mathematical operations cannot be performed. Ratio data on the other hand has a true zero. This is data such as age, weight, and height. To maximize the analysis that can be conducted afterward, it is always best to collect the highest tier data possible.

Using the example of a foot race, in a list of participants, nominal data would be the names of the schools represented by each runner. A histogram showing the number students represented by each school could be created but it would yield very little else. Ordinal data would show placement at the finish line, first/second/third. This data could provide ranking of the runners but would offer no insight for the span between the finish times for each of them. Was the result of the race a blow out or was it very close? Interval data is the amount of time between each runner, but if two runners crossed the finish line five seconds apart, is that a big gap? In a marathon where they run for hours, five seconds is almost nothing, but in a 100-meter sprint, five seconds is a lifetime. The ratio data for a race is the actual finish times. This data would enable the most complete analysis.

11.2.6 Data Collection Methods

There are many methods for collecting data within an experiment, ranging from completely manual to fully automated. The tools that are used to measure the metric and the way they are recorded should be established in advance and captured in the DCAP. What tool should be used depends on the nature of the data sets. For data sets that are very qualitative in nature a manual process might be preferred. With an expert in place to make the evaluation, a manual process can provide the details and insights needed to make an assessment. For a data set that is very quantitative in nature an automated process might be preferred. While some factors for the collection method might change by situation, it is normally a good idea for collections to be as discreet as possible and be done in a way that will ensure the integrity of the data.

11.2.6.0 Accuracy vs Precision

Accuracy refers to how close a measurement is to the true or accepted value. Precision refers to how close measurements of the same item are to each other. Precision is independent of accuracy. That means it is possible to be very precise but not very accurate, and it is also possible to be accurate without being precise.

A classic way of demonstrating the difference between precision and accuracy is with a shooting target. Think of the center of the target as the true value. The closer the shots land to the center, the more accurate they are. The tighter the cluster, the more consistent or precise. A tight grouping away from center is precise (but not accurate), a scattering that is uniform around the center is accurate but not precise. The best quality observations are both accurate and precise. If there is only one known, it is possible in the post processing to try to account for this but not always. For example, if all shots form a circle around one point, that could be accuracy but not precision. Averaging the values would provide a potential answer for the true value. On the other hand, a very tight grouping of shots says consistent or precise, but unless the target is known to be down and to the left, it will be difficult to fix the accuracy.

A key question that is often asked is how much accuracy and precision are needed, and the answer is, it depends on what is measured and the impact of being wrong. For example, the weight of a battleship is in tons, if one is off by a few pounds here or there the impact is minimal. If one is measuring the components for a satellite, the impact of being off by even a fraction of an ounce could mean being off balance and spinning out of control, with enormous impact.

11.3 Appendix C – Technology Readiness Level

As a new technology is developed, it starts as an idea in someone's head and eventually becomes a finished product. The defense acquisition community has developed a system to rate technology as it matures and provide a snapshot on its status, the Technology Readiness Levels (TRLs). The primary purpose of using TRLs is to help management in making decisions concerning the development and transitioning of technology. It should be viewed as one of several tools that are needed to manage the progress of research and development activity within an organization. Most exercises have a minimum TRL rating to be considered for inclusion, as shown in Figure 11-4. See the tables below[28] for descriptions of the TRLs and use to evaluate where the technology fits.

