Evidence-Based Methods for Privacy and Identity Management

Kovila Coopamootoo, Thomas Groß

Evidence-Based Methods for Privacy and Identity Management

Kovila P.L. Coopamootoo[1⋆] and Thomas Groß[2]

[1] University of Derby, UK
[2] Newcastle University, UK

Abstract. In the advent of authoritative experiments and evidence-based methods in security research [29,21,4,2], we are convinced that privacy and identity research will benefit from the scientific method, as well. This workshop offers an introduction to selected tools of experiment design and systematic analysis. It includes key ingredients of evidence-based methods: hallmarks of sound experimentation, templates for the design of true experiments, and inferential statistics with sound power analysis. To gauge the state of play, we include a systematic literature review of the pre-proceedings of the 2016 IFIP Summerschool on Privacy and Identity Management as well as the participants' feedback on their perception on evidence-based methods. Finally, we make our case for the endorsement of evidence-based methods in privacy and identity management.

1 Introduction

The Encyclopaedia Britannica defined *science* as a "system of knowledge that is concerned with the physical world and its phenomena and that entails unbiased observations and systematic experimentation." In general, it is a purpose of science to advance human knowledge. The scientific method is evidence-based, includes principles such as falsification or reproducibility as well as statistical tools to decide between hypotheses.

To what extent is security/privacy research a science? How does research in this field advance human knowledge? In the recent years, funding bodies have sought to strengthen evidence-based research in security and privacy and, arguably, those methods have seen adoption in the field, especially under the flag of "science of security" [29,21,4,2].

Challenges. Whereas the tenets of the scientific method are often demanded, they are easily subverted by methodological mistakes or insufficient power under the all too polished surface. Ioannidis [19] gave a harsh account of the situation, arguing "why most published research findings are false."

To make matters worse, there is a replication crisis in science. For example, 270 researchers of the Open Science Collaboration [27] have reported on a dire

⋆ Major contributions were made while the author was at Newcastle University.

situation after having sought to reproduce 100 well-known results published in three major psychology journals [28]. They could only reproduce 39% of the results. It is deemed likely that the replication crisis also pertains to other fields, including security or privacy and identity management.

Even down to the nitty-gritty of statistical inference, many misconceptions and controversies have been observed, including, for instance, a comprehensive account of null hypothesis testing by Nickerson [26].

All that glitters is not gold. While evidence-based methods hold a promise to support the pursuit of knowledge in security and privacy, they ask of us great diligence to live up to their tenets. This IFIP workshop sought to sensitize participants to the hallmarks and inference methods of evidence-based research in privacy and identity management. It includes examples for true experiments as well as systematic literature reviews as two classes of evidence that are considered as most reliable.

Scope. Research methodology for evidence-based methods is a vast topic, filling tomes in the sciences. Consequently, this workshop summary will only offer a primer—an introduction to hallmarks, experiment design and statistical inference. Given that the workshop aimed at sensitizing for evidence-based methods and its requirements, we make a number of simplifications. We only focus on (a) true experiments (inducing an experiment condition), (b) hypothesis testing (rejecting a null hypothesis), (c) two conditions (control and experiment), (d) simple statistics (difference between means, t-test). Explicitly out of scope are: qualitative methods, observational studies and complex statistical models.

Outline. This workshop summary contains two theory sections on hallmarks of empirical research and statistical inference, where each of the theory section concludes with a concise checklist of quality criteria. Section 2 contains the hallmarks discussion, leading up to hypothesis testing. Then, we interleave a section on practical experiment design in Sect. 3 which reflects a round-table discussion of the workshop. From this intermezzo, we continue our theoretical inquiry with Sect. 4 on statistical inference and power. Section 5 reports on participants responses to the workshop questionnaire. We detail areas of privacy research with interest in experimental methodology, methodological issues encountered and their personal learning objective from the workshop.

2 Hallmarks of Empirical Research

Definition 1 (True Experiment [10]). *An investigation in which the investigators have sufficient control of the system under study, in particular to be able to determine the assignment of different units of study to different conditions.*

A true experiment follows requirements contributing towards rigorous science. The requirements include (a) definition of a falsifiable hypothesis, (b) defining and controlling variables, (c) assessing internal and external validity, (d) repeatability and reproducibility of the method and analysis [21,4].

2.1 Falsifiable Hypotheses

Definition 2 (Hypothesis [12]). *Specific testable predictions made generally about the response and explanatory variables in a study.*

Testing hypotheses is one of the tenets behind scientific discovery.

Popper [30] coined the theory of falsification, whereby the researcher formulates a hypothesis such that the experiment can show it to be false. According to Popper, hypotheses cannot be inductively verified, but only empirically falsified. Falsifiable hypotheses are formulated such that they can be measured and observed.

Example 1 (Falsifiable Hypotheses).

