INSTITUTE FOR DEFENSE ANALYSES

Empirical Signal-to-Noise Ratios from Operational Test Data

Matt R. Avery

August 2014

IDA Document NS D-5243
Log: H 14-000816

INSTITUTE FOR DEFENSE ANALYSES
4850 Mark Center Drive
Alexandria, Virginia 22311-1882

About This Publication

Statistical power is a common metric for assessing experimental designs. While this metric depends on many factors, one of the most critical is the expected effect size of relevant factors and the relative noise expected in the data. Together, these values are summarized as the signal-to-noise ratio (SNR). Software packages like JMP 10 and Design Expert use SNR as a critical component in power calculations, and by general "rule of thumb," values such as 0.5, 1, and 2 are used. However, it is not clear that these values represent the true spectrum of likely outcomes from operational test data. Operational testing is the final phase prior to fielding in the DOD acquisition process for new systems. Because of the operational realism strived for in such testing, there are often many sources of uncontrolled variation, making it difficult to plan an appropriate test based on the SNR. In this briefing, we summarize observed SNRs from a wide spectrum of operational tests and offer suggestions for the use of SNR in operational test design.

INSTITUTE FOR DEFENSE ANALYSES

IDA Document NS D-5243

Empirical Signal-to-Noise Ratios from Operational Test Data

Matt R. Avery

Outline

- What is Operational Testing?
- Using signal-to-noise ratios for operational test planning
- Signal-to-noise ratios for binary responses
- Summary of results
- Recommendations
- Next steps

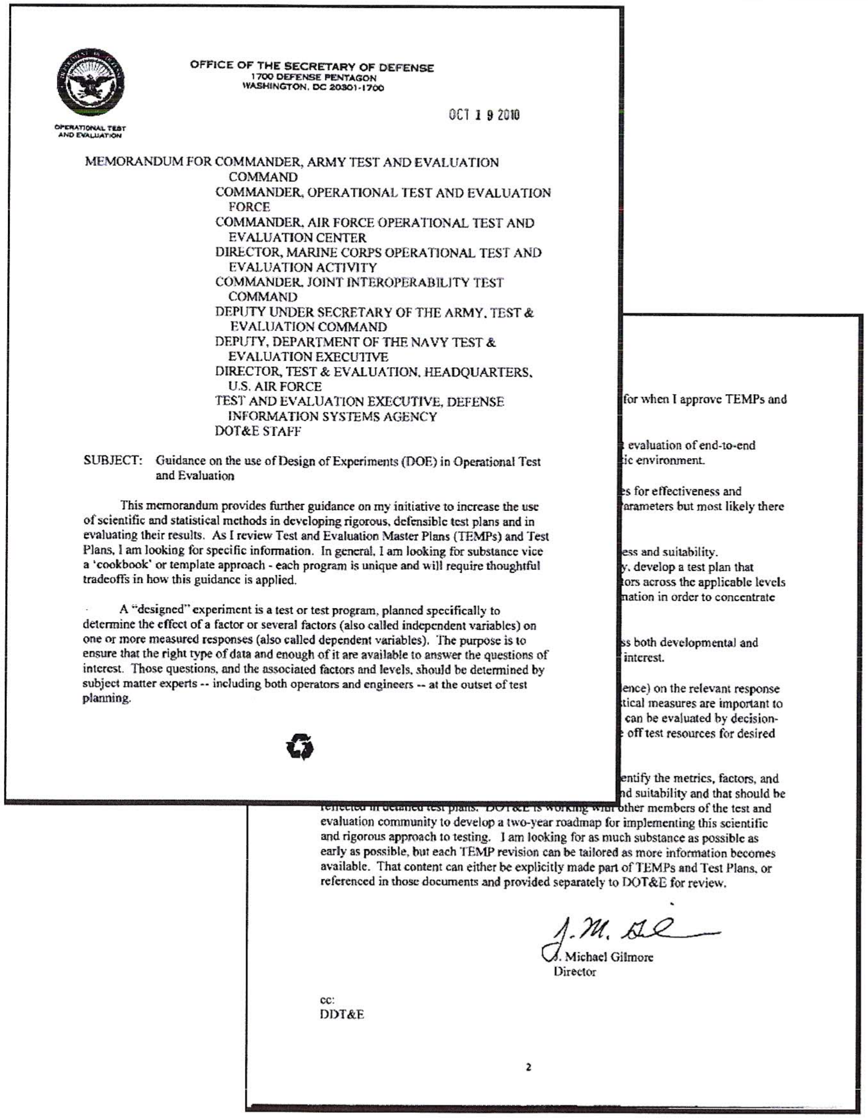
IDA

Operational Testing

- **Operational Testing plays a key role in the DoD acquisitions process**
- **Overseen by Director, Operational Tests and & Evaluations (DOT&E)**
- **Goals of Operational Testing:**
  - Determine whether the system is operationally effective and suitable
  - Demonstrate system capability in operational context
- **Careful planning is crucial for a good operational test**
  - Long time horizon
  - Resource constrained

