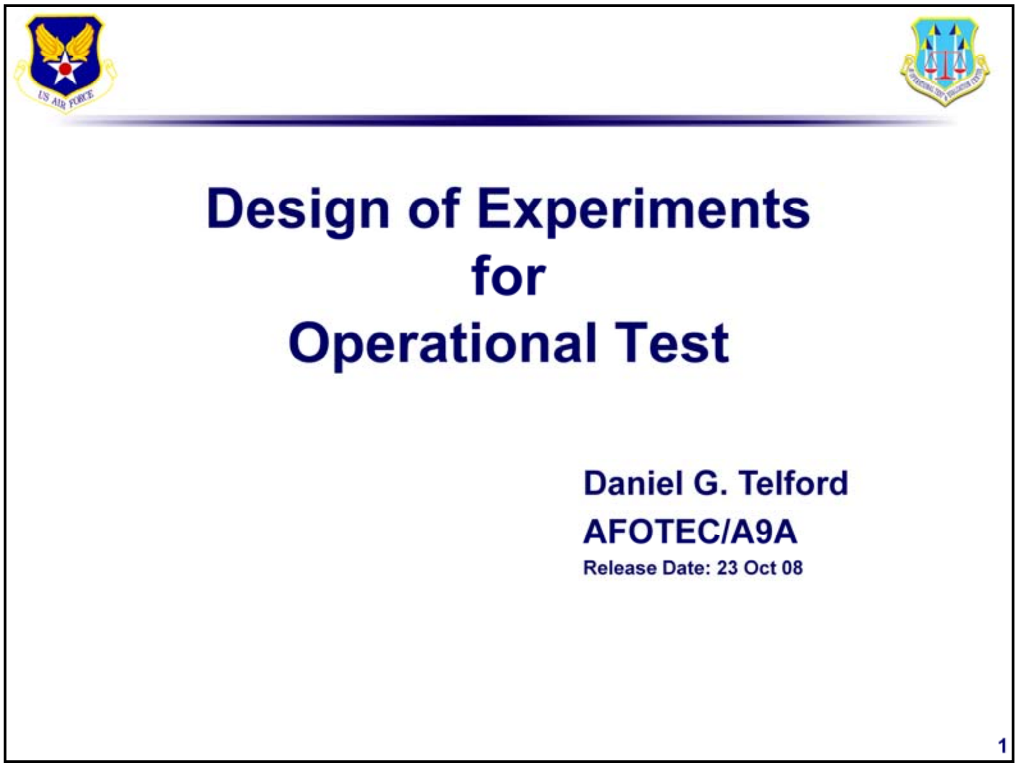
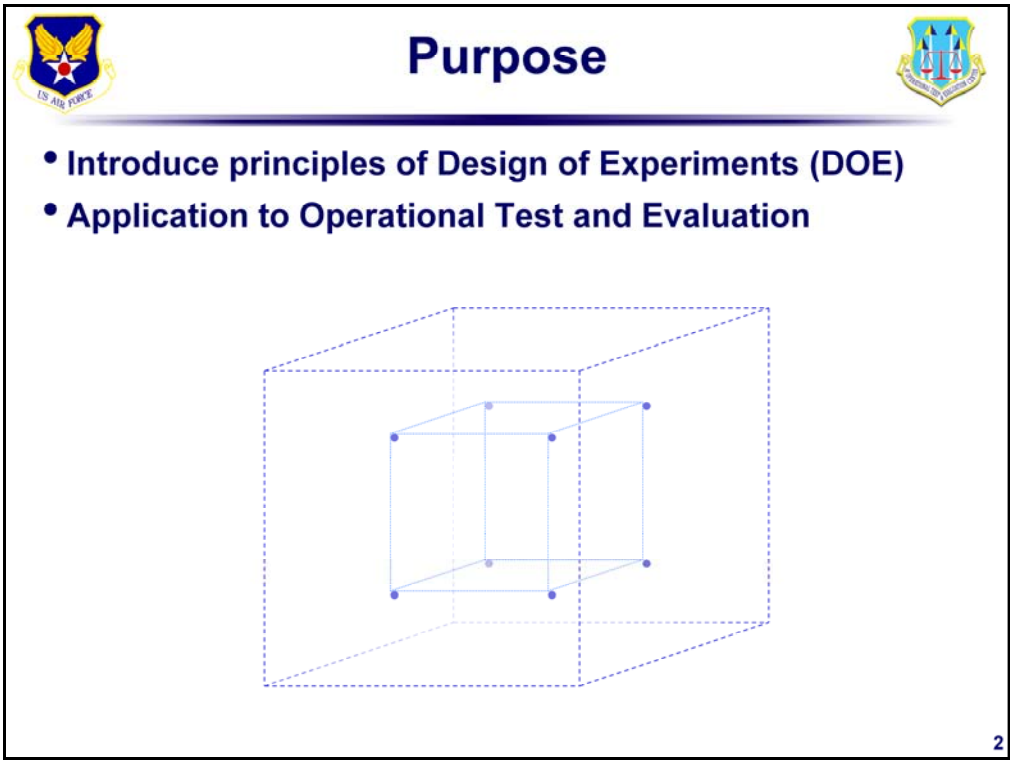


Design of Experiments for Operational Test

Daniel G. Telford
AFOTEC/A9A

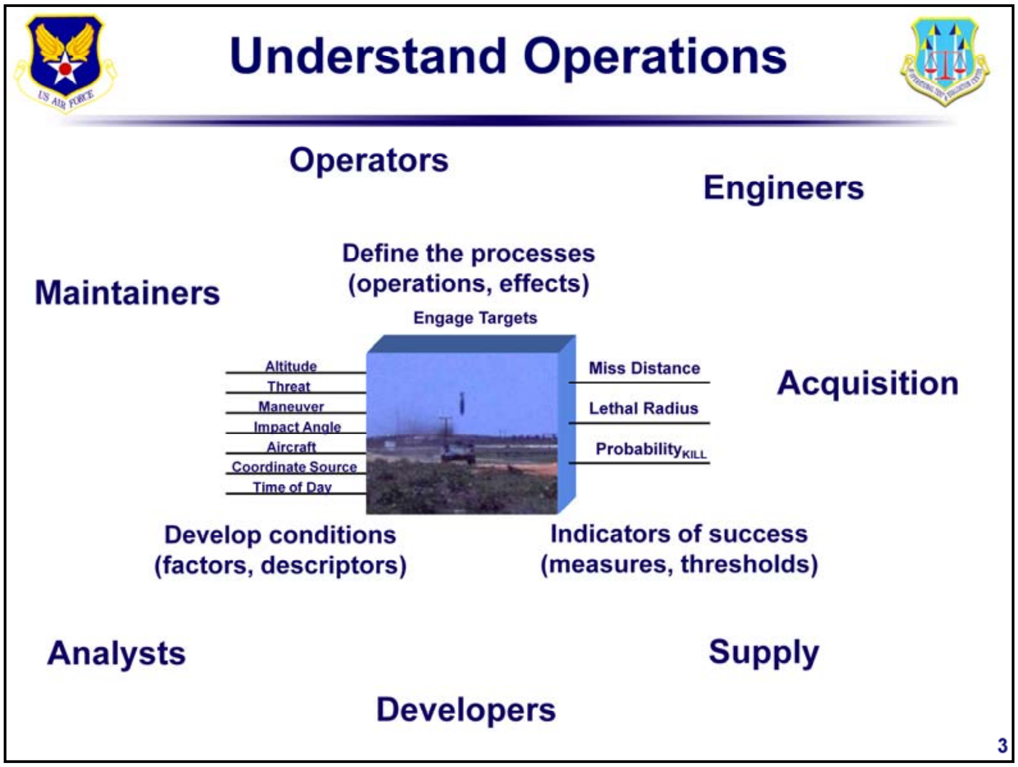
Release Date: 23 Oct 08

Design of Experiments (DOE) is an acronym and a test technique increasingly being used in the T&E community. This will present a conceptual overview of what DOE is, how it compares to other testing techniques, and how it is used for operational test.



Purpose

- Introduce principles of Design of Experiments (DOE)
- Application to Operational Test and Evaluation



Understand Operations

Operators

Engineers

Define the processes
(operations, effects)

Maintainers

Engage Targets

Altitude
Threat
Maneuver
Impact Angle
Aircraft
Coordinate Source
Time of Day

Miss Distance
Lethal Radius
$\text{Probability}_{\text{KILL}}$

Acquisition

Develop conditions
(factors, descriptors)

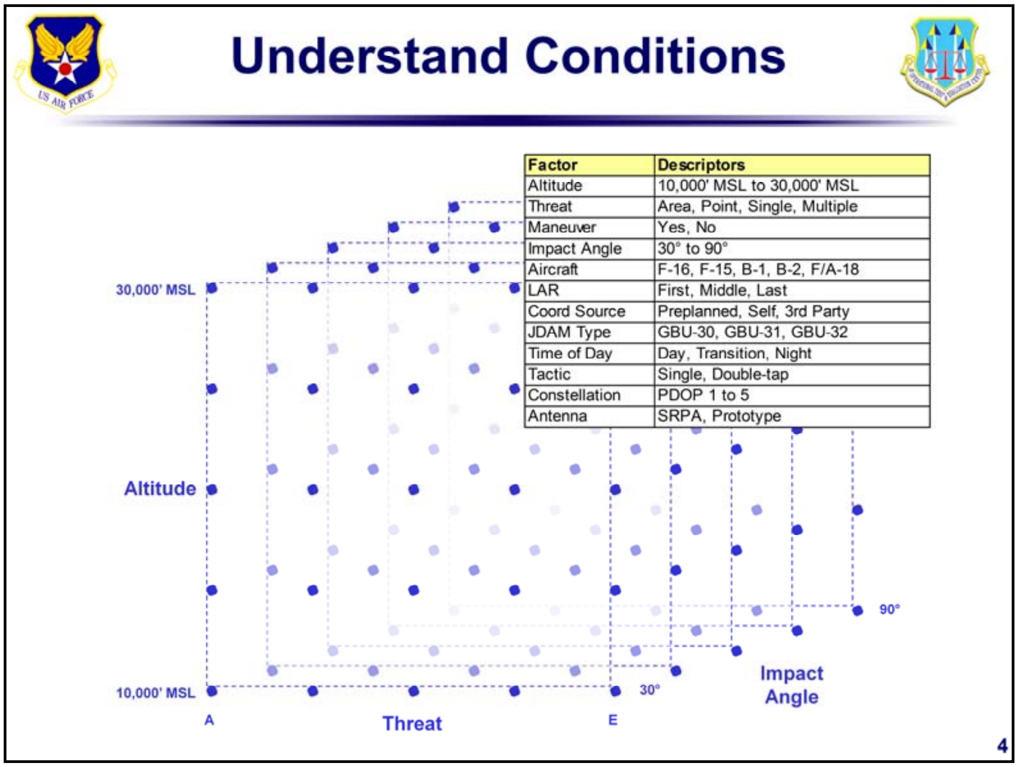
Indicators of success
(measures, thresholds)