- All swans are white. [30] (Falsifying observation: a black swan)
- Higher cognitive workload implies more click-throughs on phishing links. (Falsifying observation: experiment showing equal phishing click-troughs across workloads)

2.2 Controlled Variables

In experiments, we distinguish between three types of variables: manipulated, controlled or measured. A variable that is manipulated, the independent variable, is to predict or explain the dependent or response variable.

Definition 3 (Variable). *A* variable *is some characteristic that differs from subject to subject or from time to time [12].*

(a) The independent variable *IV is a variable that is induced/manipulated [23].*

(b) The dependent variable *DV is a variable that is observed/measured [23]. A systematic change in the IV causes a change in the DV.*

(c) A confounding variable *(short:* confounder*) is an extraneous variable whose presence affects the variables being studied, so that the results do not reflect the actual relationship between the IV and DV.*

Methods to actively control confounding variables include random assignment of subjects to conditions, restricting variation in confounders (e.g., selecting subjects of the same age eliminates confounding by age) and matching potential confounders across conditions. Statistical models can also be used to adjust for the bias introduced by a confounder during analysis.

2.3 Validity.

Validity refers to the extent to which a measuring instrument is measuring what was intended [12], where a change in the IV entails a change in the DV.

Definition 4 (Validity [9]). *The best available approximation to the truth and falsity of propositions.*

What we seek to validate are the statements, inferences and conclusions that we draw from results of empirical research [3]. We differentiate between internal and external validity.

Internal Validity. In most experiments, researchers are aiming to find out if IV A has an effect on DV B. If the experiment does not offer any alternative causes nor explanations on the outcome on B, then the experiment is internally valid.

Definition 5 (Internal Validity [3]). *The truth that can be assigned to the conclusion that a cause-effect relationship between an IV and a DV has been established within the context of the particular research setting.*

External Validity refers to the extent to which the study findings are generalizable from a laboratory setting to other settings.

Definition 6 (External Validity [3]). *The question of whether an effect (and its underlying processes) that has been demonstrated in one research setting would be obtained in other settings, with different research participants and different research procedures.*

Not all experiments can be both internally and externally valid. Depending on the purpose of the experiment, researchers need to make a trade-off.

2.4 Repeatability and Reproducibility

Replication is the attempt to recreate the conditions sufficient to obtaining a previous observed finding [28]. Scientific claims gain credence when their supporting evidence can be replicated [28].

Replication has been highlighted as a problem in scientific research. For example, the Open Science Collaboration [28] conducted a large-scale replication study ($N = 100$) of psychological journals and found that replication effects were on average half the magnitude of original effects.

We distinguish between repeatability and reproducibility as two conceptual frames for replication.

Definition 7 (Repeatability [12]). *The closeness of the results obtained in the same test material by the same observer or technician using the same equipment, apparatus and/or reagents over reasonably short intervals of time.*

Definition 8 (Reproducibility [12]). *The closeness of results obtained on the same test material under changes of reagents, conditions, technicians, apparatus, laboratories and so on.*

Remark 1 (Repeatability vs. Reproducibility).
While repeatability refers to replicating the experiment by keeping everything same (including the experimenter), reproducibility refers to altering specific components while keeping the design consistent, especially when the experiment is reproduced by an independent experimenter.

2.5 Hypothesis Testing

In this workshop summary, we limit the scope of our inquiry to *hypothesis testing* [15,25,20], a particular method of statistical inference that seeks to distinguish between hypotheses. We focus on making a decision between a null hypothesis and an alternative hypothesis.

Definition 9 (Hypothesis Testing).

(a) A statistical hypothesis test *is a method of statistical inference in which a hypothesis of a proposed statistical relationship is compared to an idealized null hypothesis that claims there is no relationship.*
(b) The null hypothesis H_0 *is the statistical hypothesis that there is no effect, no difference between conditions.*
(c) The alternative hypothesis H_1 *is the statistical hypothesis that there is an effect, a difference between conditions.*

Hypotheses are expressed on the population statistics, not the sample statistics.

Example 2 (Difference of Means). When considering the means across two conditions, the two hypotheses are:
- Null hypothesis H_0: $\mu_1 = \mu_2$,
- Alternative hypothesis H_1: $\mu_1 \neq \mu_2$.

A sound procedure for hypothesis testing will proceed as illustrated in Fig. 1.

1. State null hypothesis H_0 and alternative hypothesis H_1 explicitly, first.
2. Evaluate the statistical assumptions made, select a relevant test statistic, and select a significance level α, a probability threshold below which the null hypothesis will be rejected (cf. Sect. 4.1).
3. Evaluate the statistical inference by calculating the test statistic. Reject the null hypothesis if and only if the p-value is less than the specified significance level α.

We will discuss statistical inference and p-values in Sect. 4.1.