DOT&E Guidance

Dr. Gilmore's October 19, 2010 Memo to OTAs



OFFICE OF THE SECRETARY OF DEFENSE
1700 DEFENSE PENTAGON
WASHINGTON, DC 20301-1700

OPERATIONAL TEST AND EVALUATION

OCT 1 9 2010

MEMORANDUM FOR COMMANDER, ARMY TEST AND EVALUATION COMMAND
COMMANDER, OPERATIONAL TEST AND EVALUATION FORCE
COMMANDER, AIR FORCE OPERATIONAL TEST AND EVALUATION CENTER
DIRECTOR, MARINE CORPS OPERATIONAL TEST AND EVALUATION ACTIVITY
COMMANDER, JOINT INTEROPERABILITY TEST COMMAND
DEPUTY UNDER SECRETARY OF THE ARMY, TEST & EVALUATION COMMAND
DEPUTY, DEPARTMENT OF THE NAVY TEST & EVALUATION EXECUTIVE
DIRECTOR, TEST & EVALUATION, HEADQUARTERS, U.S. AIR FORCE
TEST AND EVALUATION EXECUTIVE, DEFENSE INFORMATION SYSTEMS AGENCY
DOT&E STAFF

SUBJECT: Guidance on the use of Design of Experiments (DOE) in Operational Test and Evaluation

This memorandum provides further guidance on my initiative to increase the use of scientific and statistical methods in developing rigorous, defensible test plans and in evaluating their results. As I review Test and Evaluation Master Plans (TEMPs) and Test Plans, I am looking for specific information. In general, I am looking for substance vice a 'cookbook' or template approach - each program is unique and will require thoughtful tradeoffs in how this guidance is applied.

A "designed" experiment is a test or test program, planned specifically to determine the effect of a factor or several factors (also called independent variables) on one or more measured responses (also called dependent variables). The purpose is to ensure that the right type of data and enough of it are available to answer the questions of interest. Those questions, and the associated factors and levels, should be determined by subject matter experts -- including both operators and engineers -- at the outset of test planning.

evaluation community to develop a two-year roadmap for implementing this scientific and rigorous approach to testing. I am looking for as much substance as possible as early as possible, but each TEMP revision can be tailored as more information becomes available. That content can either be explicitly made part of TEMPs and Test Plans, or referenced in those documents and provided separately to DOT&E for review.

J. Michael Gilmore
Director

cc:
DDT&E

2

- **The goal of the experiment**. This should reflect evaluation of end-to-end mission effectiveness in an operationally realistic environment.
- Quantitative mission-oriented **response variables** for effectiveness and suitability. (These could be Key Performance Parameters but most likely there will be others.)
- **Factors** that affect those measures of effectiveness and suitability. Systematically, in a rigorous and structured way, develop a test plan that provides good breadth of coverage of those factors across the applicable levels of the factors, taking into account known information in order to concentrate on the factors of most interest.
- **A method for strategically varying factors** across both developmental and operational testing with respect to responses of interest.
- **Statistical measures of merit** (power and confidence) on the relevant response variables for which it makes sense. These statistical measures are important to understanding "how much testing is enough?" and can be evaluated by decision makers on a quantitative basis so they can trade off test resources for desired confidence in results.

IDA

Signal-to-noise Ratios

- **DOT&E requires power analysis to justify test size/duration for all operational tests**
  - JMP and Design Expert are common tools
    - » Both require Signal-to-Noise Ratio (SNR) as an input

- **Signal: Change in response per change in a factor's level**

- **Noise: Root Mean Square Error (RMSE)**

Design

Run	Continuous	2-level	3-level
1	1	A	C
2	-1	A	D
3	-1	B	E
4	1	A	E
5	1	B	D
6	-1	A	D
7	-1	A	C
8	1	B	D
9	-1	B	E
10	1	A	E
11	0	B	C
12	0	B	C