Analysts

Supply

Developers

3



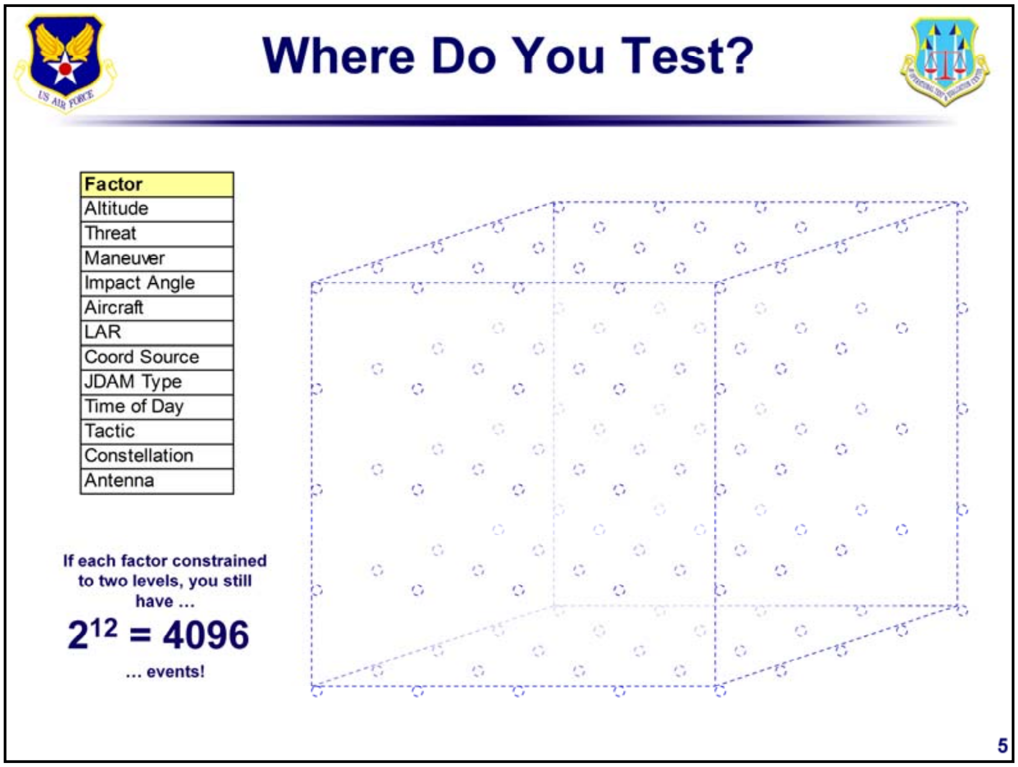
Understand Conditions

Factor	Descriptors
Altitude	10,000' MSL to 30,000' MSL
Threat	Area, Point, Single, Multiple
Maneuver	Yes, No
Impact Angle	30° to 90°
Aircraft	F-16, F-15, B-1, B-2, F/A-18
LAR	First, Middle, Last
Coord Source	Preplanned, Self, 3rd Party
JDAM Type	GBU-30, GBU-31, GBU-32
Time of Day	Day, Transition, Night
Tactic	Single, Double-tap
Constellation	PDOP 1 to 5
Antenna	SRPA, Prototype

Battlespace conditions can be specified in terms of factors and levels for factors. A 2D example, using “engage targets with a bomb” as an example, might be the altitude an aircraft is flying at and the type of threat or threats there are in the environment. Operationally realistic levels are assigned to the factors—for example, 5,000’ to 25,000 MSL for altitude.

The combination of factors and their levels describe the battlespace conditions.

Of course, the battlespace is very multidimensional, not just two factors. I’ll use just three factors to discuss DOE, but there are typically many more. For testing under operational conditions, we identify the factors we think could influence the outcome of an operation. Even with just three factors, this is a rather large set of possible operational conditions.



Where Do You Test?

Factor
Altitude
Threat
Maneuver
Impact Angle
Aircraft
LAR
Coord Source
JDAM Type
Time of Day
Tactic
Constellation
Antenna

If each factor constrained to two levels, you still have …

$2^{12} = 4096$

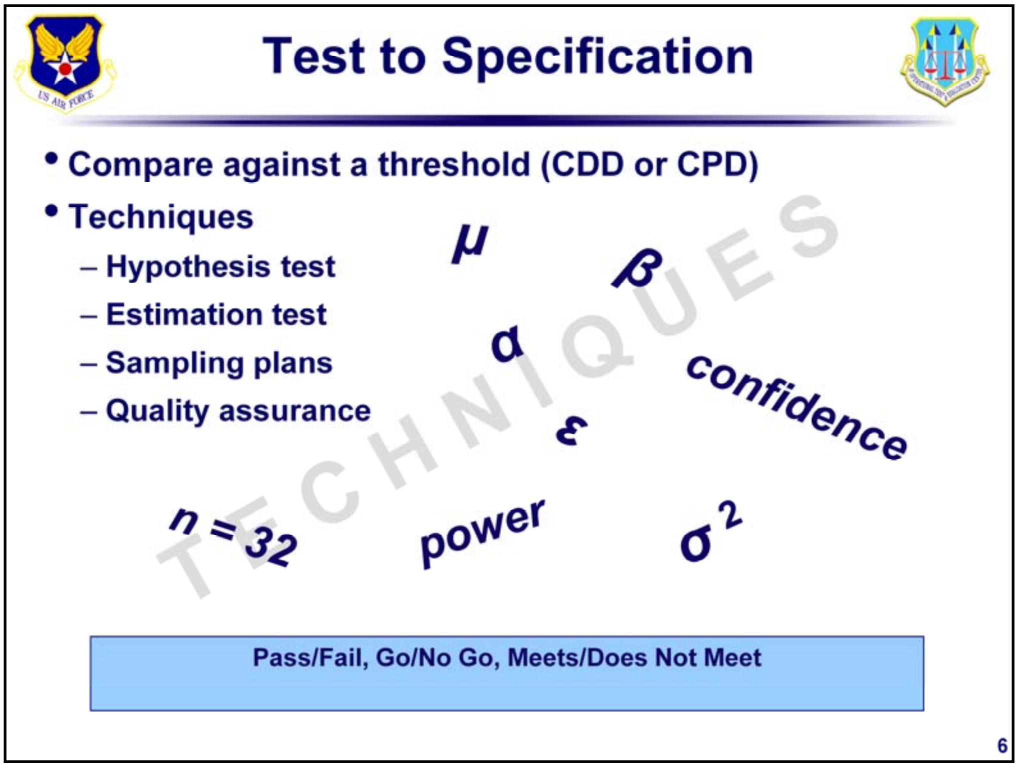
… events!

5

Out of this large, complex, operationally realistic environment, where do you test—which “points” do you pick to conduct your test. Remember, they are all operationally realistic. These are the candidate, operational conditions to help form our OT&E. So where do you test?

Part of the answer is with our “Operationally sufficient” criteria—we need to cover a breadth of conditions. But we need additional criteria to help guide the selection of test points.

This is where we introduce the “test” part of OT&E to marry with the operational part of OT&E.



Test to Specification

- Compare against a threshold (CDD or CPD)
- Techniques
  - – Hypothesis test
  - – Estimation test
  - – Sampling plans
  - – Quality assurance

TECHNIQUES

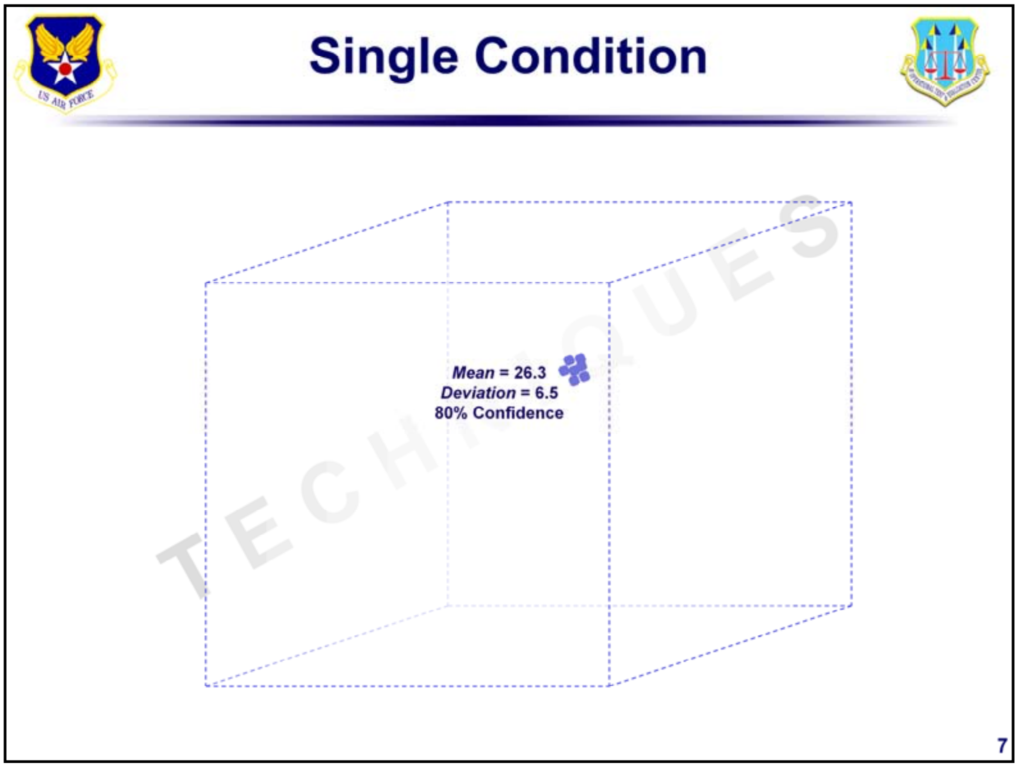
μ β α confidence ε

$n = 32$ power σ^2

Pass/Fail, Go/No Go, Meets/Does Not Meet

6

Historically, the “bread and butter” of testing. Typically done with a hypothesis test—does it meet a threshold at some confidence level. Invokes alphas, betas, power, confidence, variance, standard deviation, error, and sample size issues.

