SCHOOL OF
COMPUTING

Title: A Codebook for Evidence-Based Research
The Nifty Nine Completeness Indicators v1.1

Names: Kovila P.L. Coopamootoo - Newcastle University
Thomas Groß - Newcastle University

TECHNICAL REPORT SERIES

No. CS-TR- 1514 2017

Bibliographical details

Title and Authors

Title: A Codebook for Evidence-Based Research
The Nifty Nine Completeness Indicators v1.1

Kovila P L Coopamootoo, Thomas Groß

NEWCASTLE UNIVERSITY
School of Computing. Technical Report Series. CS-TR- 1514

Abstract
Background. Since a few years, rigorous research in privacy and security has been promoted under the flag of "science of security." While, the hallmarks for scientific research includes reproducibility, validity, statistical inferences and parameter estimation, it can be difficult to actually evaluate the quality of research studies unless a set of indicators are clearly defined.
Aim. We propose 9 Quality Indicators to evaluate experimental research.

About the authors

Dr Kovila P.L. Coopamootoo is a Research Fellow in the Secure & Resilient Systems group and the Centre for Cyber Crime and Computer Security at Newcastle University.
Her two research strands are within (1) Usable Privacy and Security and (2) Evidence-Based Methods. She designs and runs lab and online studies, in particular to understand attitude and drivers of behaviour (including emotional states, efficacy, fatigue, stress and cognitive effort). This line of research targets supporting and empowering users towards privacy and optimum security. She employs scientific methodology for cyber-security and privacy and seeks to produce clear guidelines and tools for evaluating research quality.

Current Research: Cyber Security, Privacy and Evidence-based Methods for Security

- Thomas Gross - I'm a Tenured Reader in System Security (Associate Professor) at the Newcastle University. I'm the Director of the Centre for Cybercrime and Computer Security (CCCS), a UK Academic Centre of Excellence in Cyber Security Research (ACE-CSR). I'm a member of the Secure and Resilient Systems group and the Centre for Software Reliability (CSR).

Suggested keywords

Cyber Security, Experiment, Science, Evidence-Based, Guidelines

TECHNICAL REPORT SERIES

No. CS-TR- 1514 Date. November 2017

Title: A Codebook for Evidence-Based Research
The Nifty Nine Completeness Indicators v1.1

Authors.

Kovila P.L. Coopamootoo, Newcastle University
Thomas Groß, Newcastle University

Abstract.

Background. Since a few years, rigorous research in privacy and security has been promoted under the flag of “science of security.” While, the hallmarks for scientific research includes reproducibility, validity, statistical inferences and parameter estimation, it can be difficult to actually evaluate the quality of research studies unless a set of indicators are clearly defined.
Aim. We propose 9 Completeness Indicators to evaluate experimental research.
Method. We design a codebook with criteria and success, partial and failure marks for each of the
9 Completeness Indicators. The codebook provides brief definitions together with coding examples grounded in actual study publications.
Anticipated Results. We expect the codebook to act as a valuable guideline on its own. We expect the codebook to be easily understood by researchers and to produce high inter-rater reliability results.
Anticipated Conclusions. We anticipate the codebook will support researchers in evaluating their own research and reviewers evaluating manuscripts under-review, in an effective way, and can provide solid evidence that substantiates review decisions.

A Codebook for Evidence-Based Research
The Nifty Nine Completeness Indicators
v1.1

Kovila P.L. Coopamootoo
Newcastle University

Thomas Groß
Newcastle University

Structured Abstract

Background. Since a few years, rigorous research in privacy and security has been promoted under the flag of "science of security." While, the hallmarks for scientific research includes reproducibility, validity, statistical inferences and parameter estimation, it can be difficult to actually evaluate the quality of research studies unless a set of indicators are clearly defined.

Aim. We propose 9 Completeness Indicators to evaluate experimental research.

Method. We design a codebook with criteria and success, partial and failure marks for each of the 9 Completeness Indicators. The codebook provides brief definitions together with coding examples grounded in actual study publications.

Anticipated Results. We expect the codebook to act as a valuable guideline on its own. We expect the codebook to be easily understood by researchers and to produce high inter-rater reliability results.

Anticipated Conclusions. We anticipate the codebook will support researchers in evaluating their own research and reviewers evaluating manuscripts under-review, in an effective way, and can provide solid evidence that substantiates review decisions.