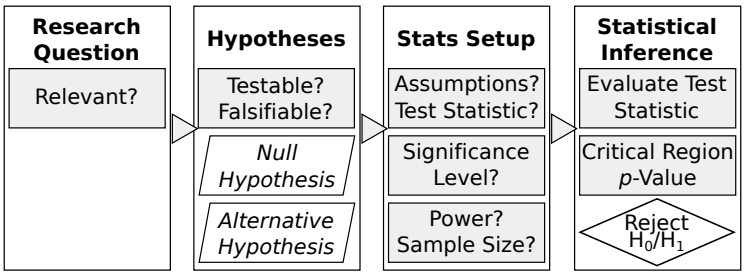


Fig. 1: Simplified process of hypothesis testing.

Remark 2 (Controversy and Criticism). There has been much controversy about hypothesis testing. Among its most vocal critics is Jacob Cohen [7]. First, we need to note that a statistically significant result only means that the effect is deemed not *nil*, nothing more. Slavishly following the "sanctification by significance" has been considered as one of great ailments in scientific reasoning. Nickerson [26] offers a comprehensive overview of the controversies around *Null Hypothesis Significance Testing (NHST)*, of which we highlight misconceptions on p-values in Remark 3 (Sect. 4.1).

Multiple proponents have argued to deprioritize hypothesis testing in favor of robust estimation with confidence intervals, e.g., as expressed by Gardner and Altman [17]. The American Psychology Association (APA) [1] has consequently made the reporting of confidence intervals a minimum standard.

While this workshop summary adheres to hypothesis testing, we advocate a cautious and diligent interpretation: Consider the size of effects investigated, the required power and sample size to detect those effects (cf. Sect. 4.2).

2.6 Checklist: Hallmarks

- ☐ Make hypotheses falsifiable, i.e., construct them such that experiments or observations can show them to be false.
- ☐ Specify independent variables (IVs) and their manipulation. Operationalise dependent variables (DVs) and specify validated measurements.
- ☐ Explicitly declare null hypothesis and alternative hypothesis *a priori*.
- ☐ If feasible, prepare a randomized controlled testing the hypotheses.
- ☐ Control for confounders, e.g., by restricting variation or matching subjects.
- ☐ Establish to what extent a change in IV entails a change in DV. Report biases and assumptions that impact this entailment.
- ☐ Make it clear whether the study is repeating or reproducing existing research. Document recruitment, sampling, procedure, experiment design, manipulations, measurements, analyses clearly for forward reproducibility.

3 An Exercise in Experiment Design

We have prepared the ground by introducing hypothesis testing. Before we proceed with statistical inference in Sect. 4, we discuss an exercise in experiment design based on a hypothetical scenario.

Example 3 (Scenario "When the cat's away, the mice do play").
A security company observes that in the evenings—when the boss—is away more dangerous sites are accessed than during day times.

3.1 Developing Research Questions.

The participants are asked to answer the following questions:

1. What is an interesting research question (RQ) for the scenario of clicks to dangerous sites?
2. Independent Variable (IV): What factor influences the number of clicks on dangerous sites?
3. Dependent Variable (DV): How can we measure the outcome reliably?
4. What is a testable null hypothesis (H_0)?
5. What is the alternative hypothesis (H_1)?

We advocate for the simple example research questions to create a core experiment design to nail down how the IV is manipulated, how he success of this manipulation is checked, and how the DV is reliably measured. We offer a template in Fig. 2. Cf. Field and Hole [14] or Montgomery [24] for experiment designs.

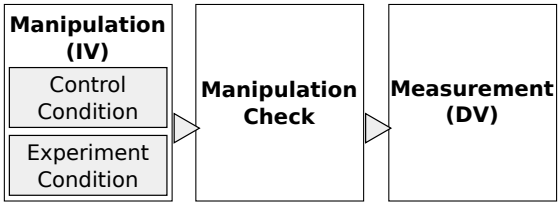


Fig. 2: A core template for a two-condition experiment with manipulation check.

The following examples were designed by workshop participants in round-table discussions in response to the questions above.

Example 4 (Design Group 1).
RQ. How does the presence of the boss impact clicking dangerous links? **IV.** Presence of the boss. **DV.** #mistakes clicking dangerous links. $\mathbf{H_0}$. The mean number of mistakes is equal between the conditions "boss present" and "boss absent." $\mathbf{H_1}$. The mean number of mistakes is greater when the boss is absent.

Example 5 (Design Group 2).
RQ. How does cognitive workload impact clicking phishing links? **IV.** Cognitive workload. **DV.** #mistakes clicking phishing links. $\mathbf{H_0}$. The mean number of phishing mistakes is equal between the depleted experiment condition and the non-depleted control condition. $\mathbf{H_1}$. The mean number of phishing mistakes is greater in the depleted condition than in the control condition.