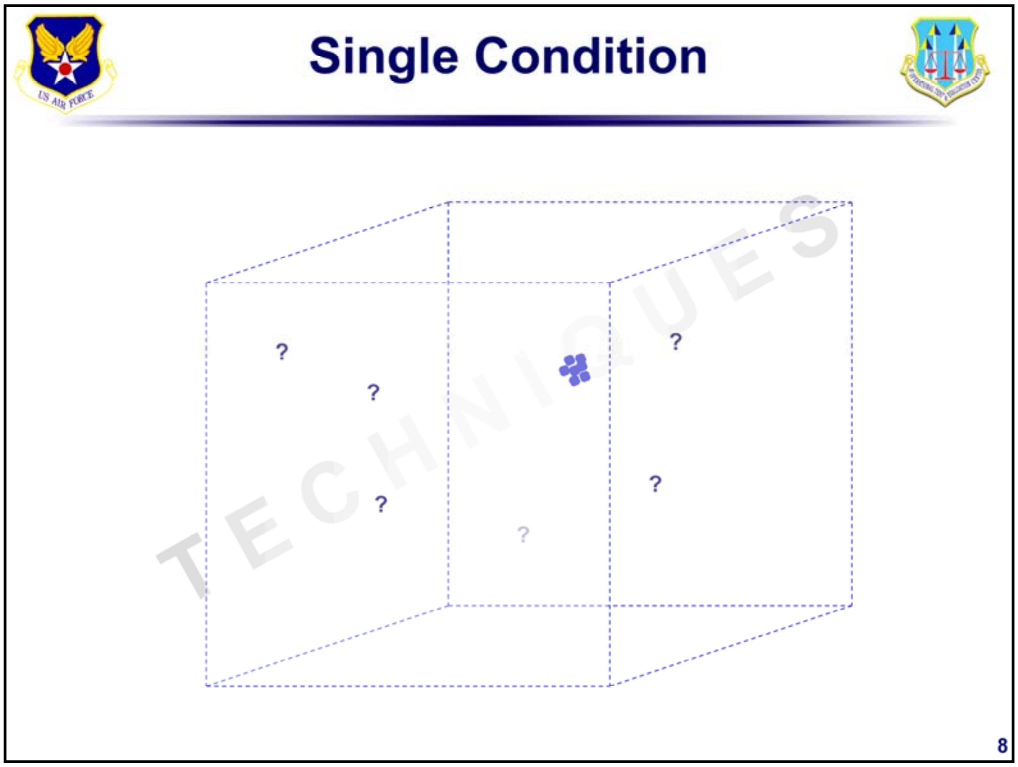


Single Condition

Mean = 26.3
Deviation = 6.5
80% Confidence

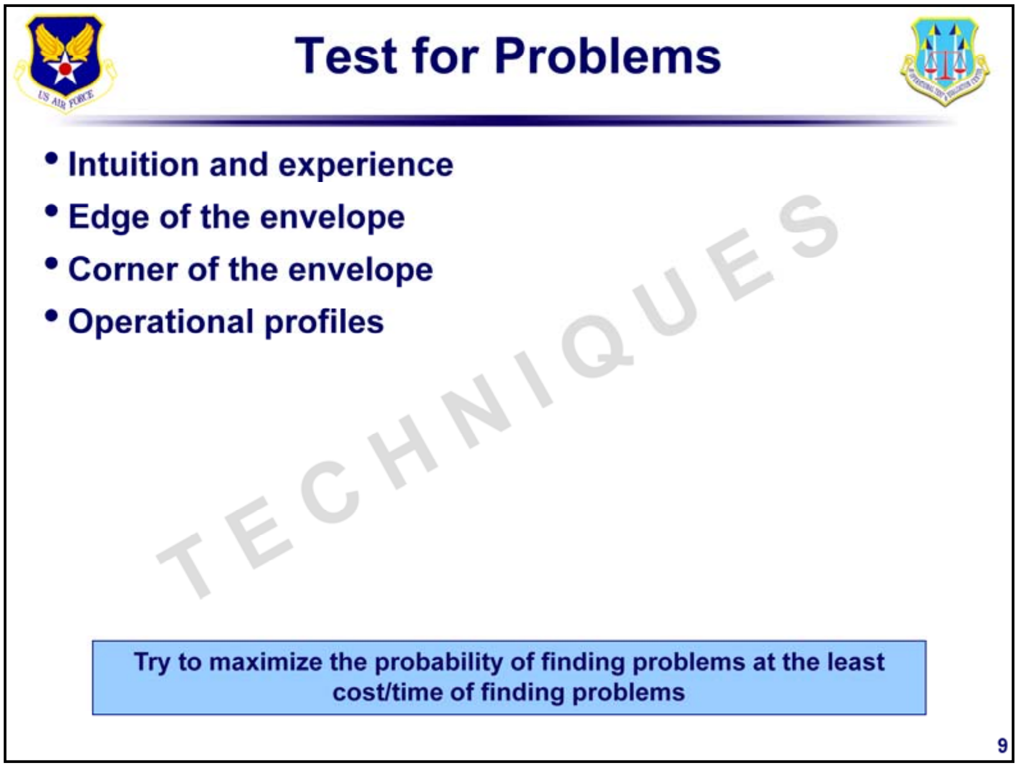
7

Done “properly,” only at one condition (otherwise any variation in results is biased, unexplained, confusing).



Single Condition

As a reminder, this should be done under one condition. This gives you a very good estimate of performance, but only under that condition. It doesn't tell you anything about performance at other points in the battlespace.



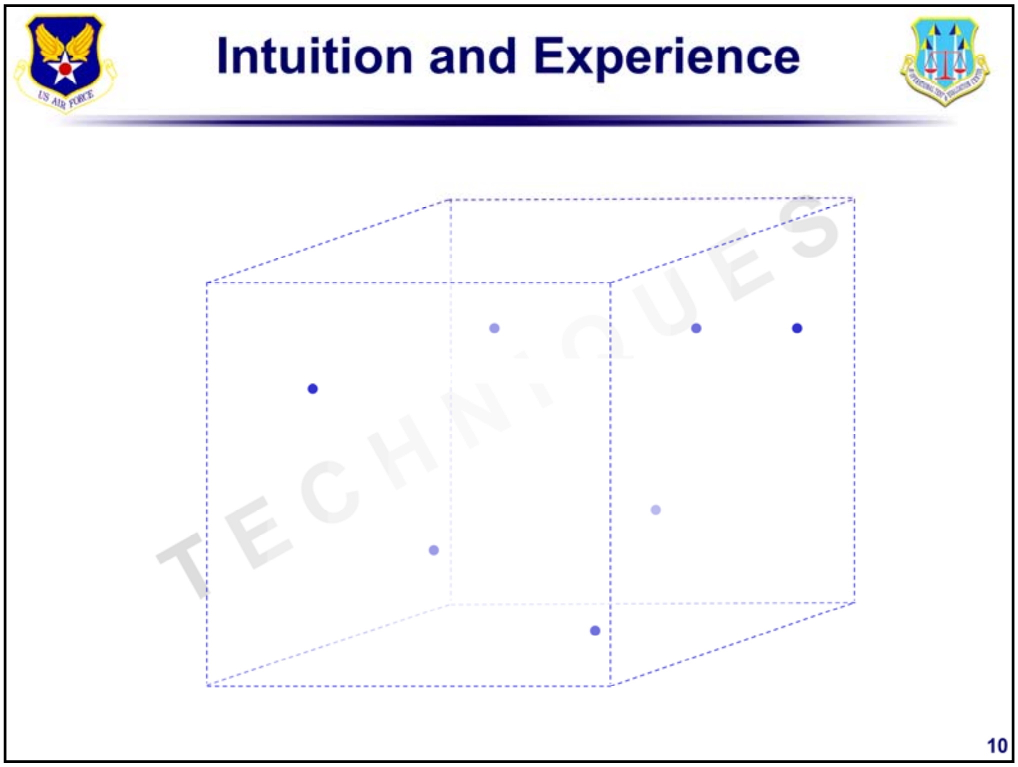
Test for Problems

- Intuition and experience
- Edge of the envelope
- Corner of the envelope
- Operational profiles

Try to maximize the probability of finding problems at the least cost/time of finding problems

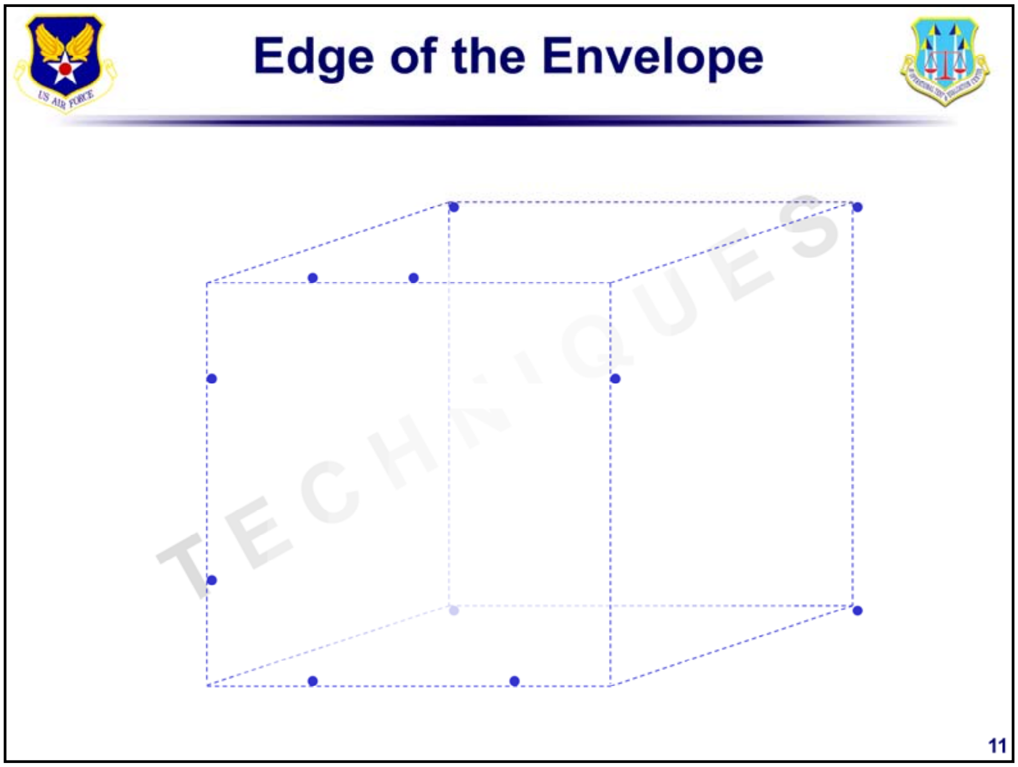
9

A test tactic commonly used, but not realized.



Intuition and Experience

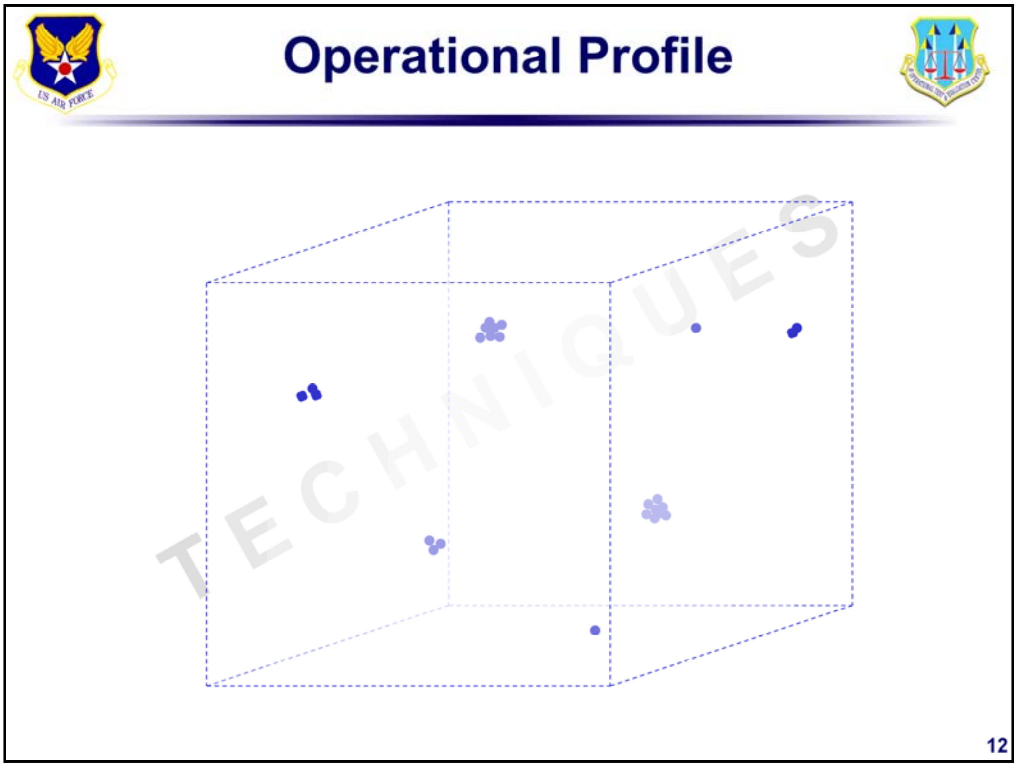
Can be very effective IF all the right SMEs are involved. Poor for computing metrics, however, since its purpose was to find problems.



Edge of the Envelope

11

aka Stress Testing. Based on the assumption that if it works at the edges/corners, then everything “inside” is OK. May not be a valid assumption. Also, you know nothing about behavior “inside” the envelope, especially if you do find problems.