1 Introduction

This codebook is aimed to facilitate evaluation of experimental research across 5 Research Questions (RQs). These are

RQ1 Did the experiment repeat or reproduce existing studies/methods? Was the experiment sufficiently reported to enable reproducibility?

RQ2 To what extent were the described studies internally valid?

RQ3 How many of the eligible papers reported results from experiments correctly according to APA guidance?

RQ4 To what extent were effect sizes and power estimates provided? How many of the studies had appropriate power?

RQ5 How many of the results reported agree with an independent recalculation of test statistics and effect sizes?

The 5 research questions lead to 9 Completeness Indicators (CI) as shown in Figure 1.

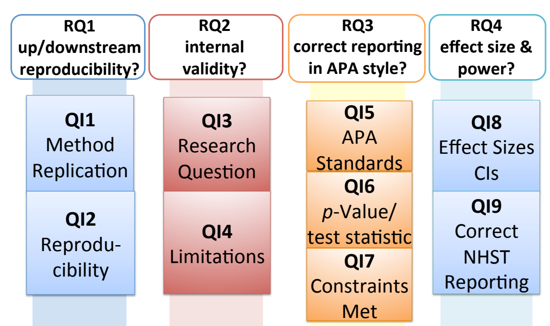

Figure 1: CIs derived from the research questions.

The full list of CIs is:

CI1 Was the study replicating existing studies or methods?
CI2 Was there correct reporting of manipulation apparatus, measurement apparatus, detailed procedure, sample size, demographics, sampling and recruitment method, contributing towards reproducibility?
CI3 Was there an explicit and operational specification of the RQs, null and alternative hypotheses, IVs, DVs, subject assignment method and manipulation checks?
CI4 Was there a discussion on the limitations, possible confounders, biases and assumptions made?
CI5 Was the result reported in the APA style?
CI6 Did the result statement include test statistic and p-value?
CI7 Were significance level α and test statistics properties and assumptions appropriately stated (e.g., "two-tailed")?
CI8 Were the appropriate the effect sizes and confidence intervals (CI) reported?
CI9 Was the significance and hypothesis testing decision interpreted correctly and put in context of effect size and sample size/power?

2 CI1: Upstream Replication

CI1 seeks to identify whether the reported study has replicated existing methods or studies as is or has adapted methods used in prior research. CI1 applies both to manipulation and measurement instruments.

2.1 Specification

Table 1: Criteria for CI1. ✓= Present, X = Absent.

#	Criteria	Success	Partial	Fail
1	Replicated existing methods as is	✓		X
2	Adapted existing methods		✓	X

2.2 Success Example

'Replicated existing methods as is': *"After participants completed the study, we asked them to fill an exit survey online, consisting of the 10-item IUIPC scale on privacy concerns [12] and an 8-item scale on privacy-protective behavior [14]"*, from Liu et al. [11].

2.3 Partial Example

'Adapted existing methods': *"Our earlier studies only tested the effectiveness of the training methodology when participants were trained once, but learning science literature suggests that if people are provided with more opportunities to learn, they tend to remember instructions better [5]. In PhishGuru, the simulated email works for both training and testing purposes; people who continue to click on the simulated phishing URLs can be presented with further training materials. Our goal was to investigate whether participants who read the training materials twice had any advantage over participants who read the training materials only once. [10]"*

3 CI2: Downstream Replication

CI1 aims to facilitate future replication of the study in question, that is if another researcher were to take the study report, he would be able to reproduce the study. Hence in this CI, we look for correct reporting of correct reporting of manipulation apparatus, measurement apparatus, detailed procedure, sample size, demographics, sampling and recruitment method. The coding criteria for success, partial success or failure of CI2 is given in Table 2.

3.1 Specification

Table 2: Criteria for CI2. ✓= Present, X = Absent.