Example 6 (Design Group 3).
RQ. How does down-time impact clicking of dangerous sites? **IV.** Down-time without customer. **DV.** #mistakes accessing dangerous sites. $\mathbf{H_0}$. The mean number of accesses to dangerous sites is independent from the measured down-time. $\mathbf{H_1}$. An increased down-time implies an increased mean number of accesses to dangerous sites.

We see from this example that for a given scenario a variety of relevant research questions and operationalizations in statistical hypotheses is possible. Consequently, it is crucial to write down precisely what is being investigated before the experiment is designed. The key points here are to commit to the independent and dependent variables, to settle the manipulation and measurement methods used, and to express the null and alternative hypotheses in the exact terms of these variables.

3.2 Structured Abstract.

We recommend a *structured abstract* as a concise tool of stating the intention of an experiment (in less than one page). The structured abstract covers

1. **Background.** The motivation and theoretical context of the experiment.
2. **Aim.** The goal of the experiment expressed in one concise sentence.
3. **Method.** The concise method of the experiment, including sample size, group design, what is manipulated (IV) and what is measured (DV).
4. **Expected Results.** The factual outcomes expected from the experiment.
5. **Expected Impact.** So what? What does the experiment mean?

A structured abstract is a superb tool in reporting findings soundly and endorsed by specialist venues, such as Learning from Authoritative Security Experiment Results (LASER)[3]. Example 7 reports the outcome of the scenario study.

4 Statistical Inference and Power

Table 1: Statistical inference decision matrix.

Decision \ Reality	H_0 is TRUE. H_1 is FALSE.	H_0 is FALSE. H_1 is TRUE.
We reject H_1.	**Confidence Level** $1-\alpha$	**TYPE II ERROR** β
We reject H_0.	**TYPE I ERROR** α	**POWER** $1-\beta$

[3] `http://www.laser-workshop.org`

Example 7 (Structured Abstract).
Background. Psychology research predicts an impact of tiredness on decision making.
Aim. We investigate the impact of tiredness on mistakes on phishing click-throughs.
Method. Two groups of 20 participants each were asked to evaluate 50 mixed e-mails (25 phishing), one group was tired, the other was not. We compared the number of mistakes across groups.
Results. The mean number of mistakes of the tired group ($M_{\mathsf{E}} = 13.9, SD_{\mathsf{E}} = 5.77$) was significantly greater than that of the control group ($M_{\mathsf{C}} = 10.75, SD_{\mathsf{C}} = 3.75$), two-tailed $t(38) = -2.047, p = .049$, 95% CI$[0.18, 6.28]$. We observed a medium effect size ($d = 0.68$).[a] The experiment achieved a power of 55%.
Impact. Tired users succumb to phishing.

[a] Reporting confidence intervals (CI) and the effect size as mandated by the APA guidelines [1].

4.1 Statistical Inference

As we have seen in Sect. 2.5, we seek to decide between the null hypothesis H_0 and the alternative hypothesis H_1. We do not know what the situation in reality is: whether H_0 is true or false. All we can do is making an observation (in an experiment) and base a decision to reject or accept H_0 on the likelihood of that observation. Because H_0 and H_1 are meant to be complements, we end up with four decision outcomes summarized in Tab. 1.

Let us consider the left-hand column of Tab. 1 first: In reality, the null hypothesis H_0 is true. We specify in advance a significance criterion α, which quantifies the likelihood of mistakenly rejecting the null hypothesis H_0. As Cohen formulates it [8], "α represents a policy: the maximum risk of attending such a rejection." If we reject H_0 even though H_0 is true, we commit a *Type I Error*.

If we correctly reject the alternative hypothesis H_1 and hence accept the null hypothesis H_0, we do so at a confidence level $1 - \alpha$.

Test Statistics and *p*-Value. We conduct statistical tests to evaluate how likely the observation is, *assuming the null hypothesis to be true.*

Definition 10 (*p*-Value [22]). *A* p-Value *is the probability of data as extreme or more extreme as that obtained, computed under the presumption of the truth of the null hypothesis* H_0*. In symbols, if we let* D *stand for data as or more extreme as that obtained, then a p-value is the conditional probability*

$$p = Pr(D|\mathsf{H}_0).$$

Hypothesis testing with significant p-values attempts a statistical *proof by contradiction* indirectly. If the p-value is smaller than the specified level of significance

α, we call a test statistic *statistically significant* and are entitled to reject the null hypothesis H_0.