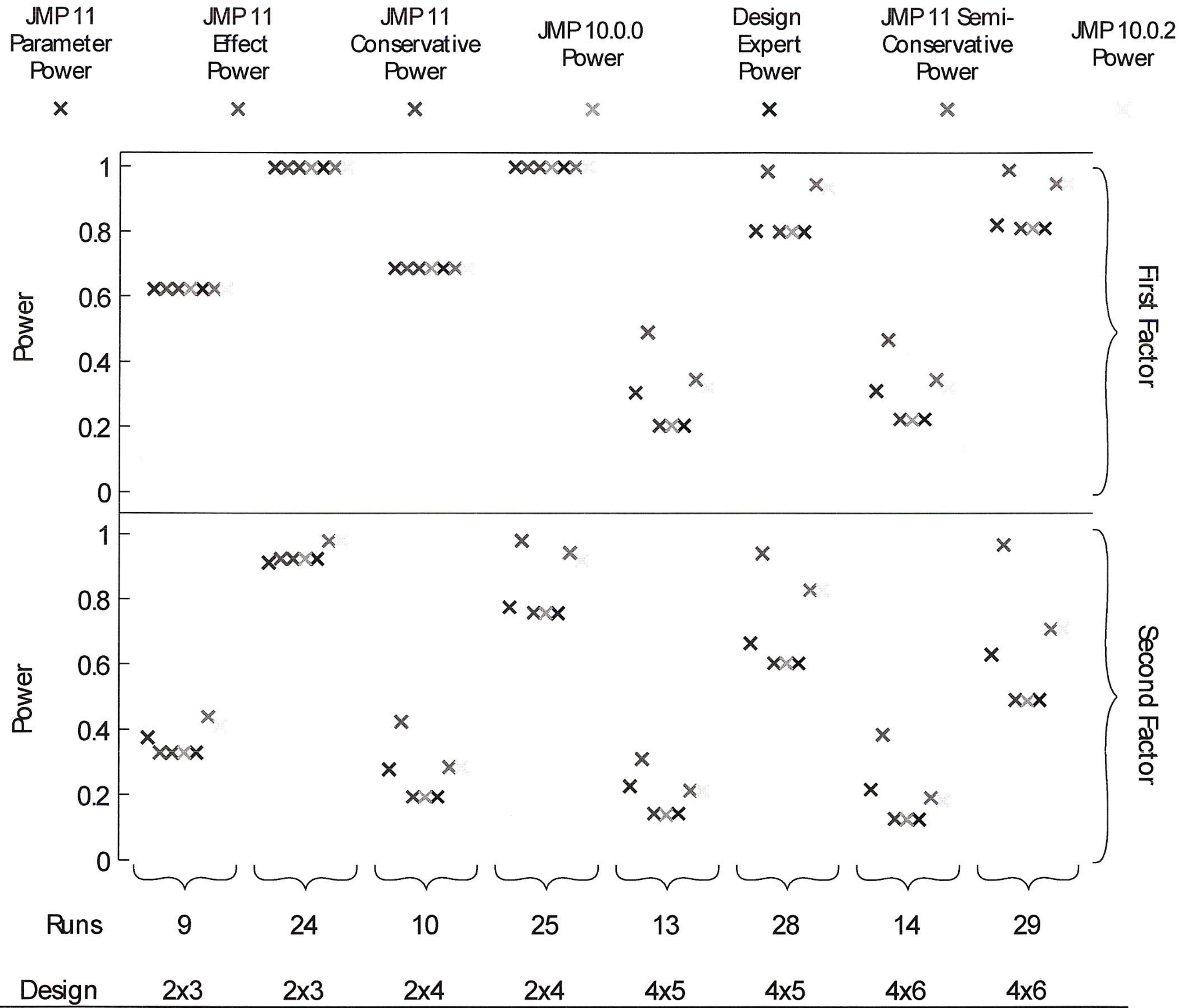
Power Analysis

Significance Level	0.05
Signal to Noise Ratio	2
Error Degrees of Freedom	7

Effect	Power Lower Bound	Numerator DF
Continuous	0.774	1
2-level	0.842	1
3-level	0.643	2

Variance Inflation Factors

Aside: Power calculations can vary dramatically by software package and version



- Different assumptions
- Different coding
- Categorical factors particularly impacted

Chart courtesy of Dr. Tom Johnson (IDA) and Dr. Jim Simpson (UA Huntsville)

