JOHNS HOPKINS
APPLIED PHYSICS LABORATORY

Experimental psychology applications for cyber T&E

power analysis, IRB's, and other notes

Nathan Bos
Johns Hopkins University Applied Physics Laboratory

My background:

- Experimental psychology
- Human-computer interaction
- Group studies, educational research
- No formal training in T&E
- Evaluation experience:
  - Forensic social network analysis
  - Computer network visualizations
  - Human terrain team data collection tools
  - Serious games for intelligence analysis
  - Blue team watchstanders
  - Blue team analysts
  - Open source data collection and use
  - Counterfactual analysis
- IRB member ~ 9 years

This talk:

- Lessons learned while trying to apply experimental methods to testing and evaluation
- Observations on common T&E mistakes
- Notes on IRB usage
- Identify some gaps in T&E research (maybe)

- *Apologies in advance if some of these are obvious to long-time T&E professionals!*

Notes on experimental psychology applied to cyber T&E:

1. Maximize statistical power by any means possible
2. Move backwards from research questions not forward from data
3. Build around performance metrics, not opinions
4. Evaluate the entire socio-technical system, not just a single operator or role
5. Consider using an IRB and publishing research

Maximize statistical power by any means possible

- Academic psychology is in the midst of dealing with a “replication crisis”
- Many prior studies, including some classics, fail to replicate
- Key reasons:
  - *Poor statistical power of original studies*
  - Selection bias in what gets published; “file drawer effect”
  - ”P-hacking”
  - Abuse of exploratory data analysis
  - Other questionable research practices

Quiz:

- You are observing a large exercise where 100 total staff are working together over 6 sites. There will be, by your estimates, 31 consequential decisions. The outcome you care about is overall mission success, measured on a scale of 1 (failure) to 7 (complete success.)
- From the perspective of an experimental psychologist, what is your N? i.e. how many subjects do you have?

- N=1
- At N=1 you do not have an experiment. You have a case study or a pilot.
- Observation #1: military exercises, as currently conducted, are overrated as data collection opportunities

Statistical power analysis

- The power of a test is the probability of the test correctly identifying an effect when a true effect exists
- Typical example: I want to test whether interface A is better than interface B at helping people search a corpus of policy documents. I expect there to be a medium effect size, d=.5, (about ½ standard deviation).
- Statistical significance is set at alpha = .05, test will be an independent samples t-test
- Data points must be statistically independent
- Power should be .8. This means that, if the effect were exactly .5, 80% of the time I would find a significant difference. .8 is the standard.
- Power analysis tells me I need 33 participants in each condition, 66 total.

Resources:

Joseph Cohen, statistical power analysis for the behavioral analysis

R statistical package ‘pwr’

Standalone package ‘GPower’

	t.test, one sample, two sided, p<.05, power=.8	**anova, 3 groups, p<.05, power=.8**
Large effect size	d=.8, n=14	d=.4, n=21
Medium effect size	d=.5, n=33	d=.25, n=52
Small efect size	d=.2, n=198	d=.1, n=322

How to improve power in a typical exercise setting?

- You have 10 operators, who are working together on a day-long exercise. The tasks were designed by someone else. They are not planning to do any task more than once. They will work with each other, and delegate tasks in unpredictable ways.