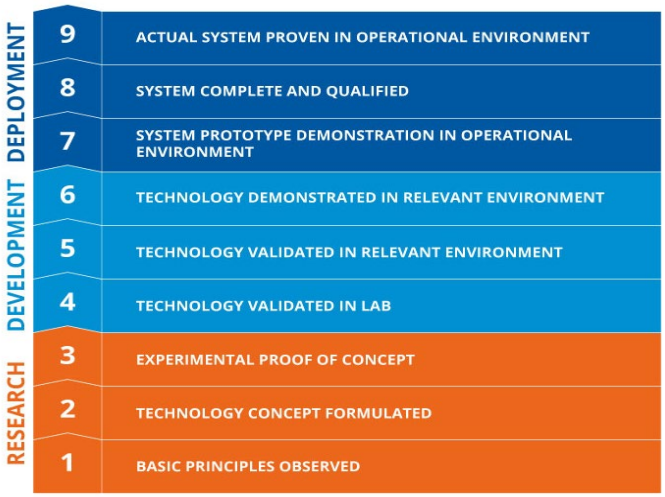

FIGURE 11-3 TRL LEVELS

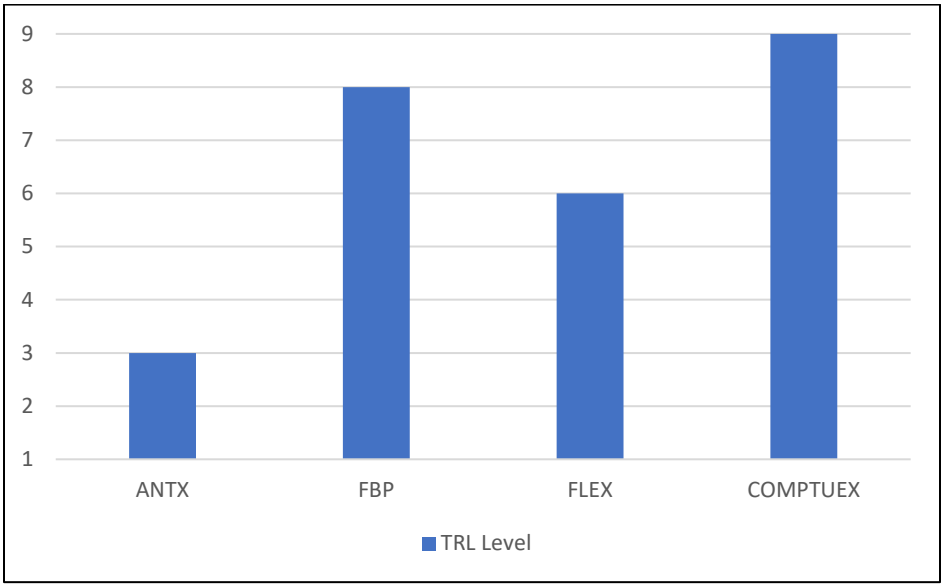

FIGURE 11-4 EVENT TRL LEVEL

[28] Technology Readiness Assessment Deskbook, 2009

Hardware TRL Definitions, Descriptions, and Supporting Information		
TRL Definition	**Description**	**Supporting Information**
1 *Basic principles observed and reported.*	Lowest level of technology readiness. Scientific research begins to be translated into applied research and development (R&D). Examples might include paper studies of a technology's basic properties.	Published research that identifies the principles that underlie this technology. References to who, where, when.
2 *Technology concept and/or application formulated.*	Invention begins. Once basic principles are observed, practical applications can be invented. Applications are speculative, and there may be no proof or detailed analysis to support the assumptions. Examples are limited to analytic studies.	Publications or other references that outline the application being considered and that provide analysis to support the concept.
3 *Analytical and experimental critical function and/or characteristic proof of concept.*	Active R&D is initiated. This includes analytical studies and laboratory studies to physically validate the analytical predictions of separate elements of the technology. Examples include components that are not yet integrated or representative.	Results of laboratory tests performed to measure parameters of interest and comparison to analytical predictions for critical subsystems. References to who, where, and when these tests and comparisons were performed.
4 *Component and/or breadboard validation in a laboratory environment.*	Basic technological components are integrated to establish that they will work together. This is relatively "low fidelity" compared with the eventual system. Examples include integration of "ad hoc" hardware in the laboratory.	System concepts that have been considered and results from testing laboratory-scale breadboard(s). References to who did this work and when. Provide an estimate of how breadboard hardware and test results differ from the expected system goals.
5 *Component and/or breadboard validation in a relevant environment.*	Fidelity of breadboard technology increases significantly. The basic technological components are integrated with reasonably realistic supporting elements so they can be tested in a simulated environment. Examples include "high-fidelity" laboratory integration of components.	Results from testing a laboratory breadboard system are integrated with other supporting elements in a simulated operational environment. How does the "relevant environment" differ from the expected operational environment? How do the test results compare with expectations? What problems, if any, were encountered? Was the breadboard system refined to more nearly match the expected system goals?
6 *System/subsystem model or prototype demonstration in a relevant environment.*	Representative model or prototype system, which is well beyond that of TRL 5, is tested in a relevant environment. Represents a major step up in a technology's demonstrated readiness. Examples include testing a prototype in a high-fidelity laboratory environment or in a simulated operational environment.	Results from laboratory testing of a prototype system that is near the desired configuration in terms of performance, weight, and volume. How did the test environment differ from the operational environment? Who performed the tests? How did the test compare with expectations? What problems, if any, were encountered? What are/were the plans, options, or actions to resolve problems before moving to the next level?
7 *System prototype demonstration in an operational environment.*	Prototype near or at planned operational system. Represents a major step up from TRL 6 by requiring demonstration of an actual system prototype in an operational environment (e.g., in an aircraft, in a vehicle, or in space).	Results from testing a prototype system in an operational environment. Who performed the tests? How did the test compare with expectations? What problems, if any, were encountered? What are/were the plans, options, or actions to resolve problems before moving to the next level?
8 *Actual system completed and qualified through test and demonstration.*	Technology has been proven to work in its final form and under expected conditions. In almost all cases, this TRL represents the end of true system development. Examples include developmental test and evaluation (DT&E) of the system in its intended weapon system to determine if it meets design specifications.	Results of testing the system in its final configuration under the expected range of environmental conditions in which it will be expected to operate. Assessment of whether it will meet its operational requirements. What problems, if any, were encountered? What are/were the plans, options, or actions to resolve problems before finalizing the design?
9 *Actual system proven through successful mission operations.*	Actual application of the technology in its final form and under mission conditions, such as those encountered in operational test and evaluation (OT&E). Examples include using the system under operational mission conditions.	OT&E reports.