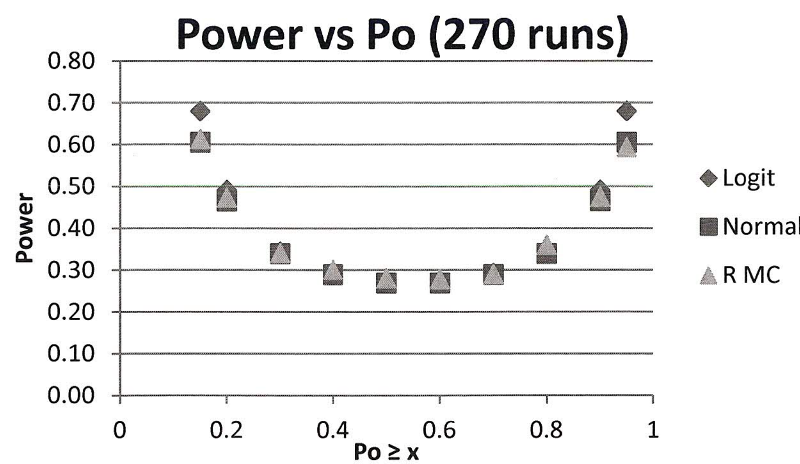
IDA

Power for binary responses

- **For some DOD systems, binary response variables are unavoidable**
  - Message completion rate
  - Torpedo hit/miss
- **SNR framework doesn't apply well to binary response variables**
  - Signal
    - » Based on change in p?
    - » Based on log odds ratio?
  - Noise depends on $\bar{p}$
  - No software solution available
- **Work-around allows use of software[1]**
  - Normal approximation conservative relative to logit method
  - Resulting power estimates close to what you'd get through simulation

Approximate SNR

P(bar)	0.8
Δ	0.2
P1	0.7
P2	0.9
δ	0.200
σ	0.400
SNR	**0.500**



[1]*Dealing with Categorical Data Types in a Designed Experiment Part II: Sizing a Designed Experiment When Using a Binary Response*, Dr. Francisco Ortiz, AFIT STAT T&E COE; www.AFIT.edu/STAT

Estimating Empirical SNRs

Goal: Determine what size effects are observed in real test data

Fitting the model

- Fit a plausible, fully estimable model
- All two-way interactions if possible
- Reduce model if necessary (estimability, degrees of freedom, model over-fit, etc.)
  - Note: Goal *is not* to fit optimal model

For continuous response variables:

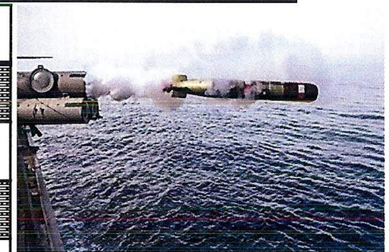
- Noise is RMSE
- Signal:
  - For categorical factor, the signal is β (R default 0-1 coding used)
  - For continuous factor, the signal is $\beta(\mu_{75} - \mu_{25})$
    - » μ_n is the nth percentile for that factor
    - » Many data sets have a few "extreme" data points

Estimating Empirical Δs

For categorical response variables:

- Using "workaround", all we need is to estimate Δ
- Begin by computing $\bar{p}$:
  - Literally estimated by taking average over all effects:
  - $\bar{p} = \beta_0 + \frac{1}{m}\sum\beta_i^*$, where m is the number of effects estimated, and $\beta^* = \frac{1}{\mathrm{m_i}}\sum\beta_j^i$
- Estimating Δ:
  - For categorical factor, the signal is $\mathrm{inverse\_logit}(\bar{\mathrm{p}} + \beta)$
  - For continuous factor, the signal is $\mathrm{inverse\_logit}(\bar{p} + \beta(\mu_{75} - \mu_{25}))$
    - » μ_q is the qth percentile for that factor