Remark 3 (p-Value Misconceptions).
Unfortunately, p-values are often misinterpreted, even in text books. Maxwell and Delaney [22, p.48] as well as Nickerson [26] offer some pointers for typical misinterpretations.
(a) We emphasize that in almost all cases, it holds that

$$p = Pr(D|\mathsf{H}_0) \neq Pr(\mathsf{H}_0|D).$$

Considering these two conditional probabilities equivalent is a fallacy, called "the confusion of the inverse."
(b) It is also a grave mistake to believe that p is the probability of the null hypothesis being true.
(c) The likelihood of the alternative hypothesis H_1 is only indirectly related to the p-value [31,19].[4] Cohen [7] is vocal that the p-value "cannot tell us anything about the probability that the [alternative] hypothesis is true."
(d) Note especially that $p = Pr(D|\mathsf{H}_0)$ is not a complement of $Pr(D|\mathsf{H}_1)$ [26].

4.2 Effect Size and Power

Cohen [8] exhorts that an effect that is statistically significant is not necessarily *scientifically significant* or important. The importance of an effect is largely linked to the magnitude of the effect. For the example of the difference between two means, we are interested how large the difference between the two populations is, and whether it constitutes a non-trivial difference.

Effect Size. We seek to quantify of the magnitude of an effect.

Definition 11 (Effect Size [8]). *The* effect size (ES) *is the degree to which* H_0 *is false. It is indexed by the discrepancy between* H_1 *and* H_0*. Each statistical test has its own ES index. All the indexes are scale free and continuous, ranging upward from zero, and for all, the* H_0 *is that ES* $= 0$*.*

The importance of a significant effect with effect sizes is considered that crucial in the science, that the American Psychology Association (APA) [1] states that "estimates of appropriate effect sizes [...] are the minimum expectations."

There are two main families of effect sizes [11]: (a) the d family, assessing the differences between groups, and (b) the r family, measuring the strength of a relationship. Effect sizes can be further specified by, for instance, regression

[4] Conditional probabilities follow Bayes' Theorem,

$$Pr(\mathsf{H}_1|D) = \frac{Pr(D|\mathsf{H}_1)Pr(\mathsf{H}_1)}{Pr(D|\mathsf{H}_1)Pr(\mathsf{H}_1) + Pr(D|\mathsf{H}_0)Pr(\mathsf{H}_0)}.$$

Nickerson [26] discusses the links and caveats in depth.

coefficients or odds ratios. In this workshop summary, we focus on the d-family of effect sizes, especially on the difference between two means, measured with Cohen's d. We refer to Cohen [8,6], Ellis [11] and Fritz et al. [16] for overviews of different effect size types and their calculations.

Power. Now we are prepared to consider the right-hand side of Tab. 1: How do we fare in a situation in which the null hypothesis H_0 is actually false?

If we accept the null hypothesis H_0 mistakenly even though the alternative hypothesis H_1 is true, then we have committed a *Type II Error*. The likelihood of committing such an error is called β.

Consequently, if we are in the case that the alternative hypothesis H_1 is actually true, and we make a correct decision to reject the null hypothesis H_0 we do so at the likelihood of the power of our test.

Definition 12 (Power [8]). *The* statistical power *of a significance test is the long-term probability, given the population ES, α, and N of rejecting H_0. Power is $1-\beta$, the probability of rejecting a false H_0.*

The four quantities sample size N, effect size (ES), significance level α and power $1-\beta$ are mathematically connected; given three of them, the fourth quantity can be computed. We recommend G*Power [13] for this computation.

In authoritative experimentation, we seek to create experiments with sufficient power (as a commonly used rule-of-thumb, $1-\beta > .8$) to have a sufficient likelihood of correctly rejecting the null hypothesis H_0. Cohen [5] and others have observed time and again an abysmal lack of power in scientific experiments.

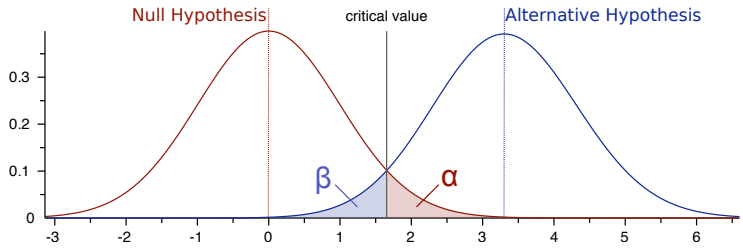


Fig. 3: Hypothesis testing with null hypothesis H_0 test distribution on the left and alternative hypothesis H_1 test distribution on the right. The null hypothesis is rejected if the a critical value is passed. The graph marks the critical areas for α and β, that is, the likelihoods to make Type I and Type II errors.

Underpowered Experiments. For research in privacy and identity management, we anticipate that the power of experiments is often too low, below $1-\beta = .5$, and the likelihood of correctly rejecting the null hypothesis basically a coin toss. We believe the experimenters underestimate the sample size, because of a missing

understanding of effect sizes and *a priori* power analysis. Fig. 4 illustrates the sample sizes needed for different levels of power.