#	Criteria	Success	Partial	Partial	Fail
1	Measurement and manipulation apparatus	✓	X		X
2	Detailed Procedure	✓			X
3	Sample Size	✓			
4	Demographics	✓		X	
5	Sampling and Recruitment	✓			

3.2 Success Example

For demographics, sampling source and sample size, as reported by Groß et al. [9]: *"The sample consisted of university students, $N = 100$, of which 50 were women. The mean age was 28.18 years ($SD = 5.241$) for the 83 participants who revealed their age. The participants were balanced by gender and assigned randomly to either the depletion ($n = 50$) or control ($n = 50$) condition. They were mostly non-computer science students from our University, of mainly international background (common countries included Oman, China and Iraq)."*

For description of replicated manipulation and measurement apparatus, see the examples in CI1. In addition, here's an example for measurement apparatus: "The State-Trait Anxiety Inventory for Adults (STAI-AD) [15] is a 40-question self-report questionnaire. We use the temporary construct of state anxiety, that is, "how you feel right now." It employs 4-point Likert items anchored on 1 – Not At All, 2 – Somewhat, 3 – Moderately So, and 4 – Very Much So."

For detailed procedure: *"The procedure consisted of (i) pre-task questionnaires for demographics and personality traits, (ii) a manipulation to induce cognitive depletion, (iii) a manipulation check on the level of depletion, (iv) a password entry for a mock-up GMail registration, and (v) a debriefing and memorability check one week after the task with a GMail login mockup. "* This was followed with a details of each section.

4 CI3: Specification & Operationalization of RQs & hypotheses

Operationalization enables systematic and explicit clarification of the predictors, IVs, and hence the cause and manipulation, while the target variable or DVs clarifies the effect, hence the measurements.

Subject assignment points to whether and how participants were randomly assigned and balanced across experimental conditions hence avoiding a bias and other possible explanations for between-subject designs. For within-subject studies, random assignment to manipulation sequences counters order effects. Manipulation check refers to verification that the manipulation has actually taken effect, hence lowering possible doubts that the observed effect did not emanate from the induced manipulation.

4.1 Specification

Table 3: Criteria for CI3. ✓= Present, X = Absent.

#	Criteria	Success	Partial	Partial	Partial	Fail
1	Research Question	✓				X
2	Hypotheses	✓			X	X
3	IVs and DVs	✓			X	X
4	Subject Assignment	✓	X			
5	Manipulation Check	✓		X		

4.2 Example

- For RQ: *"How availability of Touch ID sensor impacts users' selection of unlocking authentication secrets?"* from Cherapau et al. [4].
- For null hypotheses H_0: *"Use of Touch ID has no effect on the entropy of passcodes used for iPhone locking."* or *"Availability of Touch ID has no effect on ratio of users who lock their iPhones."*
- For corresponding alternative hypotheses H_1: *"Use of Touch ID affects the entropy of passcodes used for iPhone locking."* or *"Availability of Touch ID increases the ratio of users who lock their iPhones"*.
- For subject assignment: *"Our task scheduler presented the Captchas to Turkers in two different ways to make sure the order did not influence the results of the study (1) Random Order - Fully randomly, where any captcha from any scheme could follow any other. (2) Blocks of Three - In blocks of three Captchas from the same scheme, where the schemes were ordered randomly."*, such as from Bursztein et al. [3].
- For manipulation checks: *"In exit questions, participants confirmed that they felt they had participated in two separate studies and that it was unlikely that their responses from the first study could be linked to their responses from the second study"* from Adjerid et al. [1]

or *"We later used these photos to validate the claimed iPhone model (i.e., iPhone 4, 4S, 5S) and the locking mechanism. In addition, we also asked participants to provide us with the model number, e.g., ME302C/A,4 which has one-to-one correspondence with the marketed model, e.g., iPhone 5S"* from Cheparau et al. [4].

5 CI4: Limitations, biases & confounders

A discussion of the limits or boundaries of the experiment setup (such as missing qualitative self-reports questionnaire that back up quantitative analysis), identification of possible confounding variables whose presence affect the relationship under study, and possible assumptions made in setup, are all valuable inputs that strengthen the validity of the experiment.

5.1 Specification

Table 4: Criteria for CI4. ✓= Present, X = Absent.