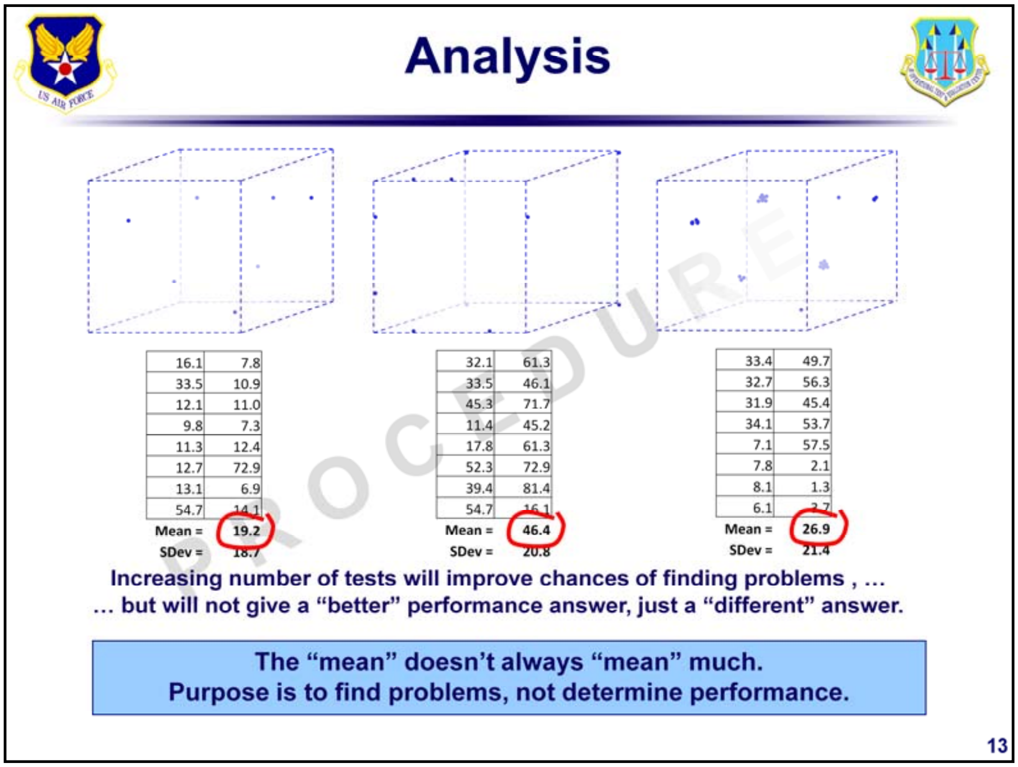


Operational Profile

12

Variation on select conditions. Number of points and location based on the expected operational profile and frequency. Good for reliability testing in that it will find the most frequent problems based on usage. Poor for computing metrics (weighted points and randomly scattered)

One size fits all, but not very well.



Analysis

16.1	7.8
33.5	10.9
12.1	11.0
9.8	7.3
11.3	12.4
12.7	72.9
13.1	6.9
54.7	14.1
Mean =	19.2
SDev =	18.7

32.1	61.3
33.5	46.1
45.3	71.7
11.4	45.2
17.8	61.3
52.3	72.9
39.4	81.4
54.7	16.1
Mean =	46.4
SDev =	20.8

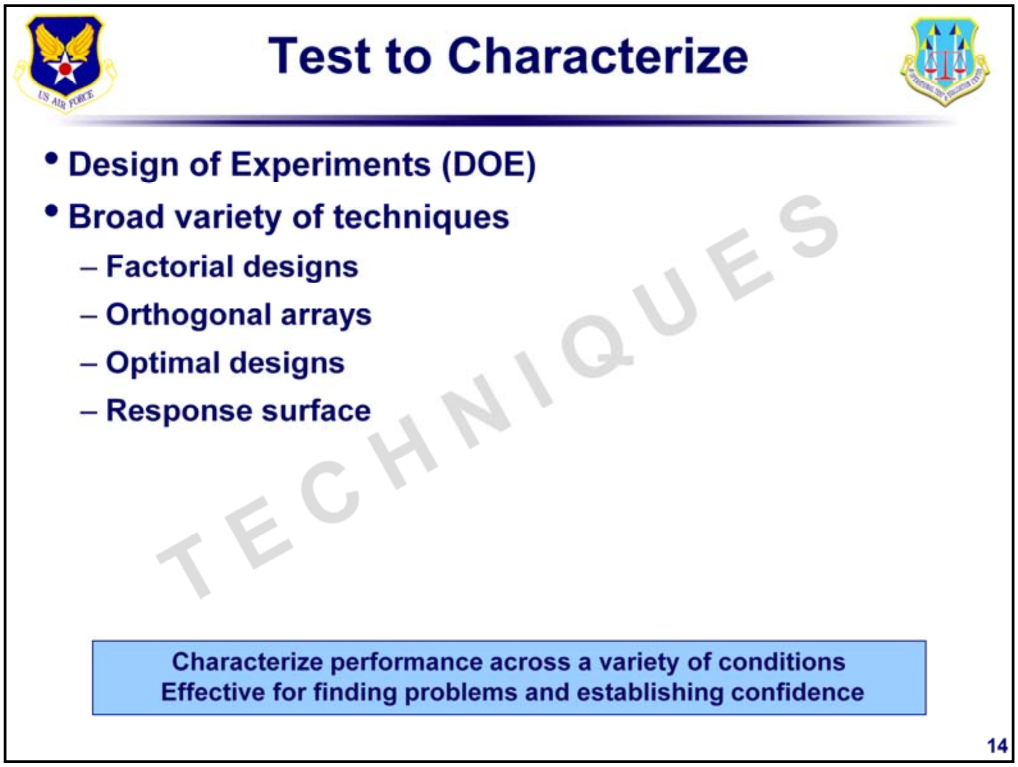
33.4	49.7
32.7	56.3
31.9	45.4
34.1	53.7
7.1	57.5
7.8	2.1
8.1	1.3
6.1	2.7
Mean =	26.9
SDev =	21.4

Increasing number of tests will improve chances of finding problems , …
… but will not give a "better" performance answer, just a "different" answer.

The "mean" doesn't always "mean" much.
Purpose is to find problems, not determine performance.

13

A challenge with all of these techniques is that they target finding problems—this does not necessarily make it a good tactic if you want to compute metrics (such as the average miss distance). Depending on exactly which points you pick, especially if you've done a "good" job of picking points with problems, the metrics you measure could vary quite a bit. Increasing the sample size may improve your chances of finding problems, but won't give you a better (more accurate or correct) performance answer, just a potentially different answer. Again, the whole purpose is to find problems, not compute a metric such as an average.



Test to Characterize

- Design of Experiments (DOE)
- Broad variety of techniques
  - Factorial designs
  - Orthogonal arrays
  - Optimal designs
  - Response surface

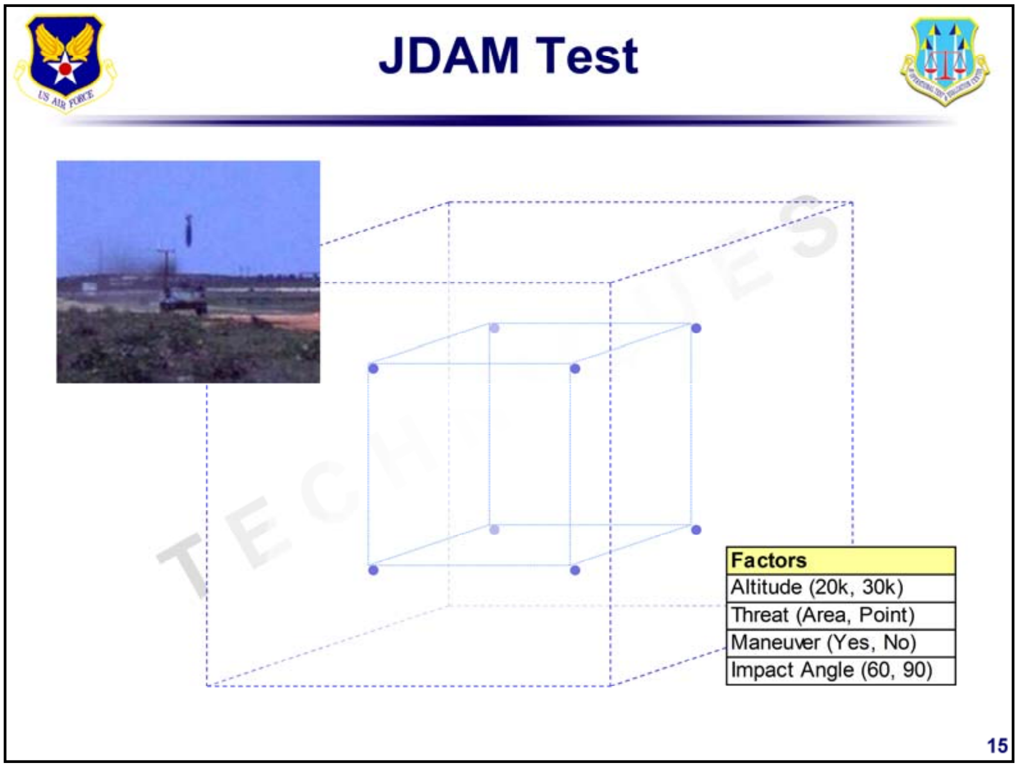
Characterize performance across a variety of conditions
Effective for finding problems and establishing confidence

14

Testing to characterize performance across the battlespace conditions is based on employing Design of Experiments or DOE. There are a variety of techniques that can be used such as factorial designs, orthogonal arrays, optimal designs, and response surface methods. The most common technique, and a very powerful and effective technique, are factorial designs. We'll use factorial designs as our example of using DOE.

The techniques we'll talk about have been used for over 100 years, having early origins in the late 1800's. Sir Ronald Fisher pioneered the principles of design of experiments in the 1930's—sometimes he is called the "father of DOE." During WWII and after, DOE was adopted by industry and has been using continuously since.

DOE is very effective and efficient at characterizing performance across a variety of conditions; it also happens to be effective finding problems and for establishing confidence or decision risk.