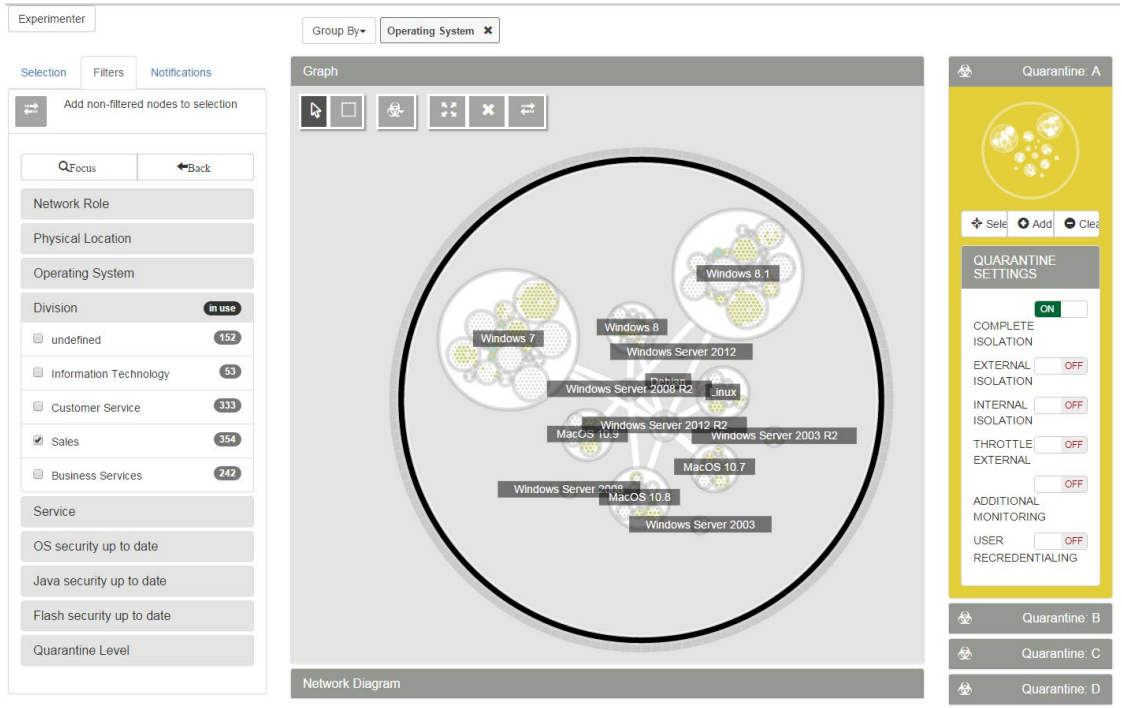
What can you do?

- Identify measurable subtasks with clear outcomes
- Add repetitive subtasks with small variations
- Develop targeted pull-out tasks: ask individuals or pairs to solve standalone problems without help outside of the exercise's main flow
- Conduct controlled testing outside of the exercise. Separate function tests, usability tests, and higher-level decision-making tasks.

Build small pull-out tasks that can be done quickly and independently. Pairs make for better verbal protocols

Example of strategic-level study

- Used PNNL's Ocelot interface to study executives' decisions about when to quarantine a network.
- Replicated a common finding of prospect theory: individuals respond to risk differently when there is a 'gain' versus 'loss' framing.



Bos, N.D., Paul, C.L. Bos, N., Paul, C. L., Gersh, J. R., Greenberg, A., Piatko, C., Sperling, S., Spitaletta, J.S., Arendt, D.L., Burtner, R. (2016). Effects of Gain / Loss Framing in Cyber Defense Decision-Making. Proceedings of the Human Factors and Ergonomics Society annual meeting 2016.

Can I just add more problems?

- Having each participant do multiple problems can, to a limited extent stretch your statistical power
- These data points are no longer independent, so dependencies must be taken into account.
- Power analysis simulations with multilevel tests are useful here. The table on the right is an example, based on some specific assumptions
- Takeaway: you can stretch power with more problems but can't escape the need for independent responses

Power variations with number of actors and problems

Number of actors (N) \ Number of problems	5	10	15	20	25	30	35	40	45	50	55	60
5	0.14	0.26	0.34	0.31	0.34	0.39	0.29	0.39	0.41	0.47	0.37	0.47
10	0.25	0.45	0.49	0.47	0.66	0.6	0.63	0.62	0.67	0.69	0.78	0.72
15	0.36	0.57	0.66	0.61	0.77	0.74	0.71	0.86	0.83	0.92	0.83	0.89
20	0.45	0.67	0.77	0.8	0.91	0.85	0.92	0.93	0.91	0.92	0.96	0.94
25	0.49	0.76	0.85	0.95	0.9	0.99	0.95	0.99	0.97	0.97	0.97	0.94
30	0.63	0.86	0.9	0.96	0.94	0.96	0.97	1	1	0.97	0.99	0.99
35	0.61	0.89	0.98	0.98	0.99	0.97	0.99	0.99	1	0.99	0.97	1
40	0.71	0.9	0.96	0.96	1	1	1	1	1	1	1	1
45	0.81	0.94	0.99	0.98	1	1	1	0.99	1	1	1	1
50	0.85	0.97	0.97	1	1	1	1	1	1	1	1	1
55	0.88	0.97	0.99	1	1	1	1	1	1	1	1	1
60	0.91	0.98	1	1	1	1	1	1	1	1	1	1

But what about…

- Conventional wisdom for usability is that if you run 5-10 tests you'll discover most of the problems. Why is this?
- Because these findings do not require statistical significance; you only need an event to occur once to record it.
  - You will not be able to estimate the relative frequency of rare problems
- If you have 5 users performing 3 tasks each, (15 trials) then you will document most problems with probability > 10% chance of occurring on any one trial.