#	Criteria	Success	Partial	Partial	Partial	Fail
1	Research Limitations	✓	X			X
2	Confounders	✓		X		X
3	Biases (sampling)	✓			X	X

5.2 Example

For limitations: *"Due to privacy constraints, we could not collect information about users' personal demographics or browsing habits. Consequently, we cannot measure whether user behavior differs based on personal characteristics, the target site, or the source of the link to the site. We also cannot identify SSL false positives due to captive portals, network proxies, or server misconfigurations"*, from Akhawe & Felt [2]. For sample bias [2]: *"The participants in our field study are not a random population sample. Our study only represents users who opt in to browser telemetry programs. This might present a bias. The users who volunteered might be more likely to click through dialogs and less concerned about privacy. Thus, the clickthrough rates we measure could be higher than population-wide rates."*

6 CI5: Standardized reporting

Reporting standards provide a degree of comprehensiveness in the information that is reported for empirical investigations. Uniform reporting standards make it easier to generalize within and across fields, to understand implications of individual studies and to allow for techniques of meta-analysis.

6.1 Specification

Table 5: Criteria for CI5. ✓= Present, X = Absent.

#	Criteria	Success	Partial	Fail
1	APA guidelines for all results	✓		X
2	APA guidelines for some results		✓	X

6.2 Example

"We computed an one-way ANOVA with the password strength score as dependent variable. There was a statistically significant effect of the experiment condition on the password strength score, $F(2,63) = 6.716$*,* $p = .002 < .05$*. We measure the effect size in Cohen's* $f = .42$ *from (*$\eta^2 = .176$*, 95% CI* $[0.043, 0.296]$*) and Cohen's* $\omega^2 = 0.148$*. This constitutes a large effect. [7]"*

7 CI6: Test statistic & p-value

This CI supports reproducibility of the analysis and foundations for research evidence and quality.

7.1 Specification

We provide the specifications for CI6in Table 6.

Table 6: Criteria for CI6. ✓= Present, X = Absent.

#	Criteria	Success	Partial	Fail
1	Actual $p-value$ reported	✓	X	X
2	Test-statistics reported	✓	✓	X
3	Mean & standard Dev. reported	✓		

7.2 Example

Same example as CI6 would be okay: *"We computed an one-way ANOVA with the password strength score as dependent variable. There was a statistically significant effect of the experiment condition on the password strength score,* $F(2,63) = 6.716$*,* $p = .002 < .05$*. We measure the effect size in Cohen's* $f = .42$ *from (*$\eta^2 = .176$*, 95% CI* $[0.043, 0.296]$*) and Cohen's* $\omega^2 = 0.148$*. This constitutes a large effect. [7]"*

8 CI7: Test statistics properties & assumptions

To ascertain whether the statistical analyses were correctly employed on the data, statistical assumptions need to be made explicit in reporting. For example, the assumptions for parametric tests, in general, are normally distributed data, homogeneity of variance, interval data and independence [8].

8.1 Specification

Table 7 shows our coding for partial or complete fulfillment of CI7 and for failure. The impact of CI7 is proof of appropriateness and correct deployment of the statistical methods used.

Table 7: Criteria for CI7. ✓= Present, X = Absent.

#	Criteria	Success	Partial	Partial	Partial	Fail
1	Significance level	Yes			No	No
2	Test assumptions	Yes		No		No
3	Test Properties	Yes	No			No

8.2 Example

Example for test properties is "one-tailed" or "two-tailed".

9 CI8: Effect size & confidence intervals

An effect that is statistically significant is not necessarily scientifically significant or important, where the importance of an effect is linked to the magnitude of the effect [5]. In addition, the APA makes reporting of confidence intervals a minimum standard.

9.1 Specification

Table 8: Criteria for CI8. ✓= Present, X = Absent.

#	Criteria	Success	Partial	Partial	Fail
1	Effect sizes for all results	✓			
2	Effect sizes for some results			✓	X
3	Confidence intervals	✓	✓		X

9.2 Example

"There was a statistically significant effect of the experiment condition on the password strength score, $F(2,63) = 6.716$*,* $p = .002 < .05$*. We measure the effect size in Cohen's* $f = .42$ *from (*$\eta^2 = .176$*, 95% CI* $[0.043, 0.296]$*) and Cohen's* $\omega^2 = 0.148$*. [7]"*

10 CI9: Null Hypothesis Significance Testing

Was the significance and hypothesis testing decision interpreted correctly and put in context of effect size and sample size/power? To facilitate correct interpretation, we develop a few recommendations, building from Nickerson [13] and the previous CIs. We postulate reporting of (a) p-value vis-a-vis significance level, (b) a-prior sample size computation, (c) Type I error correction, (d) explicit definition of null and alternative hypotheses, (e) and population specification.