Summary of programs involved in this study

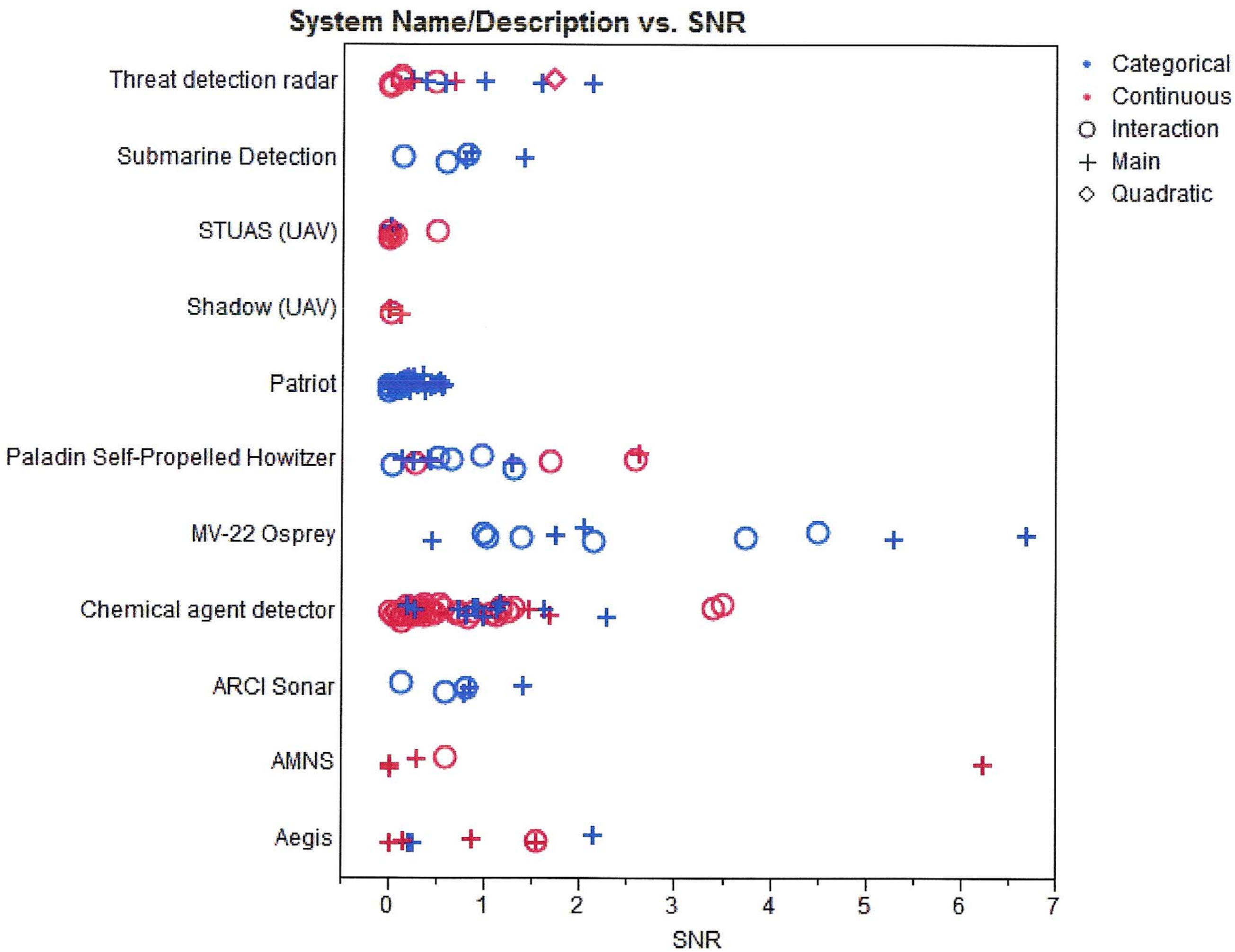




System	Response Variable	n	
Aegis	P(Raid Annihilation)	22	
Airborne Mine Neutralization System	Time to neutralize	33	
Virginia Class Submarine	Bearing Prediction Error	147	256
Chemical Agent Detector	Time to Detection	9,461	
LPD-17 (amphibious combat ship)	P(Impact)	296	
Mk54 CBASS Torpedo	P(Hit)	115	
Mk48 Torpedo	P(Hit)	35	
ARC-I Sonar	Difference in detection time	100	
Patriot	P(Intercept)	3,472	
RQ-21a Tactical UAV	Target Location Error	32	
Stryker Mobile Gun System	Correct Target Classification	464	
Global Broadcasting System	P(Successful Communication)	358	87
MV-22 Osprey	Mission Success Score	38	
Paladin Self-Propelled Howitzer	Miss Distance	71	
Shadow Tactical UAV	Target Location Error	285	