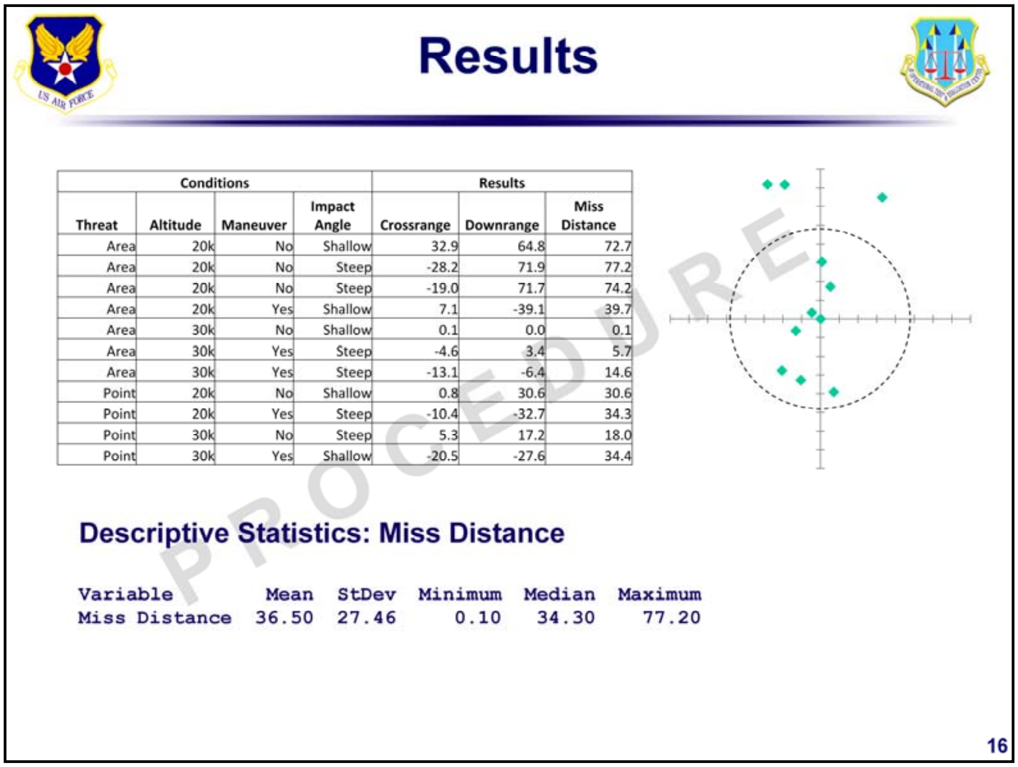
JDAM Test

Factors
Altitude (20k, 30k)
Threat (Area, Point)
Maneuver (Yes, No)
Impact Angle (60, 90)

An example, based on an actual test, is the JDAM. DOE principals were used to test the JDAM across a variety of conditions in this “structured” manner. Numerous factors of the battlespace were identified. Based on previous testing, experience of operators, planners, engineers, etc., and an understanding of the physics, four factors were highlighted as possibly affecting the outcome. As we’ll see, DOE allowed us to tie these factors and combinations to the outcomes—though the outcome wasn’t always what people intuitively thought it should be.

FAQs

- Non-linear behavior across the battlespace can be checked by adding “center points” on each central axis, both within and outside the “cube.”
- Generally, you don’t pick either the edges of the battlespace or the center of the battlespace, but something “in between.”



Results

Conditions				Results		
Threat	Altitude	Maneuver	Impact Angle	Crossrange	Downrange	Miss Distance
Area	20k	No	Shallow	32.9	64.8	72.7
Area	20k	No	Steep	-28.2	71.9	77.2
Area	20k	No	Steep	-19.0	71.7	74.2
Area	20k	Yes	Shallow	7.1	-39.1	39.7
Area	30k	No	Shallow	0.1	0.0	0.1
Area	30k	Yes	Steep	-4.6	3.4	5.7
Area	30k	Yes	Steep	-13.1	-6.4	14.6
Point	20k	No	Shallow	0.8	30.6	30.6
Point	20k	Yes	Steep	-10.4	-32.7	34.3
Point	30k	No	Steep	5.3	17.2	18.0
Point	30k	Yes	Shallow	-20.5	-27.6	34.4

Descriptive Statistics: Miss Distance

```
Variable         Mean  StDev  Minimum  Median  Maximum
Miss Distance   36.50  27.46     0.10   34.30    77.20
```

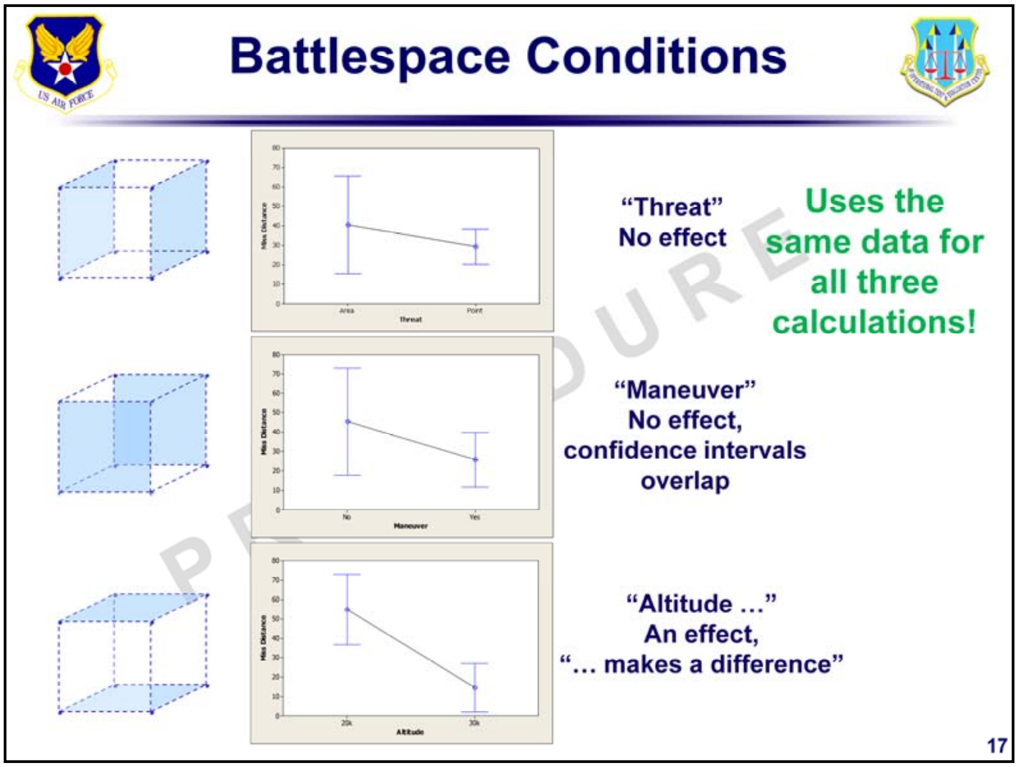
16

These are the results from the JDAM test (they've been "scaled" to keep the results unclassified). There are actually 11 points, not just eight, because after the initial eight were dropped, we did three additional confirmatory drops because of the results of the analysis. We'll point out what those were.

A typical approach to analyzing this would be to figure out the mean or median and draw a circle around the points with that radius.

A word about averages or means—often they are meaningless; "One size fits all, but doesn't fit anybody." Nobody is "average." This is the drawback of many of the other techniques if they're applied to the wrong tactic. DOE is a very robust technique, however. Because of the structured, factorial design, we can do much more than just compute the grand average.

You can see a hint of what the analysis showed by looking at the three impacts well outside the circle—is there anything which is causing those extreme misses? A factorial design let's us analyze the data and determine if there is something causing this or if they're just the random misses.



Battlespace Conditions

“Threat”
No effect

Uses the same data for all three calculations!

“Maneuver”
No effect,
confidence intervals overlap

“Altitude …”
An effect,
“… makes a difference”

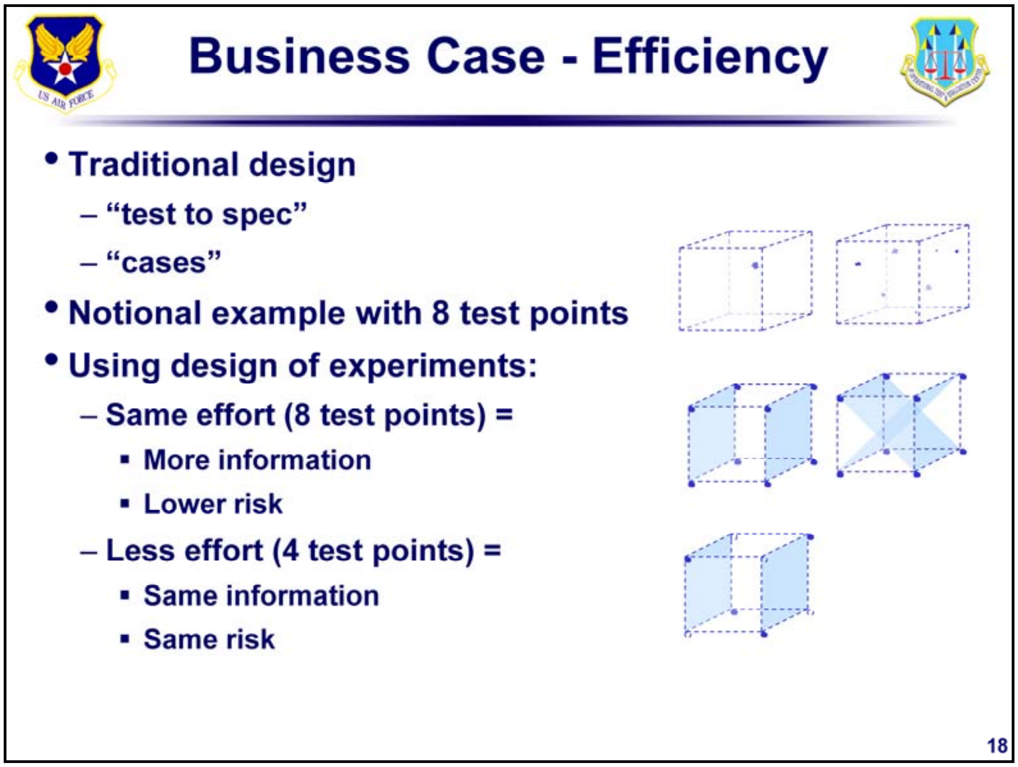
17

Since we structured this test using a factorial design, we can go beyond just reporting a grand average or mean. We can look at individual factors and determine which ones, if any, influence the outcome. For the factor “Threat,” there isn’t much of a difference between Area or Point and the confidence interval for Area actually subsumes the Point confidence interval—we determine that the type of threat doesn’t make a difference. The factor “Maneuver” hints that there might be an effect, but the confidence intervals overlap so much we determine that maneuvering doesn’t make a difference. Although not shown here, we also determined that “Impact Angle” did not make a difference—contrary to what the intuitive belief was. However, the factor “Altitude” appears to make a big difference AND the confidence intervals don’t overlap at all. So altitude has an effect.

FAQs

-These charts give a visual indication of whether there is an effect or not. There are statistical tests to determine if there is an “effect” or not and what the decision risk or confidence is.

- Sample size is a consideration for factorial designs—they use the confidence interval to determine if there is an effect. If the sample size is too small, then the confidence intervals are very wide and you will usually conclude there is no effect; unless the effect is very, very large and “overcomes” the wide confidence intervals.



Business Case - Efficiency

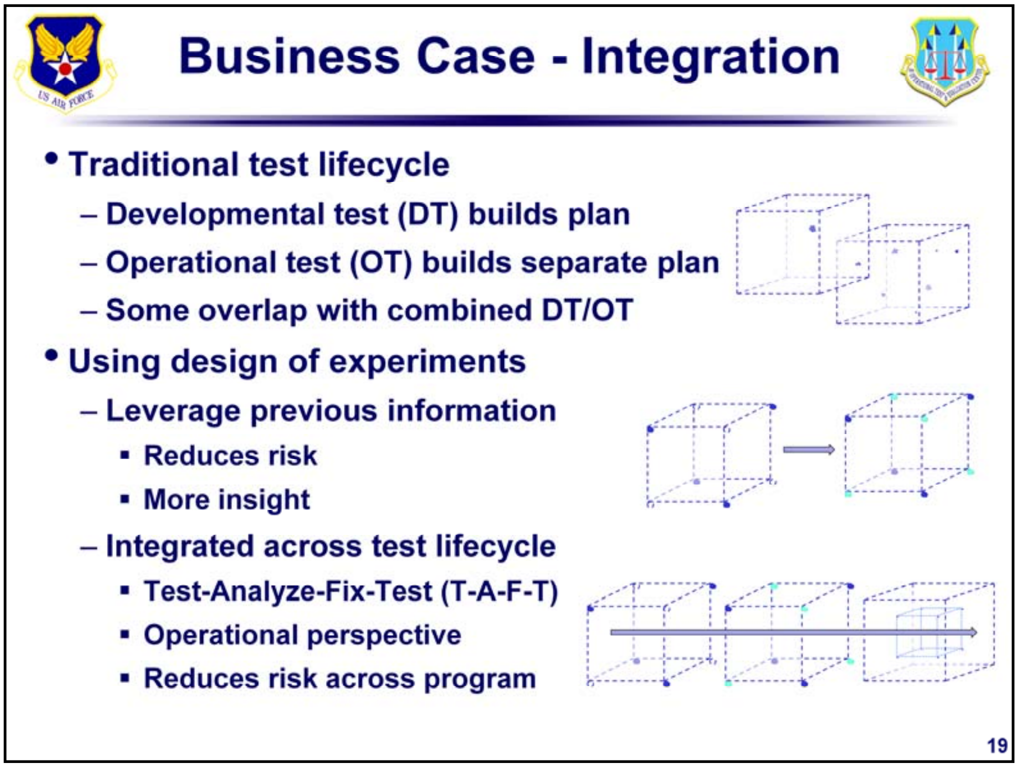
- Traditional design
  - "test to spec"
  - "cases"
- Notional example with 8 test points
- Using design of experiments:
  - Same effort (8 test points) =
    - More information
    - Lower risk
  - Less effort (4 test points) =
    - Same information
    - Same risk

18

DOE offers significant gains in efficiency over the traditional "test to spec" or "cases" approaches. Using a notional example with a test having eight test points (e.g. eight bombs, eight missiles, eight images, etc.). Theses efficiencies can be either more information for the same effort or the same information for less effort.