10.1 Specification

We provide the specification for marking this CI as success, partial success or failure in Table 9.

Table 9: Criteria for CI9. ✓= Present, X = Absent.

#	Criteria	Success	Partial	Partial	Fail
1	p-value interpretation	✓	✓	✓	
2	A-priori sample specification	✓			X
3	Type I error correction	✓	✓		X
4	Null hypothesis specification	✓			X
5	Alternative hypothesis specification	✓		✓	X
6	Population specification	✓			X

10.2 Example

For Type I error correction: *"Given the number of comparative* t-tests *computed on the data set, we compute a multiple comparisons correction, where differences marked with a dagger † in Table 1 are statistically significant under Bonferroni-Holm correction for all comparisons made." [6]*

References

[1] ADJERID, I., ACQUISTI, A., BRANDIMARTE, L., AND LOEWENSTEIN, G. Sleights of privacy: Framing, disclosures, and the limits of transparency. In *Proceedings of the Ninth Symposium on Usable Privacy and Security* (2013), ACM, p. 9.

[2] AKHAWE, D., AND FELT, A. P. Alice in warningland: A large-scale field study of browser security warning effectiveness. In *USENIX security symposium* (2013), vol. 13.

[3] BURSZTEIN, E., BETHARD, S., FABRY, C., MITCHELL, J. C., AND JURAFSKY, D. How good are humans at solving captchas? a large scale evaluation. In *Security and Privacy (SP), 2010 IEEE Symposium on* (2010), IEEE, pp. 399–413.

[4] CHERAPAU, I., MUSLUKHOV, I., ASANKA, N., AND BEZNOSOV, K. On the impact of touch id on iphone passcodes. In *SOUPS* (2015), pp. 257–276.

[5] COHEN, J. A power primer. *Psychological bulletin 112*, 1 (1992), 155.

[6] COOPAMOOTOO, K. P., AND GROSS, T. Why privacy is all but forgotten - an empirical study of privacy and sharing attitude. *Proceedings on Privacy Enhancing Technologies 4* (2017), 39–60.

[7] COOPAMOOTOO, K. P., GROSS, T., AND PRATAMA, MUHAMMAD, F. An empirical investigation of security fatigue - the case of password choice after solving a captcha. 2017.

[8] FIELD, A. *Discovering statistics using IBM SPSS statistics*. Sage, 2013.

[9] GROSS, T., COOPAMOOTOO, K., AND AL-JABRI, A. Effect of cognitive depletion on password choice. In *The LASER Workshop: Learning from Authoritative Security Experiment Results (LASER 2016)* (2016), USENIX Association, pp. 55–66.

[10] KUMARAGURU, P., CRANSHAW, J., ACQUISTI, A., CRANOR, L., HONG, J., BLAIR, M. A., AND PHAM, T. School of phish: a real-world evaluation of anti-phishing training. In *Proceedings of the 5th Symposium on Usable Privacy and Security* (2009), ACM, p. 3.

[11] LIU, B., ANDERSEN, M. S., SCHAUB, F., ALMUHIMEDI, H., ZHANG, S., SADEH, N., ACQUISTI, A., AND AGARWAL, Y. Follow my recommendations: A personalized privacy assistant for mobile app permissions. In *Symposium on Usable Privacy and Security* (2016).

[12] MALHOTRA, N. K., KIM, S. S., AND AGARWAL, J. Internet users' information privacy concerns (iuipc): The construct, the scale, and a causal model. *Information systems research 15*, 4 (2004), 336–355.

[13] NICKERSON, R. S. Null hypothesis significance testing: a review of an old and continuing controversy. *Psychological methods 5*, 2 (2000), 241.

[14] PEW RESEARCH CENTER. Internet project/gfk privacy panel. 2014.

[15] SPIELBERGER, C. D., GORSUCH, R. L., LUSHENE, R., VAGG, P. R., AND JACOBS, G. *State-trait anxiety inventory for adults*. Mind Garden, 1983.