SNR for different program types

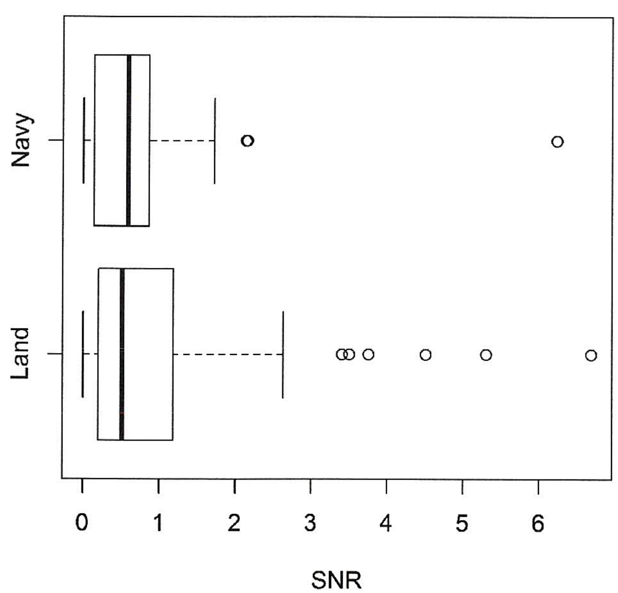


Summary Statistics for Empirical SNRs

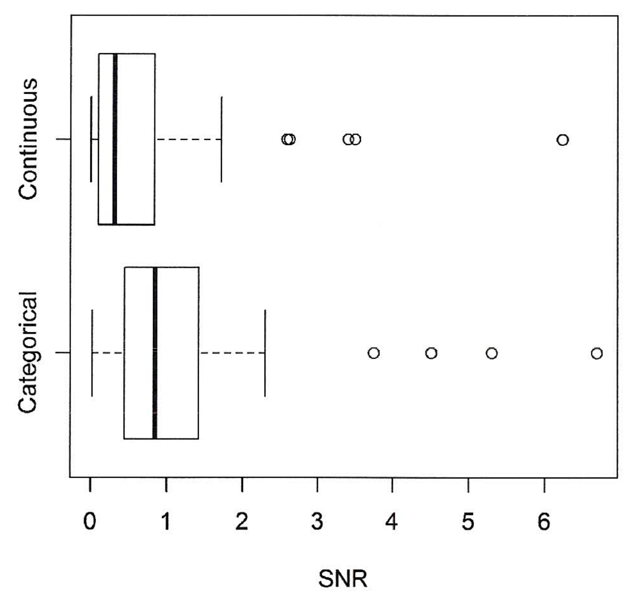
Mean	0.888
Median	0.534
75^{th} percentile	1.151
90^{th} percentile	2.026

- Over 90% of observed effects have $SNR < 2$
- Minimal variation across warfare group
- Categorical factors had higher SNR
  - » Possibly an artifact of estimation method