Using eight test points, you can generally garner more information from them using a DOE based design than the traditional designs. This information is the effects due to varying conditions. You also maintain the confidence in the results you may have had with the traditional design. Together, this translates to lower risk.

Alternatively, you can generally garner the same information as a traditional design, but with fewer test points (not necessarily half, as shown here). This also maintains the confidence and the risk associated with the traditional designs.



Business Case - Integration

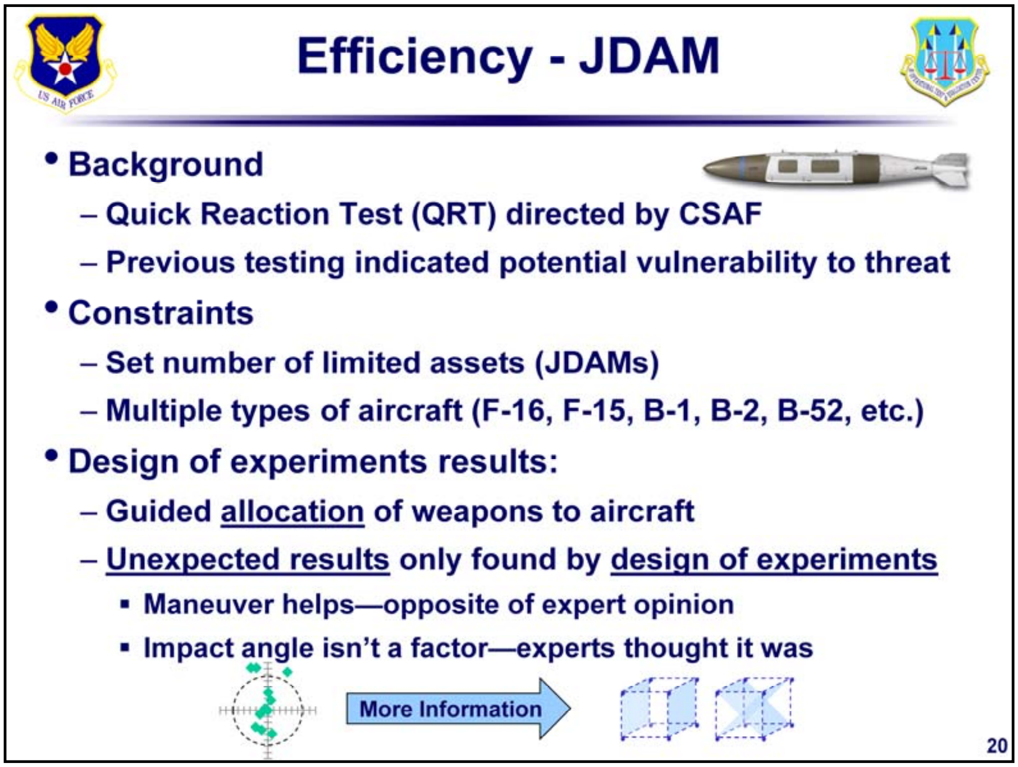
- Traditional test lifecycle
  - – Developmental test (DT) builds plan
  - – Operational test (OT) builds separate plan
  - – Some overlap with combined DT/OT
- Using design of experiments
  - – Leverage previous information
    - Reduces risk
    - More insight
  - – Integrated across test lifecycle
    - Test-Analyze-Fix-Test (T-A-F-T)
    - Operational perspective
    - Reduces risk across program

19

DOE can also provide an avenue to integrate CT, DT, and OT across the test life cycle. A traditional approach to integrating DT and OT is for each to build their (large) test plan independently and then look for opportunities where the plans overlap. This area is combined DT/OT. Often, it involves identifying which DT test points used an operationally representative asset in an operationally representative environment. This DT data can then be "qualified" for use in the OT plan, analysis, and report. Note this can result in a slight reduction in total number of assets.

DOE can further this integration in a couple of ways. One approach is to leverage the previous, DT information to build a smaller operational test. At the simplest, this may mean supplementing or augmenting test points from DT with OT test points. The savings is in not repeating test points DT has already accomplished. Other techniques is to use DT information to focus where OT should (and should not) test. A fuller integration would have OT influence DT events (using DOE) to reach a more complete DOE based design. This would provide more insight to the performance and help reduce risk (as well as cost and schedule).

Extending this last thought further, the whole process could be based on a sequential Test-Analyze-Fix-Test process. Although not highlighted, a fundamental tenant of DOE is this sequential testing. Combining this with operational perspectives of the conditions a system must operate in, leads to reducing risk across the program and test lifecycle. It also has the potential to reduce cost and schedule.

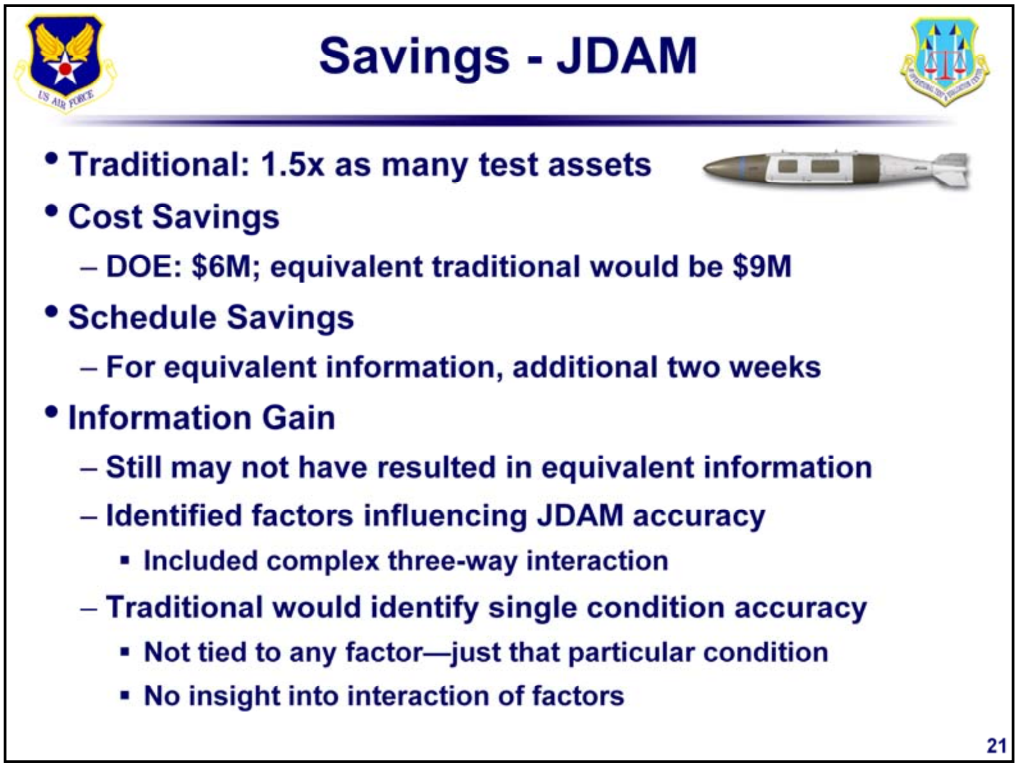


Efficiency - JDAM

- Background
  - Quick Reaction Test (QRT) directed by CSAF
  - Previous testing indicated potential vulnerability to threat
- Constraints
  - Set number of limited assets (JDAMs)
  - Multiple types of aircraft (F-16, F-15, B-1, B-2, B-52, etc.)
- Design of experiments results:
  - Guided <u>allocation</u> of weapons to aircraft
  - <u>Unexpected results</u> only found by <u>design of experiments</u>
    - Maneuver helps—opposite of expert opinion
    - Impact angle isn't a factor—experts thought it was

20

JDAM – The previous examples were base on the JDAM Quick Reaction test. Previous testing had shown a possible vulnerability to certain threats. A test was planned and executed in about two months using the principles from design of experiments. The test involved about a dozen different aircraft; consequently the number of JDAMs allocated to each aircraft was limited. This resulted in a complex allocation of JDAMs to each—design of experiments guided this allocation rather than some "arbitrary" scheme. Unexpected results (opposite of what expert opinion thought would happen) were found; only design of experiments would find these results. This lead to modified TTPs for employment as well as engineering improvements to the JDAM and aircraft.



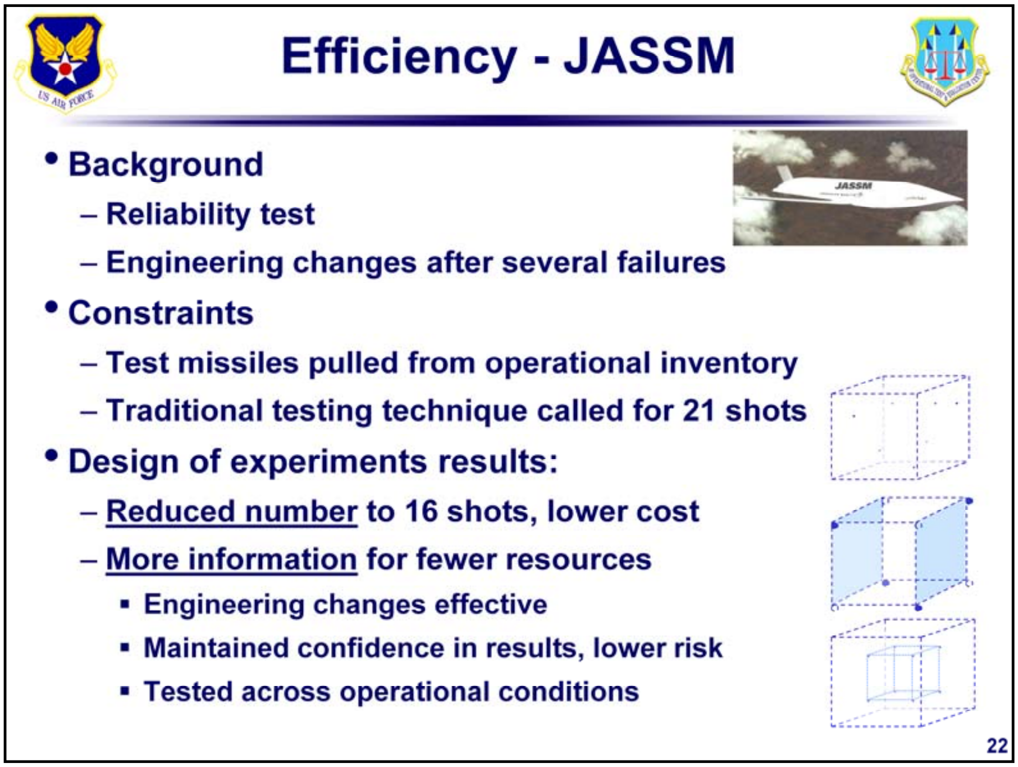
Savings - JDAM

- Traditional: 1.5x as many test assets
- Cost Savings
  - DOE: $6M; equivalent traditional would be $9M
- Schedule Savings
  - For equivalent information, additional two weeks
- Information Gain
  - Still may not have resulted in equivalent information
  - Identified factors influencing JDAM accuracy
    - Included complex three-way interaction
  - Traditional would identify single condition accuracy
    - Not tied to any factor—just that particular condition
    - No insight into interaction of factors

Savings are not just AFOTEC. They are Air Force or “taxpayer” savings.