TABLE 4 HARDWARE TRL LEVEL DESCRIPTIONS

Software TRL Definitions, Descriptions, and Supporting Information		
TRL Definition	**Description**	**Supporting Information**
1 *Basic principles observed and reported.*	Lowest level of software technology readiness. A new software domain is being investigated by the basic research community. This level extends to the development of basic use, basic properties of software architecture, mathematical formulations, and general algorithms.	Basic research activities, research articles, peer-reviewed white papers, point papers, early lab model of basic concept may be useful for substantiating the TRL.
2 *Technology concept and/or application formulated.*	Once basic principles are observed, practical applications can be invented. Applications are speculative, and there may be no proof or detailed analysis to support the assumptions. Examples are limited to analytic studies using synthetic data.	Applied research activities, analytic studies, small code units, and papers comparing competing technologies.
3 *Analytical and experimental critical function and/or characteristic proof of concept.*	Active R&D is initiated. The level at which scientific feasibility is demonstrated through analytical and laboratory studies. This level extends to the development of limited functionality environments to validate critical properties and analytical predictions using non-integrated software components and partially representative data.	Algorithms run on a surrogate processor in a laboratory environment, instrumented components operating in a laboratory environment, laboratory results showing validation of critical properties.
4 *Module and/or subsystem validation in a laboratory environment (i.e., software prototype development environment).*	Basic software components are integrated to establish that they will work together. They are relatively primitive with regard to efficiency and robustness compared with the eventual system. Architecture development initiated to include interoperability, reliability, maintainability, extensibility, scalability, and security issues. Emulation with current/legacy elements as appropriate. Prototypes developed to demonstrate different aspects of eventual system.	Advanced technology development, stand-alone prototype solving a synthetic full-scale problem, or standalone prototype processing fully representative data sets.
5 *Module and/or subsystem validation in a relevant environment.*	Level at which software technology is ready to start integration with existing systems. The prototype implementations conform to target environment/interfaces. Experiments with realistic problems. Simulated interfaces to existing systems. System software architecture established. Algorithms run on a processor(s) with characteristics expected in the operational environment.	System architecture diagram around technology element with critical performance requirements defined. Processor selection analysis, Simulation/Stimulation (Sim/Stim) Laboratory buildup plan. Software placed under configuration management. Commercial-of-the-shelf/government-off-the-shelf (COTS/GOTS) components in the system software architecture are identified.
6 *Module and/or subsystem validation in a relevant end-to-end environment.*	Level at which the engineering feasibility of a software technology is demonstrated. This level extends to laboratory prototype implementations on full-scale realistic problems in which the software technology is partially integrated with existing hardware/software systems.	Results from laboratory testing of a prototype package that is near the desired configuration in terms of performance, including physical, logical, data, and security interfaces. Comparisons between tested environment and operational environment analytically understood. Analysis and test measurements quantifying contribution to system-wide requirements such as throughput, scalability, and reliability. Analysis of human-computer (user environment) begun.
7 *System prototype demonstration in an operational high-fidelity environment.*	Level at which the program feasibility of a software technology is demonstrated. This level extends to operational environment prototype implementations, where critical technical risk functionality is available for demonstration and a test in which the software technology is well integrated with operational hardware/software systems.	Critical technological properties are measured against requirements in an operational environment.
8 *Actual system completed and mission qualified through test and demonstration in an operational environment.*	Level at which a software technology is fully integrated with operational hardware and software systems. Software development documentation is complete. All functionality tested in simulated and operational scenarios.	Published documentation and product technology refresh build schedule. Software resource reserve measured and tracked.
9 *Actual system proven through successful mission-proven operational capabilities.*	Level at which a software technology is readily repeatable and reusable. The software based on the technology is fully integrated with operational hardware/software systems. All software documentation verified. Successful operational experience. Sustaining software engineering support in place. Actual system.	Production configuration management reports. Technology integrated into a reuse "wizard."