SNR for Land vs. Navy Programs

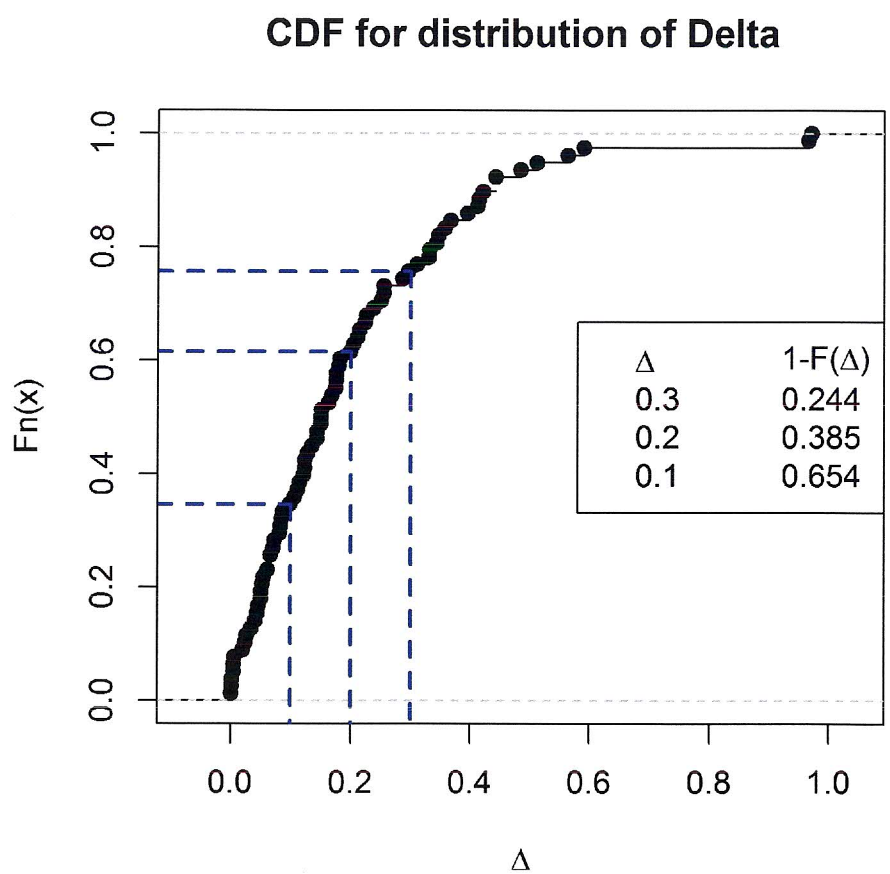


SNR by Parameter Type



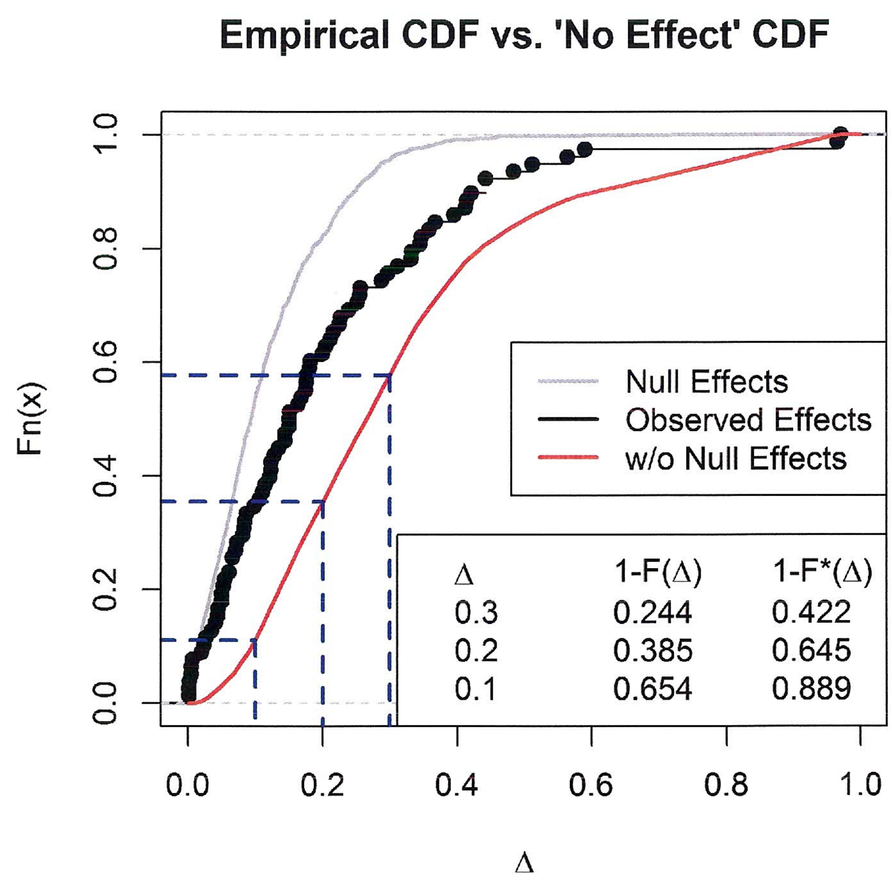
CDF for Categorical Responses

- **Some effects are very large**
  - Largest come from continuous factors observed over large ranges
- **Typical values for Δ when sizing tests: 0.3, 0.2, 0.1**
  - Median effect size: 0.151
- **Many effect sizes very close to 0**
  - Most (11/14) with $\Delta < 0.05$ are interactions
  - How many are just "noise"?



Comparison to Null Model

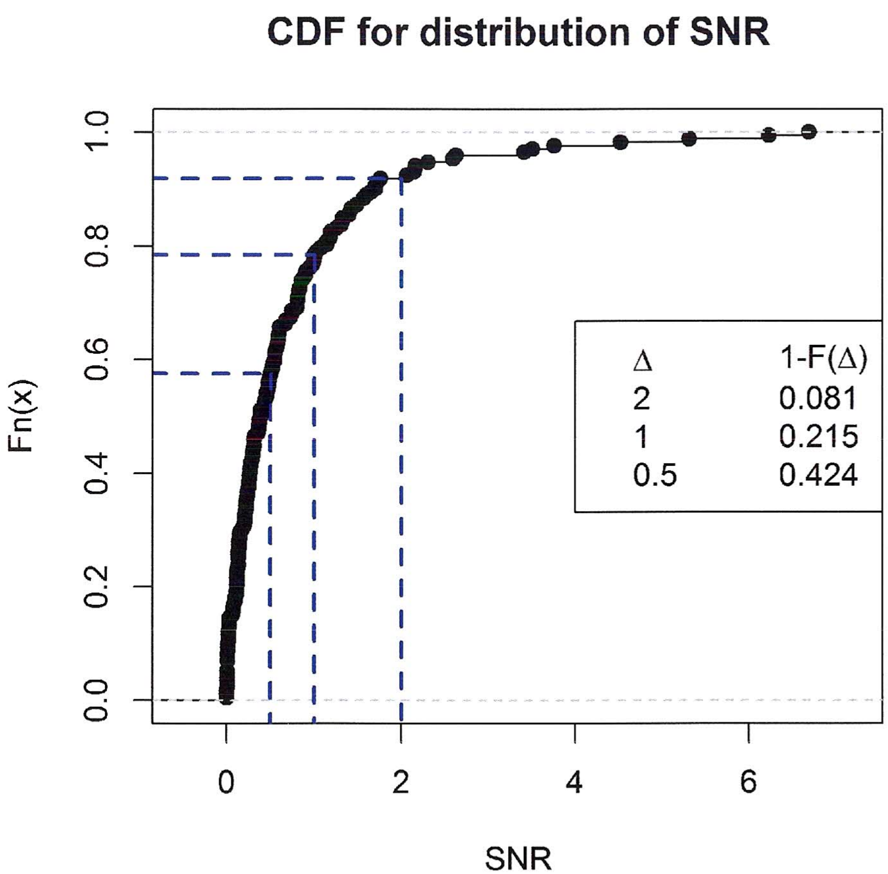
- **Gray curve: Simulated data where "null" model is true**
  - Most effects are small
  - Median=0.093
- **Subtracting "null" effects and normalizing yields red curve**
  - Distribution of true effects
  - Most are greater than 0.2
  - Nearly all greater than 0.1