Using traditional test design techniques, such as test-to-spec or special cases, which was the original design, it would have taken about 1.5x as many munitions to learn the equivalent information. This would translate roughly as a $3M increase over the $6M cost using DOE-based designs. The schedule would increase by two weeks to the two-month test.

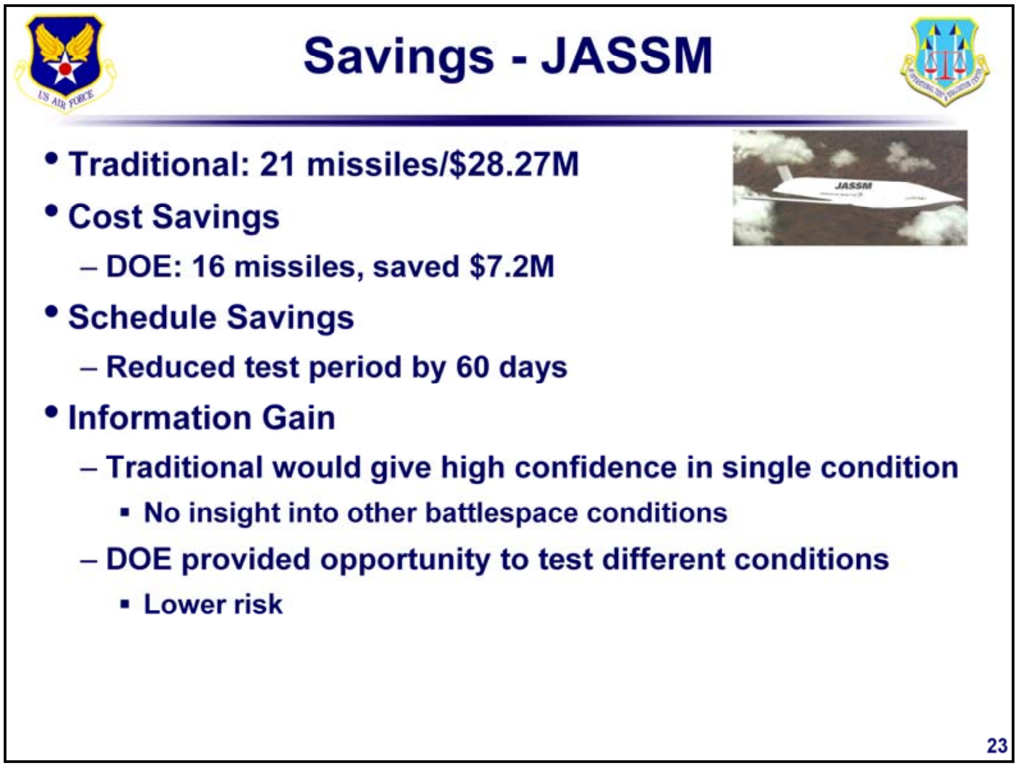
There is a risk that even this expanded test would not result in equivalent information. The DOE-based test designs allowed us to identify factors (e.g. altitude) affecting the accuracy of JDAMS. Additionally, it allowed identification of complex interactions between multiple factors. Traditional techniques MIGHT have found the less accurate conditions, but even still would blend or confound all the factors affecting performance.



Efficiency - JASSM

- Background
  - Reliability test
  - Engineering changes after several failures
- Constraints
  - Test missiles pulled from operational inventory
  - Traditional testing technique called for 21 shots
- Design of experiments results:
  - Reduced number to 16 shots, lower cost
  - More information for fewer resources
    - Engineering changes effective
    - Maintained confidence in results, lower risk
    - Tested across operational conditions

JASSM – During an exercise, several JASSMs had reliability failures. Missiles were modified to remove the suspected failure cause. The traditional "test to spec" approach called for 21 missiles. This was about 25% of the total operational inventory of this version of JASSM. Using design of experiments, this was reduced to 16 shots in a structured design (vice the proposed, arbitrary selection of test conditions). A rough order of magnitude (ROM) of savings was about $4 million between the cost of the JASSM and the costs to test (range, instrumentation, personnel, etc.). Not only were the number of assets reduced providing the same level of statistical confidence, but the operational conditions that could affect the reliability was characterized. More information for fewer resources.



Savings - JASSM

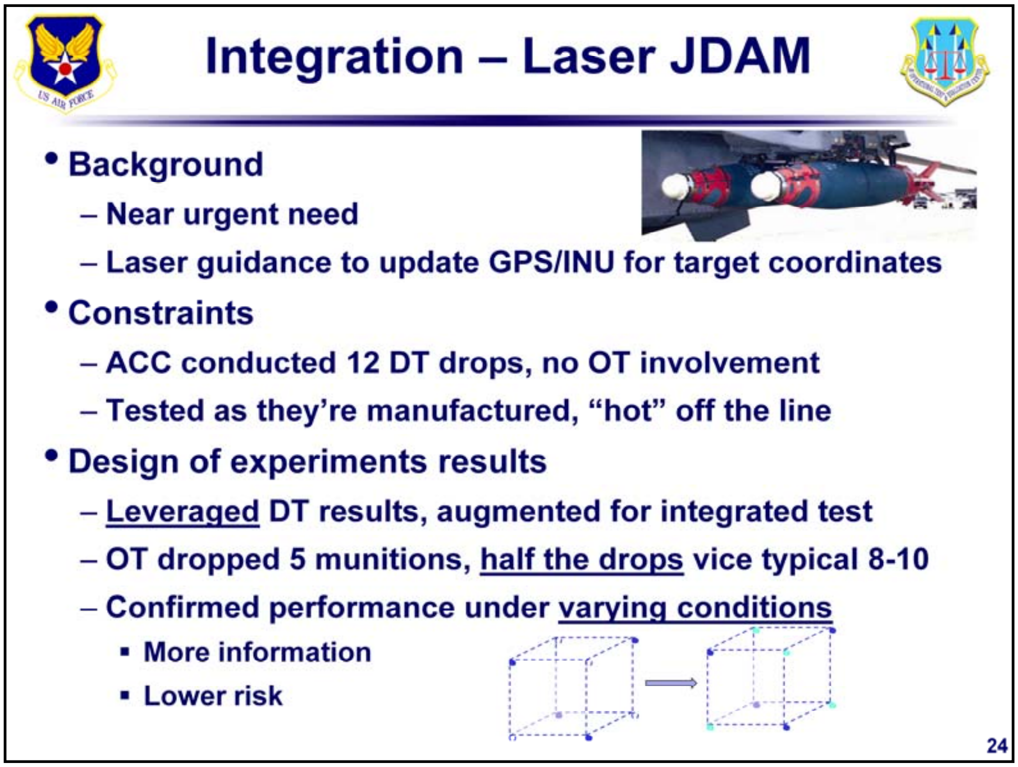
- Traditional: 21 missiles/$28.27M
- Cost Savings
  - DOE: 16 missiles, saved $7.2M
- Schedule Savings
  - Reduced test period by 60 days
- Information Gain
  - Traditional would give high confidence in single condition
    - No insight into other battlespace conditions
  - DOE provided opportunity to test different conditions
    - Lower risk

Savings are not just AFOTEC. They are Air Force or “taxpayer” savings.

Final OA costs for 16 missiles:

Range costs:	$ 1.0M
Target costs:	$ 1.2M
JASSM costs:	$ 14.1M
Telemetry kit costs:	$ 5.2M
General support:	$ 42k
Total	$ 21.54M

Reduction from 21 to 16 missiles saved $7.2M



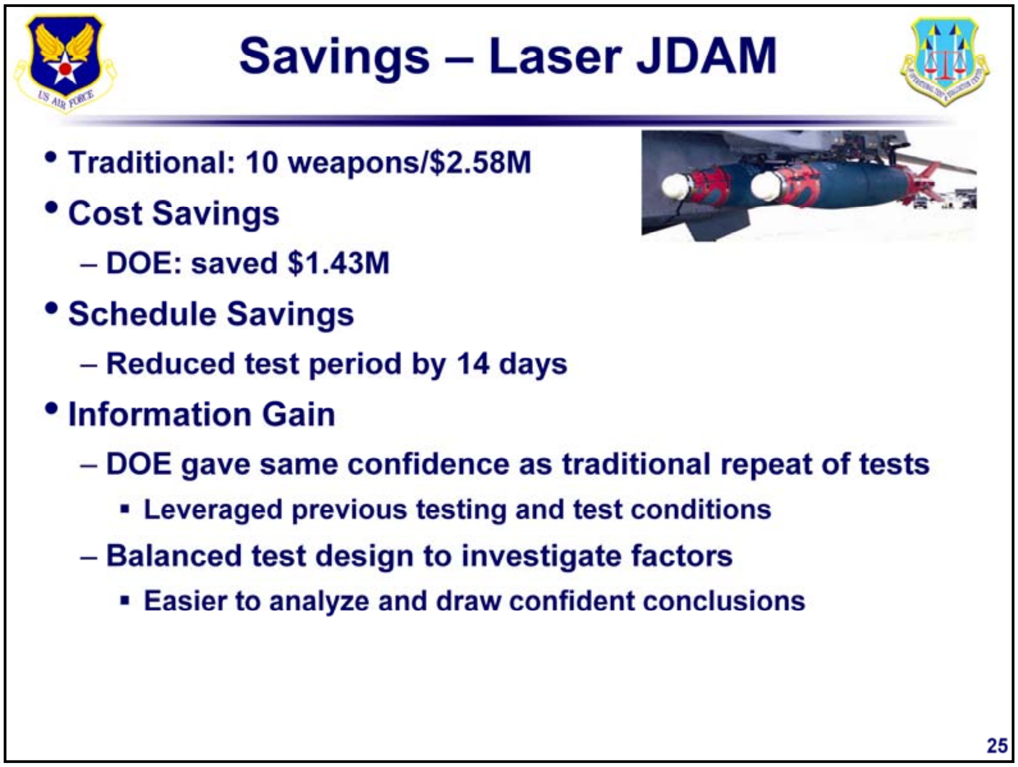
Integration – Laser JDAM

- Background
  - Near urgent need
  - Laser guidance to update GPS/INU for target coordinates
- Constraints
  - ACC conducted 12 DT drops, no OT involvement
  - Tested as they're manufactured, "hot" off the line
- Design of experiments results
  - Leveraged DT results, augmented for integrated test
  - OT dropped 5 munitions, half the drops vice typical 8-10
  - Confirmed performance under varying conditions
    - More information
    - Lower risk

24

Laser JDAM – AFOTEC was asked, on short notice, to conduct an operational test on the Laser JDAM. ACC had already dropped 12 munitions. While not done with design of experiments, it did cover a variety of conditions. We were able to augment or complement some of the original shots with 4 additional drops (a fifth drop was a demo). This confirmed the performance of the Laser JDAM and reduced the number of munitions to half of what would typically been used. No need to drop the "necessary" eight by leveraging the info from DT and using four of their shots to form a complete design.