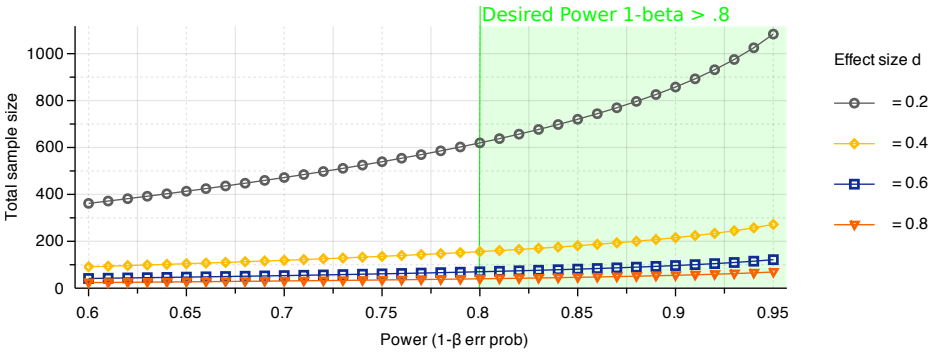


Fig. 4: Power achieved for different effect sizes and sample sizes. It is apparent that a desired power of more than $1 - \beta = .8$ needs large sample sizes N for smaller effect sizes d. (Here for an one-tailed independent samples t-test, $\alpha = .05$)

Remark 4 ($N = 30$ debunked). There was a myth of a "rule-of-thumb" to run experiments with a per-group sample size of $N = 30$. This was debunked by Jacob Cohen [7], *Some Things You Learn Aren't So*. The sensitivity of two-tailed independent samples t-test for significance level $\alpha = .05$ and power $1 - \beta = .95$ implies required effect size: $d = 0.94$ (large). Smaller effect sizes d will not be detected at this power. At a medium effect size ($d = .5$), such an experiment will only achieve a power $(1 - \beta) = .48$, a coin toss.

High-Powered Experiments. As Cohen argues [7], the null hypothesis—that there is no effect whatsoever—is never actually true in reality. With a large enough sample—and thereby large enough power—even infinitesimal effects can still be detected with statistical significance (cf. Fig. 5). Consequently, it is crucial to put the hypothesis testing and the rejection of the null hypothesis with statistical significance into context of the sample size and achieved power.

4.3 Checklist: Statistical Inference

Table 2 contains further reading.

- ☐ Specify the exact contents of the statistical inference precisely, e.g., in a structured abstract naming IVs, DVs and hypotheses.
- ☐ Choose relevant test statistics and evaluate their assumptions carefully.
- ☐ Conduct an *a priori* power analysis to determine the required sample size for a committed significance level α and an appropriate power $1 - \beta > 80\%$.
- ☐ Exercise diligence in interpreting p-values and significance, putting them into context with effect sizes and the post-hoc power the experiment achieved.

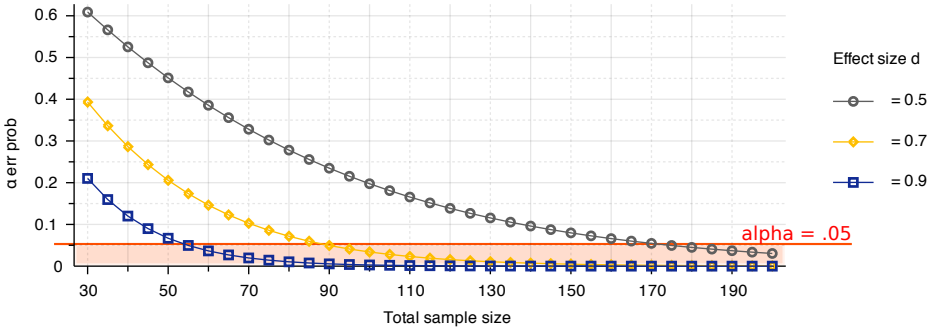


Fig. 5: α-probability for different sample sizes N and effect sizes d (one-tailed independent-samples t-test at $1-\beta = .95$), illustrating that—with a large enough sample size—even smallest effects can be detected with statistical significance.

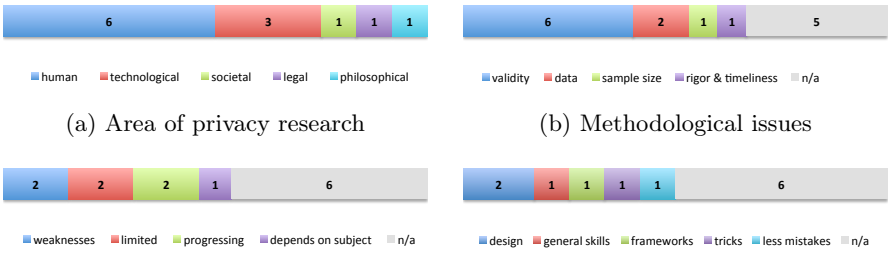
Table 2: Further reading on statistical inference and power.