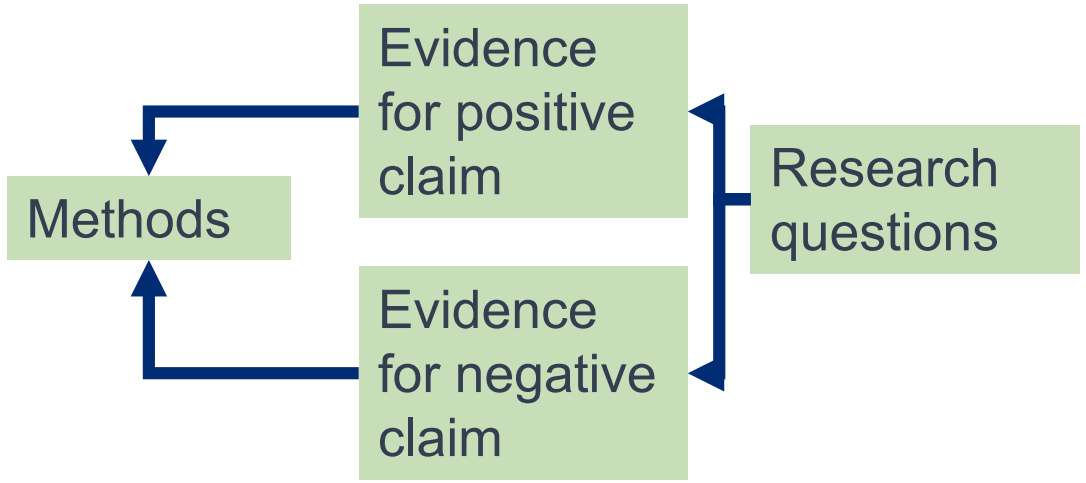
5% chance = 32 trials. 1% chance = 161 trials.

- **Need power analysis techniques designed for T&E!**

Move backwards from research questions to findings

The proper reaction if someone says, “we’ll just {videotape | record | observe |active badge} everything and make sense of it later”

- Don’t start an assessment without clearly articulated research questions, *and* a very clear mapping of how the data you are collecting will provide evidence to answer these questions
  - Make sure you can provide evidence for results in either direction, positive or negative!
- You *should* use open-ended observations, pre-interviews, etc. to form research questions. Similar to *grounded theory*, see Strauss & Corbin
- Design observations and interviews to fill in gaps, document reasons for actions



Build around performance metrics, not opinions

- The key part of behavioral evaluations are **not** surveys, interviews, brain scans, or other metrics that are added on to a test event
- The best behavioral measures are embedded in the tasks the be performed, such that *task success is directly equitable to systems success*