This is an example of using DOE to augment previous testing where there was little or no influence on previous testing, but the events were done under a variety of conditions. This could have been a more efficient test if the original 12 had been integrated and done based on design of experiments.



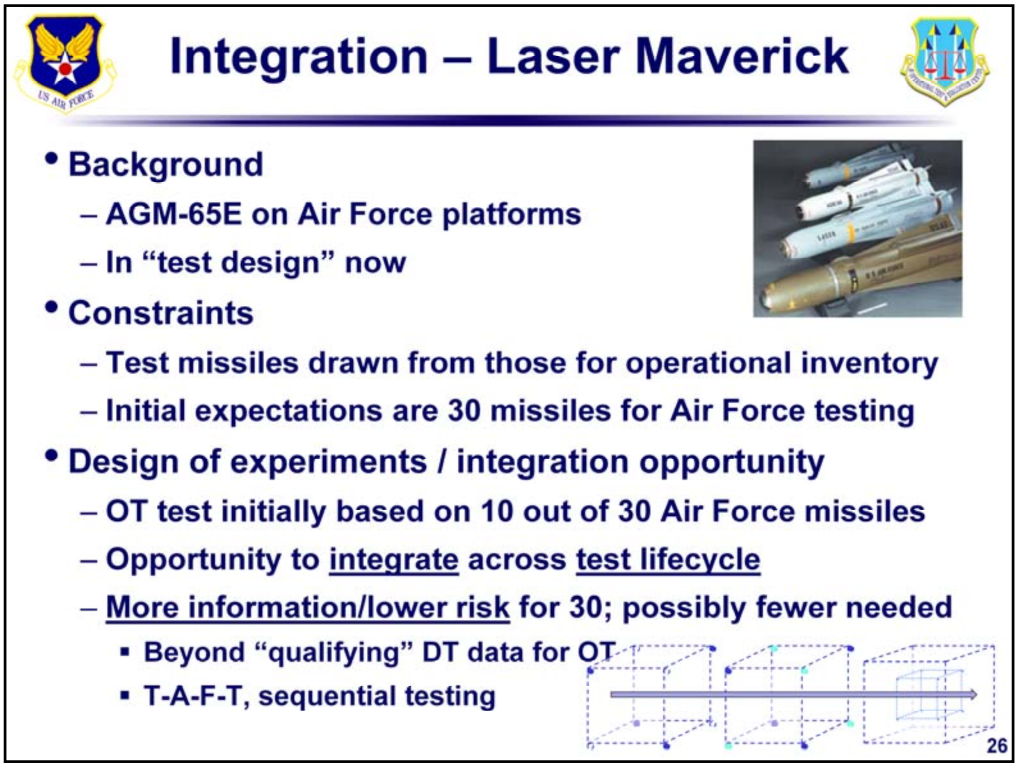
Savings – Laser JDAM

- Traditional: 10 weapons/$2.58M
- Cost Savings
  - DOE: saved $1.43M
- Schedule Savings
  - Reduced test period by 14 days
- Information Gain
  - DOE gave same confidence as traditional repeat of tests
    - Leveraged previous testing and test conditions
  - Balanced test design to investigate factors
    - Easier to analyze and draw confident conclusions

25

Savings are not just AFOTEC. They are Air Force or “taxpayer” savings.

	Traditional	Integrated
Range costs:	$ 400k	$ 200k
Target costs:	$ 1.5M	$ 630k
Weapon costs: $ 450k	$ 225k	
Telemetry kit costs:	$ 189k	$ 81k
General support:	$ 39k	$ 13k
Total	$ 2.58M	$ 1.43M



Integration – Laser Maverick

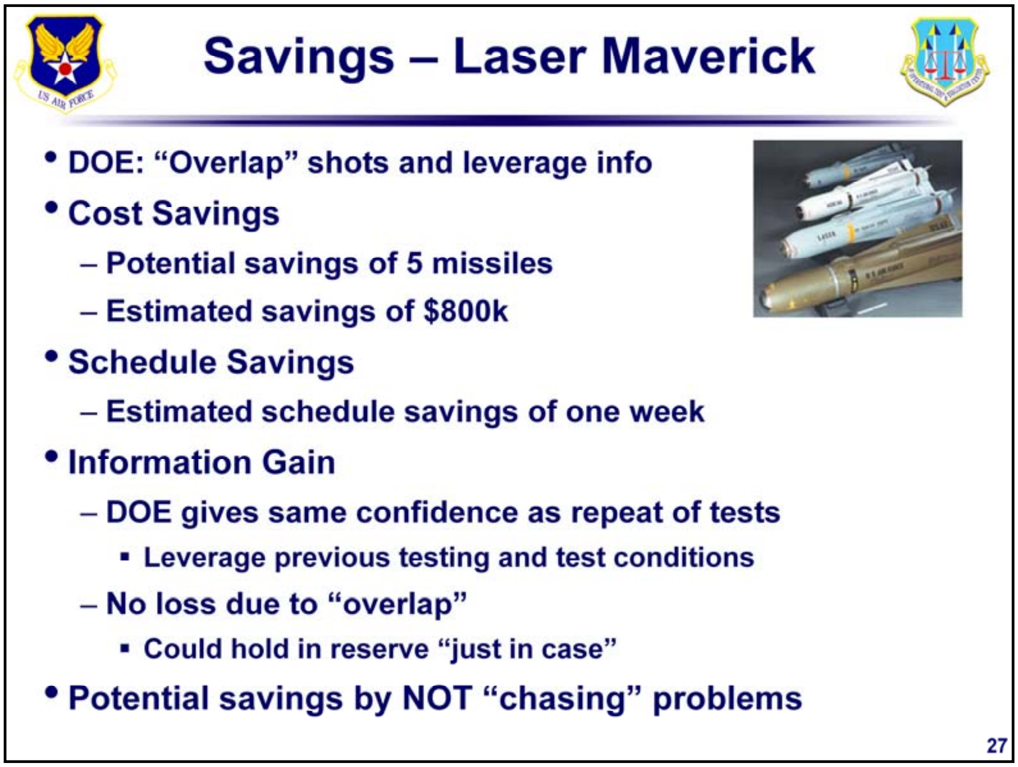
- Background
  - AGM-65E on Air Force platforms
  - In “test design” now
- Constraints
  - Test missiles drawn from those for operational inventory
  - Initial expectations are 30 missiles for Air Force testing
- Design of experiments / integration opportunity
  - OT test initially based on 10 out of 30 Air Force missiles
  - Opportunity to integrate across test lifecycle
  - More information/lower risk for 30; possibly fewer needed
    - Beyond “qualifying” DT data for OT
    - T-A-F-T, sequential testing

26

Laser Maverick – AFOTEC has just completed an initial test design for the Laser Maverick. Expectations are about 30 munitions for Air Force DT and OT; OT would expend approximately 10 of those munitions plus numerous captive carries. Laser Maverick provides an excellent opportunity for integrated testing AND application of design of experiments. While there is a possibility of reducing the number of munitions, the benefit may be in a more efficient test (fewer repeat or regression shots, schedule flexibility,) with more information (e.g. lower risk) for the same 30 munitions.

This goes beyond the current approach of “qualifying” DT data for OT purposes. This does not mean one big design that DT does part of and OT does part of; that is the same as current combined DT/OT. It means using more of a T-A-F-T process with a sequence of progressive test events learning from the preceding tests. A key part of executing this, however, will be to avoid the “special” cases and “pet rock” conditions. Too many of these, and the benefits of both integration and design of experiments is lost.

This is an example of an integrated use of design of experiments where the entire test lifecycle is managed.



Savings – Laser Maverick

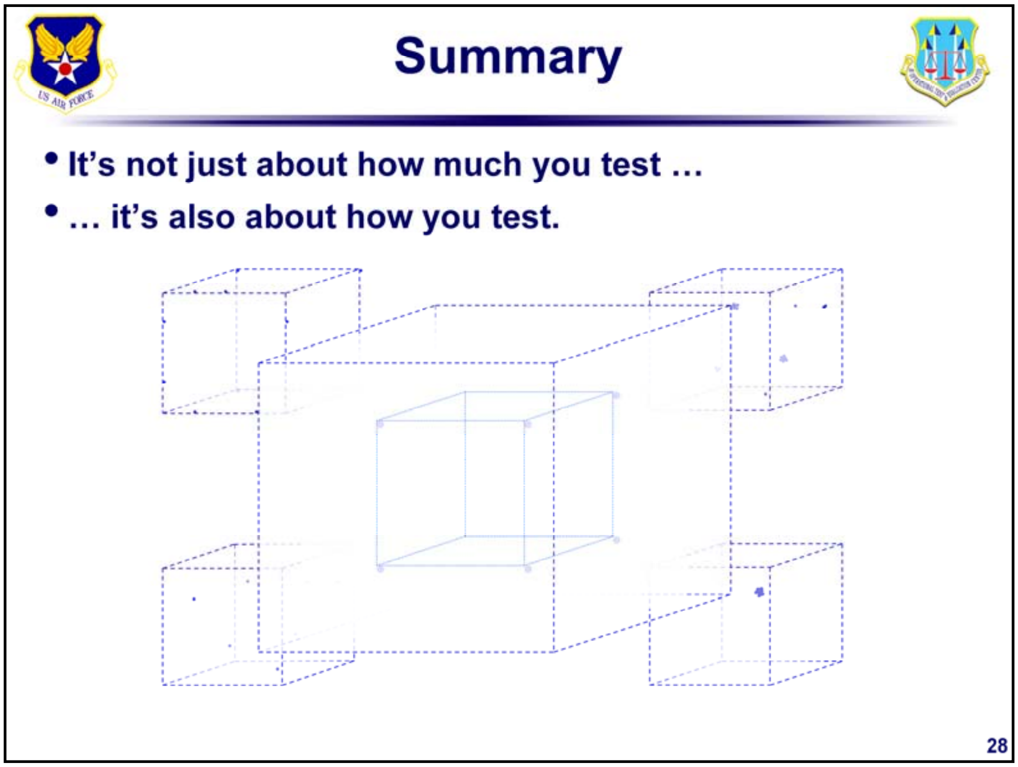
- DOE: “Overlap” shots and leverage info
- Cost Savings
  - Potential savings of 5 missiles
  - Estimated savings of $800k
- Schedule Savings
  - Estimated schedule savings of one week
- Information Gain
  - DOE gives same confidence as repeat of tests
    - Leverage previous testing and test conditions
  - No loss due to “overlap”
    - Could hold in reserve “just in case”
- Potential savings by NOT “chasing” problems

27

Savings are not just AFOTEC. They are Air Force or “taxpayer” savings.

	Unit Cost	For 5 Missiles
AGM-65E test article:	$ 110k	$550k
Targets:	$ 10k	$ 50k
Range:	$ 75k	$150k
F-16 support:	$ 5k	$ 25k
General support:	$ 10k	$ 25k
Total		$800k

There are potentially unquantified savings by NOT chasing problems. DOE allows you to focus in on the factors or interactions causing poor performance. Traditional methods just tell you (at best), if you are having performance problems, but provide little insight into what to fix. Consequently, there is much speculation and “chasing” possible causes. DOE can help reduce this “chasing.”



Summary

- It’s not just about how much you test …
- … it’s also about how you test.

28