Codesheet for The Nifty Nine Completeness Indicators
Coding for Evidence-Based Research
v1.1

Kovila P.L. Coopamootoo
Newcastle University

Thomas Groß
Newcastle University

1 Instructions

This coding sheet is meant to evaluate easily checkable Completeness Indicators CIs for evidence-based research papers. For each of the 9 CIs, code the research article across each of the criteria:

- for each of the criteria, mark it as 'Present' with a ✓or 'Absent' with an X,
- use the codebook specification for the CI to make a decision on succeeding, partially succeeding or failing the CI,
- mark the CI as success, partial success or failure.

2 CI1 Criteria

2.1 Specification

Was the study replicating existing studies/methods?

Table 1: Criteria for CI1. Mark ✓for 'present,' and X for 'absent'.

#	criteria	Marking
1	Replicated existing methods as is	☐
2	Adapted existing methods	☐

2.2 Marking

Overall Marking for CI1: ______.

3 CI2 Criteria

3.1 Specification

Was there correct reporting of manipulation apparatus, measurement apparatus, detailed procedure, sample size, demographics, sampling and recruitment method, contributing towards reproducibility?

Table 2: Criteria for CI2. Mark ✓for 'present,' and X for 'absent.'

#	Criteria	Marking
1	Measurement and manipulation apparatus	☐
2	Detailed Procedure	☐
3	Sample Size	☐
4	Demographics	☐
5	Sampling and Recruitment	☐

3.2 Marking

Overall Marking for CI2: ______.

4 CI3 Criteria

4.1 Specification

Was there an explicit and operational specification of the RQs, null and alternative hypotheses, IVs, DVs, subject assignment method and manipulation checks?

Table 3: Criteria for CI3. Mark ✓for 'present,' and X for 'absent.'

#	Criteria	Marking
1	Research Question	☐
2	Hypotheses	☐
3	IVs and DVs	☐
4	Subject Assignment	☐
5	Manipulation Check	☐

4.2 Marking

Overall Marking for CI3: ______.

5 CI4 Criteria

5.1 Specification

Was there a discussion on the limitations, possible confounders, biases and assumptions made?

Table 4: Criteria for CI4. Mark ✓for 'present,' and X for 'absent.'

#	Criteria	Marking
1	Research Limitations	☐
2	Confounders	☐
3	Biases (sampling)	☐

5.2 Marking

Overall Marking for CI4: ______.

6 CI5 Criteria

6.1 Specification

Was the result reported in the APA style?

Table 5: Criteria for CI5. Mark ✓for present, and X for 'Absent'.

#	Criteria	Marking
1	APA guidelines for all results	☐
2	APA guidelines for some results	☐

6.2 Marking

Overall Marking for CI5: ______.

7 CI6 Criteria

7.1 Specification

Did the result statement include test statistic and p-value?

Table 6: Criteria for CI6. Mark ✓for present, and X for 'Absent'.

#	Criteria	Making
1	Actual $p-value$ reported	☐
2	Test-statistics reported	☐
3	Mean & standard Dev. reported	☐

7.2 Marking

Overall Marking for CI6: ______.

8 CI7 Criteria

Were significance level α and test statistics properties and assumptions appropriately stated (e.g., "two-tailed")?

8.1 Specification

Table 7: Criteria for CI7. Mark ✓ for present, and X for 'Absent'.

#	Criteria	Marking
1	Significance level	☐
2	Test assumptions	☐
3	Test Properties	☐

8.2 Marking

Overall Marking for CI7: ______.

9 CI8 Criteria

9.1 Specification

Were the appropriate the effect sizes and confidence intervals (CI) reported?

Table 8: Criteria for CI8. Mark ✓ for present, and X for 'Absent'.

#	Criteria	Marking
1	Effect sizes for all results	☐
2	Effect sizes for some results	☐
3	Confidence intervals	☐

9.2 Marking

Overall Marking for CI8: ______.

10 CI9 Criteria

10.1 Specification

Was the significance and hypothesis testing decision interpreted correctly and put in context of effect size and sample size/power?

Table 9: Criteria for CI9. Mark ✓ for present, and X for 'Absent'.

#	Criteria	Marking
1	p-value interpretation	☐
2	A-priori sample specification	☐
3	Type I error correction	☐
4	Null hypothesis specification	☐
5	Alternative hypothesis specification	☐
6	Population specification	☐

10.2 Marking

Overall Marking for CI9: ______.