Build performance metrics into the task

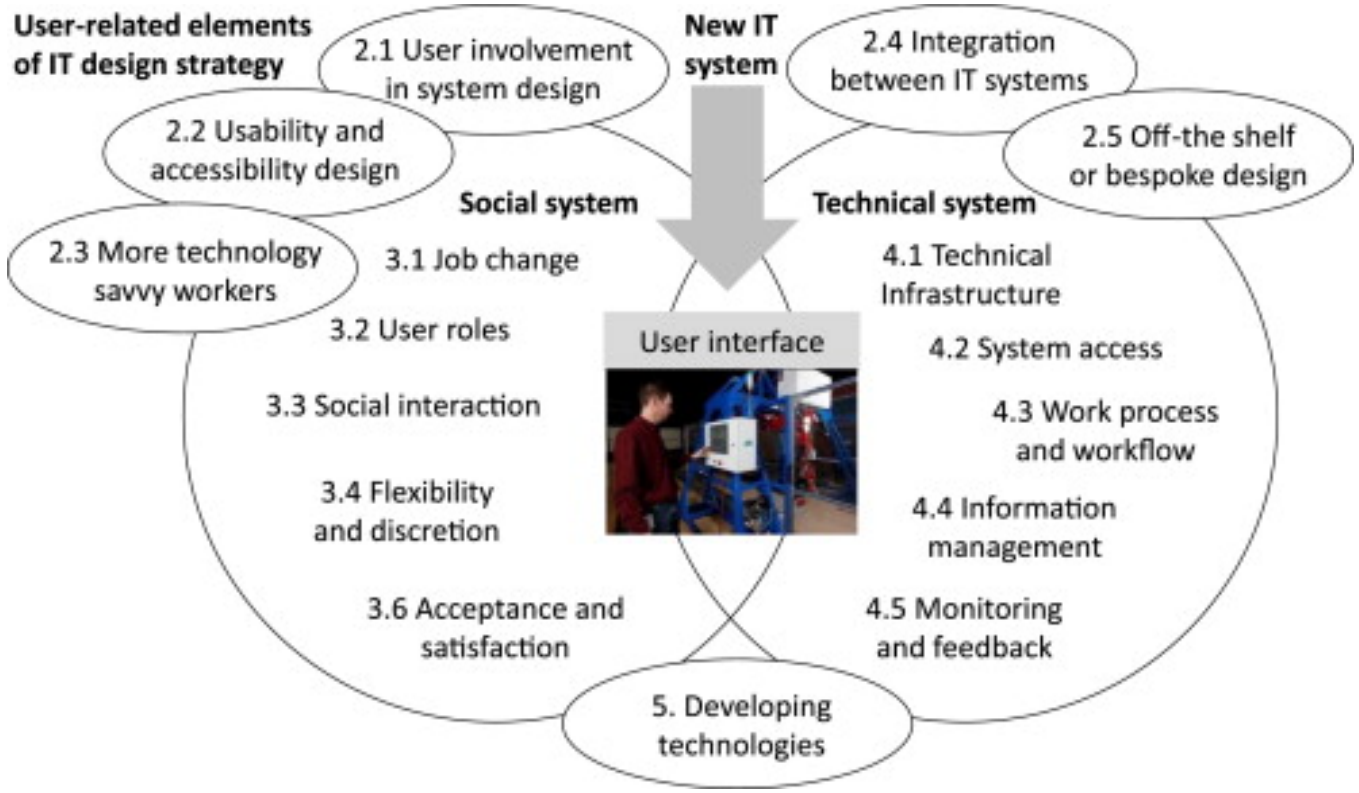
- Start with a cognitive task analysis or cognitive work analysis
- Be involved in developing the synthetic data and scenarios
- Identify key points where individuals/ teams are most likely to struggle
- Pre-test, a lot. The larger and more complex task is, the more piloting and pretesting is required. This tends to increase the expense and time needed for assessment. To control costs:
- Develop strong logging tools and evaluation routines
- Develop a ‘library’ of tasks and, if possible, range environments that you know well

- My opinion: formal exercises are often not very good test opportunities

Evaluate the entire socio-technical system

Consider:

- User interface
- Multiple use cases
- Training, support
- Usage under high stress, workload, multitasking
- Administrator interface
- Back end data import, curation, management
- Data consumers: who ultimately makes decisions with this data? Who is the user reporting to?
- Organizational structure
- Organizational culture



Maguire, M. (2014) Socio-technical systems and interaction design – 21st century relevance. From Advances in Socio-Technical Systems Understanding and Design: A Festschrift in Honour of K.D. Eason Edited by Patrick Waterson. Applied Ergonomics Volume 45, Issue 2, Part A, Pages 133-220 (March 2014)