TABLE 5 SOFTWARE TRL LEVEL DESCRIPTIONS

11.4 Appendix D – Installation Processes

Each experiment requires some sort of engineering rigor and risk management process that should be followed. Several risk assessment processes exist, based on the experiment's risk to the ship and what must be done to mitigate it. For example, the fewer DoD components used in an experiment, the easier the process. Similar assessments and testing are applied to aircraft and submarines to understand and mitigate any risks and impacts presented by the experiment. Many of these tests and planning considerations are **common to all experiments or technical demonstrations whether they are on surface ships, shore facilities, aircraft, or submarines**. As stated above, not all of these will be required for every experiment; but the following is a list of considerations likely to be required for any fleet experiment or technical demonstration. For more in-depth information, please consult SBIR 103: Installation Guidebook and the forthcoming Quick Reference Guides.

- The **Navy Risk Management Framework (RMF) for Cybersecurity** applies to all systems – without exception – that receive, process, store, display, or transmit DoD information, including systems participating in Navy experimentation or technical demonstrations with the goal of obtaining Interim Authorization to Test (IATT) prior to the install date for the event. A streamlined RMF process for experimentation has been developed to achieve IATT authorizations and fulfill Cybersecurity requirements using best practices from DoD partners and the Center for Internet Security (CIS) with the goal of improving RMF IATT processing times in support of experimentation requirements and timelines. The streamlined process may not be guaranteed in every circumstance, so the emphasis should be on the earliest possible start for RMF processing to avoid having the experiment stopped due to lack of IATT.[29]
- **Application Integration (AI) Assessment**: SBIR communities should be aware that early planning is crucial for acceptance into the lab testing environment; there are criteria and cyber accreditation requirements needed prior to lab environment entry. AI assessment is required for computing hardware or software integrated on any afloat network. The process is started through submission of the Afloat Service Request Form (SRF). It is then scheduled and executed in government labs with government personnel assisted by commercial submitters. PMW 160 holds sponsorship for the legacy and CANES networks but there are several trusted agents with the ability to accomplish this assessment.[30]
- **Mission Readiness Assessment (MRA)/Combat Systems Integrated Testing**: Assessment and testing provides evidence that systems, software applications, and hardware are functioning properly. It is important to note that if the experiment impacts the Integrated Combat System (ICS), the submitter (sponsor, PARM, or assignee) needs to enter it into the ICS Configuration Control Board (CCB) **before** formal submittal of the SCD through the Navy Data Environment (NDE).
- **Weapon System Explosive Safety Review Board (WSESRB)**: Initial installation testing, qualification testing, physical fit checks, status ground fire testing, systems integration lab (SIL), safety analysis, safe separation test certification, and Non-Nuclear Munitions Safety Board (NNMSB) may be gathered for review and concurrence through this board.
- **Ship Checks and Shore Site Visits**: These are performed in conjunction with the planning yard and the sponsor. Tasking and funding must be in place before they can begin. Ship Installation Drawings (SIDs), if needed, will be developed from information obtained through the Ship Check.[31]
- **Standard Frequency Action Format (SFAF)**: This may be required because some systems receive but do not transmit a signal. Electromagnetic Spectrum (EMS) support is often not required for these receive-only systems; however, these systems can be vulnerable to emissions from other devices.