Reference	Title	Comment
Montgomery 2012 [24]	Design and Analysis of Experiments	Detailed treatment of design and analysis of experiments.
Howell 2012 [18]	Statistical Methods for Psychology	Statistics for experiments with human factors.
Cohen 1992 [8]	A Power Primer	The quintessential concise introduction to effect size and power.
Fritz et al. 2012 [16]	Effect size estimates	Survey of the use of effect size types incl. best practices for their computation and transformation.

☐ Report the results following the APA Guidelines [1], especially by reporting appropriate effect sizes and confidence intervals. Include all data needed to recompute the results and their effect sizes (test statistics, standard deviations, coefficients, etc.).

5 Participant Feedback

We asked participants to fill in a questionnaire just before starting the workshop. We summarize the outcomes of the 12 respondents in Fig. 6.



(a) Area of privacy research

(b) Methodological issues

(c) State of the field

(d) Requirements

Fig. 6: Feedback of 12 participants

5.1 Area of Privacy Research

Second we asked participants *"What area(s) of privacy do you research/interests you?"* Of the 12 participants, six reported the human dimension of privacy, three reported technological aspects (geolocation and transparency), one reported the societal dimension and personal data (P3: "personal data, social exclusion, effect on the use of personal data"), one reported legal dimension in the context of health-care (P12: "Privacy for health-care systems, compliance with legal frameworks of privacy and data protection"), one philosophical dimension.

The human dimension responses covered aspects of P1: "attitude, behaviour, decision-making;" P2: "HCI, usability;" P8: "views of privacy among 'normal' people;" P9: "perceptions of privacy, how to make people more aware;" P10: "genomic privacy;" P11: "corruption and human behaviour."

5.2 Methodological Issues

Third, we asked participants *"What methodological issues have you encountered?"* From the 12 response sheets, five were excluded as they did not answer the question. Validity was the most recurring response (six), with three participants pointing towards internal validity: P11 "measuring what is intended;" P4

"software can have errors and it is unclear if experimental results can be caused by error programs." One participant stated on validity and confounders, P9 "not understanding behavioural issues [...] may ruin months of data gathering." Three responses were about external validity. P4 and P9: "generalisability;" P4 and P5: "representativeness." Two responses were about data: P1: "difficulty investigating categorical data" or large-scale data gathering (P3) The other responses included P1: "sample size;" P8 mentions running timely rigorous experiments: "experiments that can be done rigorously yet in a timely manner."

5.3 Opinion on Experimental Research in Privacy

Fourth, we asked *"What is the state of experimental research in privacy?"* From the 12 participants, five were new to privacy or reported they did not know. Two participants pointed to weaknesses, with P9 suggesting the state to be "dubious" and P11: "fairly poor." Two participants suggested an early state, P10, suggested experimental research in privacy is in "early development;" P5 stated it is "limited." Two participants suggested the state is progressing, with P1: "there are many longitudinal studies on privacy behaviour" and P6: "in progress." One participant, P4, suggested that it depends on the subject.

5.4 Requirements on Workshop

Fifth, we asked *"What would you like to learn in this workshop today?"* Of the 12 participants, six either did not provide a response or understand the question; six other thought the workshop would improve their skills in one way or another. Two participants were concerned with experiment design; P8 thought of "Good ways of running experiments;" P9 expected "better ideas to design experiments." P11 sought to learn about: "frameworks." P2 mentioned "tips and trades (*sic*) in ways forward;" and P1 expected help on "how to perform research with less mistakes."

6 Conclusion

From the participants' feedback in Sect. 5 we see that, on the one hand, participants are interested in research that lend themselves to evidence-based methods, such as human dimensions of privacy and identity. On the other hand, they report a dire situation of the state of play in the field and a need to learn more on research methodology.

> *"[Perhaps], we should simply study our Mr. Pritchard and learn our rhyme and meter and go quietly about the business of achieving other ambitions."* — John Keating, Dead Poets Society

Should we simply run studies that receive "pass" marks in our community—ignoring the depths of evidence-based methods—and go quietly about the business of achieving other ambitions? Mastering evidence-based methods is a challenging prospect, daunting at times. However, what is at stake here is our capacity as a community to truly learn from our research and advance our field's

body of knowledge. Consequently, we certainly advocate going deep in research methodology.

How? From the workshop experience, we believe there are three key ingredients that are reinforcing each other. (a) First, we would focus on the tenets of reaching clarity on research questions, hypotheses and variables, ideally specified in concise structured abstracts. (b) Second, we advocate the specification of sound experiment designs that not only replicate validated methods but also specify their components in such detail that they propagate forward reproducibility by other investigators. (c) Finally, we stress the importance of the quantitative tools from correct statistical inference, over effect sizes and power, to interval estimation, all strengthening the reliability of the reported results. We are convinced that these three ingredients are essential to advance the body of knowledge of our field.