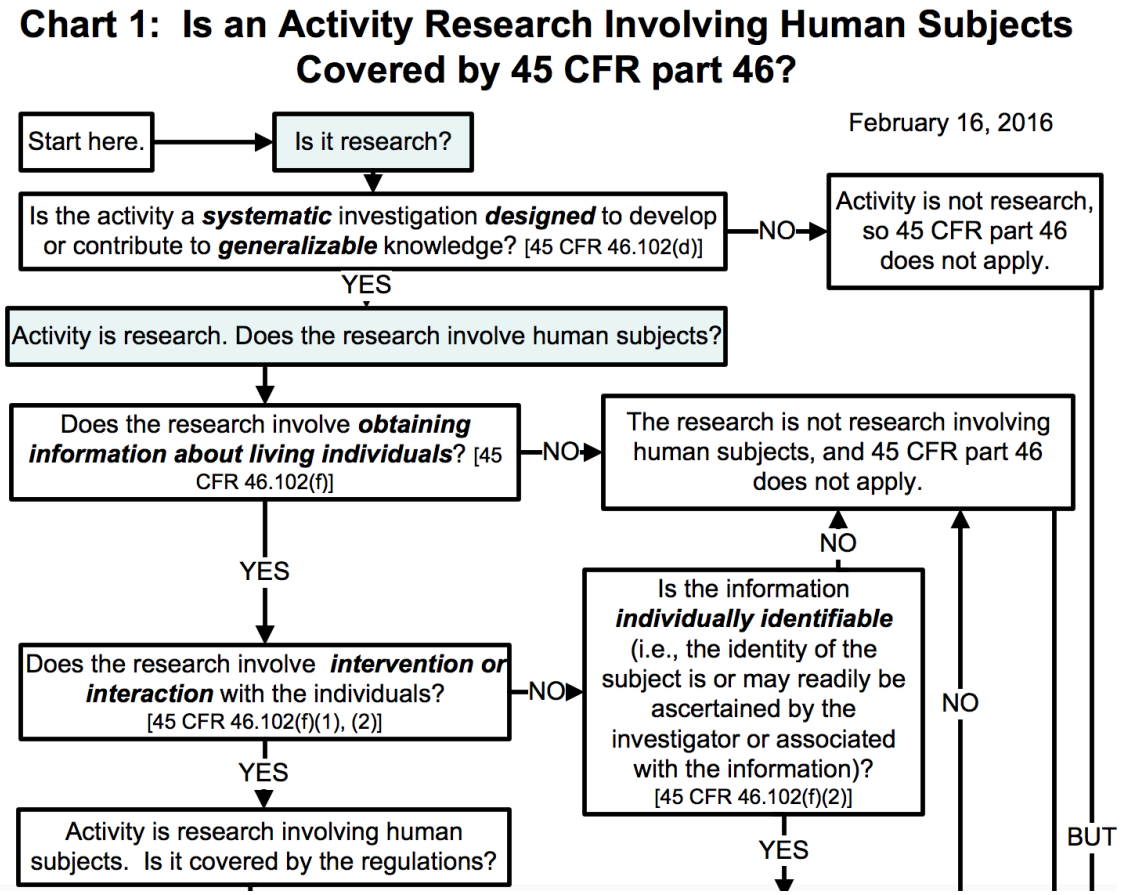
Evaluate the entire socio-technical system

Focus some attention on:

- Flow of information before it gets to the system and after it leaves the system
- Critical support roles that are not thought of as ‘users’
- Decision makers and actual decisions
- Limitations on use in the ‘real world’ due to culture, authorities, etc. These are all part of the system.
- Real use cases, especially ones that deviate from test cases

Consider using an IRB and publishing your research!

- There is very little good published research on human factors in cybersecurity!
- More work could be sanitized/ anonymized and published
- Must be research, not just evaluation. Topics include: tool design, problem-solving, risk assessment, human-computer interaction, interacting with autonomy, information-seeking behavior, collaboration processes, One impediment is getting Institutional Review Board approval for human subjects research.
- Tool evaluations may not be human subjects research even if users are involved. Is the focus on the tool or the person?



Pro tips for IRB applications

Pro tips for IRB applications:

- Anonymize as much as possible from the beginning
- The IRB cares about these things:
  - **Consent**– did the participants agree to everything that they are going to be asked to do, and did they understand what they were agreeing to? Can they quit partway through?" Note that they might be able to quit the data collection without quitting the exercise.
  - **Data handling and privacy**– have a plan for data collection and storage. Think through how data will be moved around. Anonymize/sanitize it as soon as possible.
  - **Recruitment**– how are you finding people, and are you breaking any rules/ regulations etc. in doing so? Don't be a spammer.
- The IRB does not care about every detail of your protocol, (no matter what they tell you).
  - Don't say: "At 1:05 seconds into the protocol we will ask the participant whether they have found what they were looking for in their first query."
  - Do say: "The observer will follow a protocol with specific questions related to their process. Observer will also use their own judgment follow up on interesting events."
- Using DoD personnel as participants is still problematic because HRPO vary widely in speed and competence. Work with the command and the IRB on this.

Revised common rule as of January 2018

- More research is exempt, including most behavioral research
- Exempt may still require some permissions; work with your local IRB and relevant command
- Multi-site collaboration should be easier
- Add a sentence in your consent form about data reuse and you can more freely share data with others