[29] Appendix Q, Fleet Experimentation and Technology Demonstration, p. 23
[30] PMW 160 Fact Sheet
[31] Appendix Q, Fleet Experimentation and Technology Demonstration, SL720-AA-MAN-030, p. Q-23

Standard frequency action format (SFAF) records are a way to identify the location of these devices for their protection as receive-only systems.[32]

A typical timeline can be seen in Figure 11-5. For more information and descriptions, consult the forthcoming Shipboard Installations Guidebook. FLEX and TECH DEMOs have a slightly different set of processes and requirements. Surface ships and unmanned surface vehicles (USVs) will follow the FLEX or TECH DEMO process described below.

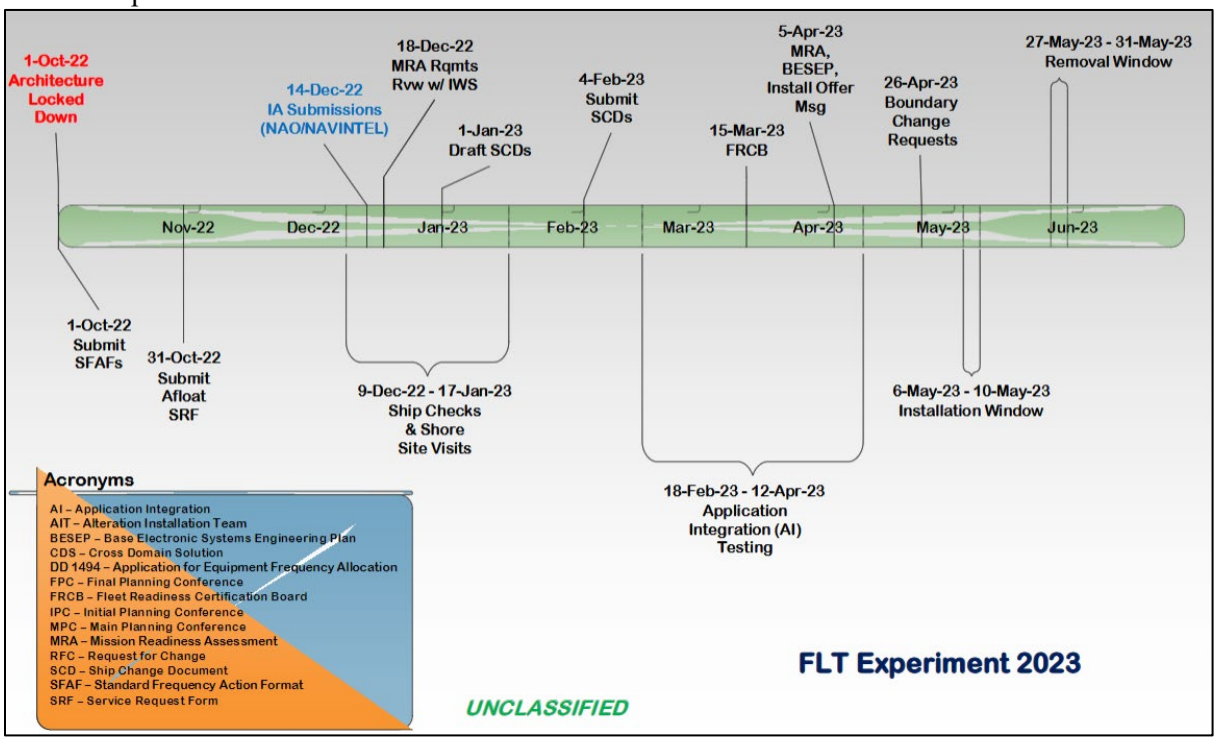

FIGURE 11-5 SAMPLE INSTALLATION TIMELINE

11.4.1 Surface Ships and AEGIS Ashore

Both FLEX and TECH DEMOs will follow the streamlined NMP process for experimentation which either will fall into one or more of 12 installation types and four scenarios (or combinations of each). For FLEX, the experiment will be entered into the FLEX Information Management System (FIMS). A TECH DEMO will not go into FIMS; but will be entered into the Navy Data Environment (NDE). Each of the 12 installation types and scenarios will generate a set of requirements and deliverables that must be met to mitigate risk to the ship. These must be accomplished before execution of the experiment or demonstration on the ship. Low complexity experiments will generate a short list of requirements and high complexity experiments will generate a longer list of requirements.

For most installations that will occur on a ship for the purpose of experimentation or technical demonstration, a ship change document (SCD) will be required. Whether or not a full SCD is necessary will be determined by the type of installation and its level of complexity. The SCD will be filled out in NDE and will serve as a platform for the technical data package (TDP) and produce a tracking number for coordination and communication in relation to reviews and approvals. Almost all TECH DEMOs will need a SCD and most FLEXs will need one in addition to entry into FIMS. In the case of a less complex installation, a header-only SCD will be set up mainly to obtain a tracking number. The timeframe for approval will vary depending on the

[32] CJCSM 3320.01C, Joint Electromagnetic Spectrum Management Operations in the Electromagnetic Operational Environment, Enclosure C, p. C-6