Acknowledgments

We are grateful for the contributions and feedback from participants of the workshop. We are grateful for the discussions with Roy Maxion on evidence-based methods for cyber security. The preparation of the evidence-based methods workshop was in parts funded by the EPSRC Research Institute in Science of Cyber Security (RISCS), grant *EP/K006568/1*. This work was supported by a Newcastle-sponsored International Research Collaboration Award (IRCA) for work with Carnegie-Mellon University (CMU).

References

1. American Psychological Association (APA): Publication Manual of the American Psychological Association. American Psychological Association (APA), 6th revised edn. (July 2009)
2. Balenson, D., Tinnel, L., Benzel, T.: Cybersecurity Experimentation of the Future (CEF): Catalyzing a new generation of experimental cybersecurity research – final report. Tech. rep., SRI International and USC Information Sciences Institute (Jul 2015)
3. Brewer, M.B.: Research design and issues of validity. Handbook of research methods in social and personality psychology pp. 3–16 (2000)
4. Carroll, T.E., Manz, D., Edgar, T., Greitzer, F.L.: Realizing scientific methods for cyber security. In: Proceedings of the 2012 Workshop on Learning from Authoritative Security Experiment Results. pp. 19–24. ACM (2012)
5. Cohen, J.: The statistical power of abnormal-social psychological research: a review. The Journal of Abnormal and Social Psychology 65(3), 145 (1962)
6. Cohen, J.: Statistical power analysis for the behavioral sciences. Psychology Press, Taylor & Francis Group, LCC, New York, 2nd edn. (1988)
7. Cohen, J.: Things i have learned (so far). American psychologist 45(12), 1304 (1990)
8. Cohen, J.: A power primer. Psychological bulletin 112(1), 155 (1992)
9. Cook, T.D., Campbell, D.T.: Quasi-experimentation: Design and analysis for field settings. Rand McNally (1979)

10. Dodge, Y. (ed.): Oxford Dictionary of Statistical Terms. Oxford University Press on Demand (2006)
11. Ellis, P.D.: The essential guide to effect sizes: Statistical power, meta-analysis, and the interpretation of research results. Cambridge University Press (2010)
12. Everitt, B.: Cambridge dictionary of statistics. Cambridge University Press (1998)
13. Faul, F., Erdfelder, E., Lang, A.G., Buchner, A.: G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behavior research methods 39(2), 175–191 (2007)
14. Field, A., Hole, G.: How to design and report experiments. Sage (2003)
15. Fisher, R.A.: Statistical methods for research workers. Genesis Publishing Pvt Ltd (1925)
16. Fritz, C.O., Morris, P.E., Richler, J.J.: Effect size estimates: current use, calculations, and interpretation. Journal of Experimental Psychology: General 141(1), 2 (2012)
17. Gardner, M.J., Altman, D.G.: Confidence intervals rather than p values: estimation rather than hypothesis testing. Br Med J (Clin Res Ed) 292(6522), 746–750 (1986)
18. Howell, D.C.: Statistical methods for psychology. Cengage Learning, 8th edn. (2012)
19. Ioannidis, J.P.: Why most published research findings are false. PLoS Med 2(8), e124 (2005)
20. Lehmann, E.L.: The Fisher, Neyman-Pearson theories of testing hypotheses: One theory or two? In: Selected Works of EL Lehmann, pp. 201–208. Springer (2012)
21. Maxion, R.: Making experiments dependable. Springer (2011)
22. Maxwell, S.E., Delaney, H.D.: Designing experiments and analyzing data: A model comparison perspective, vol. 1. Psychology Press, 2nd edn. (2004)
23. Miller, S.: Experimental design and statistics. Routledge (2005)
24. Montgomery, D.C.: Design and analysis of experiments. John Wiley & Sons, 8th edn. (2012)
25. Neyman, J., Pearson, E.S.: On the use and interpretation of certain test criteria for purposes of statistical inference: Part i. Biometrika pp. 175–240 (1928)
26. Nickerson, R.S.: Null hypothesis significance testing: a review of an old and continuing controversy. Psychological methods 5(2), 241 (2000)
27. Open Science Collaboration: An open, large-scale, collaborative effort to estimate the reproducibility of psychological science. Perspectives on Psychological Science 7(6), 657–660 (2012)
28. Open Science Collaboration: Estimating the reproducibility of psychological science. Science 349(6251), aac4716 (2015)
29. Peisert, S., Bishop, M.: How to design computer security experiments. In: Fifth World Conference on Information Security Education. pp. 141–148. Springer (2007)
30. Popper, K.: The logic of scientific discovery. Routledge (2005)
31. Wacholder, S., Chanock, S., Garcia-Closas, M., Rothman, N., et al.: Assessing the probability that a positive report is false: an approach for molecular epidemiology studies. Journal of the National Cancer Institute 96(6), 434–442 (2004)