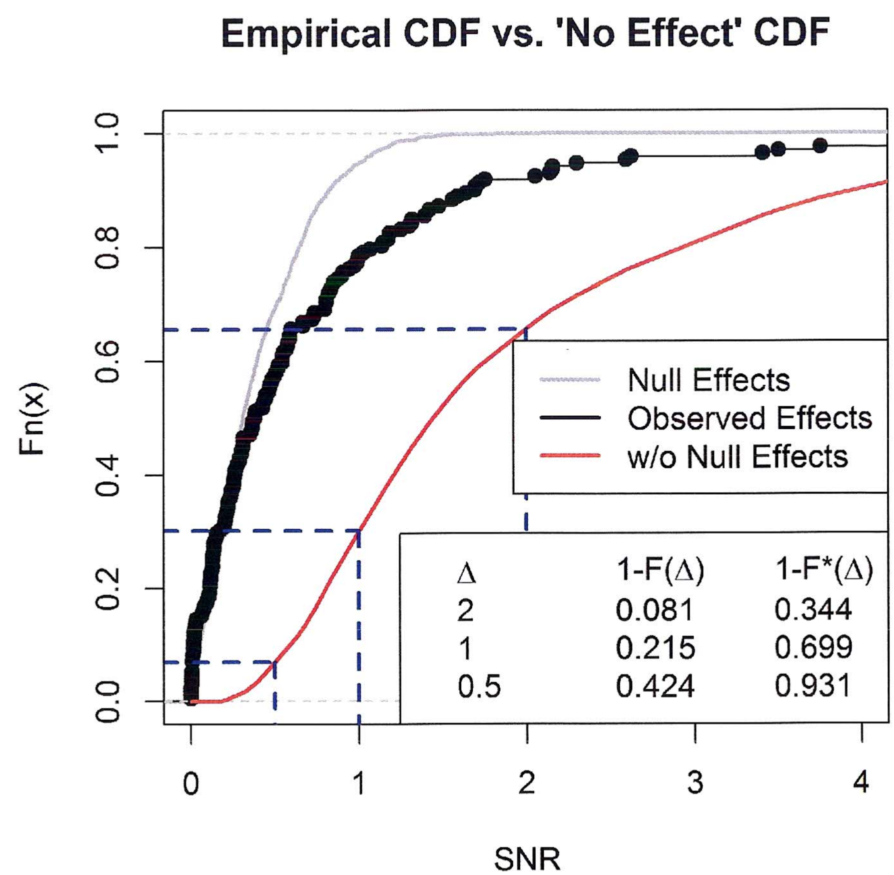
Future Work & Conclusions

- **Future Work**
  - Additional data sets can be added for additional breadth and depth
  - Assess accuracy of *a priori* estimates of SNR
    - » Are the values currently being used in test plans reflective of the SNRs observed once the tests have been conducted?
- **Major Conclusions**
  - After normalizing:
    - » **59%** of SNRs between **0.5** and **2**
    - » **46%** of Δs between **0.1** and **0.3**
- **Recommendations**
  - *Ceteris paribus*, use SNR no greater than 1.5 for power calculations
  - *Ceteris paribus*, use Δ no greater than 0.2 for power calculations

Backup



CDF for distribution of SNR
Fn(x)
SNR
0.0
0.2
0.4
0.6
0.8
1.0
0
2
4
6
Δ 1-F(Δ)
2 0.081
1 0.215
0.5 0.424



Empirical CDF vs. 'No Effect' CDF
Fn(x)
SNR
0.0
0.2
0.4
0.6
0.8
1.0
0
1
2
3
4
Null Effects
Observed Effects
w/o Null Effects
Δ 1-F(Δ) 1-F*(Δ)
2 0.081 0.344
1 0.215 0.699
0.5 0.424 0.931

Backup

- **Choosing appropriate SNRs for test planning can have long-ranging implications**
  - Resource-constrained environments make accurate assessment of costs and benefits of additional testing critical
  - Best practice is to use existing data or data from previous tests to estimate SNR wherever possible
  - When this isn't possible, we can use SNRs aggregated over numerous systems to determine a plausible range
  - Focus on similar systems (similar response variable, same type of parameters, same warfare group, etc.)
  - Further updates to the database will increase robustness
- **For continuous variables, the range over which the variable will be observed can be a crucial determinant of the effect size**
  - This can be misleading, as some of these "small" effects were highly significant