- CNS/ATM certification is required for Research, Development, Test and Evaluation (RDT&E) activities that occur on rotary wing avionics mission systems.[57] [58]

11.4.4 Submarines

11.4.4.0 Temporary Submarine Alterations

The Supervisor of Shipbuilding (SUPSHIP) will need to be notified of a submarine Temporary Submarine Alterations (TEMPALT). SUPSHIP assigns the TEMPALT number and contact should be maintained throughout the process. SUPSHIP will require payment for review of the TEMPALT, so it is important to have funding in place first. In addition to those requirements listed above for ALL experiments, other possible requirements for submarines are as follows: [59]

- Accreditation Package Development/Risk Management Framework (RMF): In addition to the streamlined process that is in place entitled, "NAO – Streamlined Process for Experimentation 2020" described above, submarines will go through a TEMPALT (TA) Cybersecurity Evaluation.[60]
- EMI Surveys: In addition to the EMI considerations described above, EMI surveys for submarines cannot occur in a manner that would impede forward sonar and communications systems access or cause a power-down of systems, unless notification has been provided ahead of time to the EMC technician. To obtain an accurate assessment, forward electronics must be energized like the possible at-sea lineup. EMI surveys must be conducted by NAVSEA or NAVSEA designates.[61]
- Other TEMPALT items to consider are found on the *PMS 392 TEMPALT Submission Checklist* and the *Technical Requirements Manual for Temporary Submarine Alterations.* Both sources are not generally viewable by the public, but the sponsor will likely be able to access them.

The processes for these areas will be tailored toward the unit or domain in which the system will be installed, most have similar guidelines and sub-processes. As discussed above, some processes and assessments are common to any type of experimentation or demonstration. These are listed to provide some help in forecasting possible requirements and the timelines involved to move more efficiently through the processes.

More information, including due regard processes, test plans, and aircraft/submarine/surface ship operations can be found in the SBIR 103: Installation Guidebook.

[57] MIL-HDBK-516C, pp. 49, 51, 114, 395-396, and 489-490
[58] NAVAIR MANUAL M-13034.1, pp. 2-2 and 3-7
[59] SUPSHIP Operations Manual (SOM), S0300-B2-MAN-010 Rev 2, Change #21, Chapter 10
[60] PMS392 TempALT Submission Checklist, p. 1
[61] Joint Fleet Maintenance Manual, Vol. VI, 4.3.2.2