Human-Computer Interaction

Research Reporting & Writing + Open Science

Professor Bilge Mutlu

Today's Agenda

- **Why reporting is a credibility problem:** the reading's case
- **Writing it up:** structure, one claim per section
- **Reporting results:** effect size + CI, not a bare *p*
- **Open science:** pre-registration, open data and materials
- **In class (30 min):** skeleton, report, release

Today's readings

Required

Cockburn, Dragicevic, Besançon & Gutwin (2020). *Threats of a Replication Crisis in Empirical Computer Science.* CACM 63(8).

The case for today, in CS terms.

Recommended

Dragicevic (2016). *Fair Statistical Communication in HCI.* Section 4.

How to report effect sizes with CIs in practice.

Why reporting is a *credibility* problem

Is CS research at risk? *(Spoiler: yes)*[1]

The authors read the 10 most-downloaded articles from each of 41 ACM *Transactions* journals: 410 articles.

- **61 (15%)** used *p* as a yes/no criterion, across **21 journals (51%)**
- **59 of those 61** reported at least one $p < .05$: a sign that null results go unpublished
- Not only HCI: graphics, software engineering, AI, performance analysis

Elsewhere: cancer research could not reproduce 47 of 53 landmark findings.

[1] Cockburn, Dragicevic, Besançon & Gutwin (2020). Threats of a Replication Crisis in Empirical Computer Science. *CACM* 63(8), 70–79. doi:10.1145/3360311

How a result goes wrong[1]

Publication bias (nulls go in the file drawer) invites:

- **HARKing:** hypothesizing after the results are known
- **Mid-study changes:** adding participants until $p < .05$ (56% of psychologists admit it)
- **Flexible analysis:** dropping outliers, transforming, testing subsets

Test 15 measures when nothing has an effect: a 54% chance one reaches $p < .05$ anyway.

[1] Cockburn, Dragicevic, Besançon & Gutwin (2020). Threats of a Replication Crisis in Empirical Computer Science. *CACM* 63(8), 70–79. doi:10.1145/3360311

What the reading recommends for CS[1]

Better analysis

- Report **effect sizes** with **interval estimates**
- Name the test and its df; give the **exact p-value**
- Help readers judge for themselves

More openness

- **Pre-register** (e.g., on OSF)
- **Registered reports:** reviewed before data exist
- **Data and materials** released
- Encourage **replications**

[1] Cockburn, Dragicevic, Besançon & Gutwin (2020). Threats of a Replication Crisis in Empirical Computer Science. *CACM* 63(8), 70–79. doi:10.1145/3360311

Writing it up

What are reporting norms in HCI research?

Because HCI is interdisciplinary, reporting norms are borrowed from several fields.

Aspect	**Norm**
Paper structure	APA (loosely)
Results of statistical analyses	APA (strictly)
Tables, figures	APA (very loosely)
Citations	Depends on publisher (ACM, IEEE, …)
Style	APA (loosely); high writing standards

How is an HCI paper structured?

- Abstract
- Introduction
- Related Work / Background
- *Hypotheses (quant. empirical)*
- *System / Design (design-based)*
- Method
- Results
- Discussion
- Conclusion
- Acknowledgements
- References
- Appendices

What is an abstract?[2]

A brief but complete summary; check the venue's word limit. One or two sentences each:

- Literature summary
- Problem / RQs
- Hypotheses
- Methods
- Results
- Implications

[2] APA Style: Abstract & Keywords Guide. https://apastyle.apa.org/instructional-aids/abstract-keywords-guide.pdf

Each section makes *one* claim

Before you write, name the single claim each section has to make.

- **Intro:** the problem and why it matters *now*
- **Related work:** the gap only you fill
- **Method:** why this design answers the RQ
- **Results:** the one finding that carries the paper
- **Discussion:** what it means, and what it does *not* show

Example: a *structured* abstract[5]

Background: CBT apps work, but adherence is poor; conversational agents may be more engaging.

Objective: feasibility, acceptability, and early efficacy of Woebot for students with anxiety and depression symptoms.

Methods: unblinded RCT; 70 students, 2 weeks of Woebot (n = 34) vs. an NIMH ebook (n = 36); symptom scales before and after.

Results: Woebot users reduced depression symptoms (F = 6.47; P = .01); the ebook group did not. Both groups reduced anxiety.

Conclusions: Conversational agents appear to be a feasible, engaging, and effective way to deliver CBT.

Your skeleton looks like this: one labeled claim per section.

[5] Fitzpatrick, Darcy & Vierhile (2017). Delivering CBT ... Using a Fully Automated Conversational Agent (Woebot): A Randomized Controlled Trial. *JMIR Mental Health* 4(2), e19. doi:10.2196/mental.7785

Reporting results

How do we report statistics?

Descriptive: distribution via summary statistics, in text, tables, graphs.

Inferential: test parameters and results in text or tables; significance highlighted in graphs.

In *text*, APA is followed **strictly**; in *graphs*, you can be creative.

Report the *size*, not just the *p*

A bare *p*-value tells you *whether*, not *how much*.

Convey effect sizes, and "the uncertainty around those effect sizes by reporting interval estimates."[1]

- $p < .05$ alone can't tell a trivial effect from a meaningful one
- The CI shows precision; the effect size shows practical stakes
- "Human + AI beats human alone" (Nov 30 crit) needs both: by how much, and how sure

[1] Cockburn, Dragicevic, Besançon & Gutwin (2020). Threats of a Replication Crisis in Empirical Computer Science. *CACM* 63(8), 70–79. doi:10.1145/3360311

Fair statistical communication[3]

Communicate "in recognition of the right of other free minds to utilize them in making their own decisions." *(Fisher, 1955)*

- **Clarity:** a reader can't judge what they can't follow
- **Transparency:** state every analysis decision
- **Simplicity:** prefer the simpler analysis
- **Noncontingency:** no "if the data look like this, report that"

[3] Dragicevic (2016). Fair Statistical Communication in HCI. In Robertson & Kaptein (eds.), *Modern Statistical Methods for HCI*, Springer, 291–330. Author PDF: dragice.fr/fairstats-last.pdf

Reporting conventions to get right

- **Name the test with its df:** $t(182.3) = 2.33$
- **Exact *p*:** $p = .021$, not $p < .05$; no leading zero (APA)
- **Descriptives** next to every inferential result
- **Estimate first, then the test**

"Stories written with five AI ideas were rated 0.31 points more novel than stories written alone, 95% CI [0.05, 0.58], d = 0.34; t(182.3) = 2.33, p = .021."

Open science

What is open science?

Making the research process **inspectable and reusable**, not just the paper readable.

- **Pre-registration:** commit to hypotheses and an analysis plan *before* data collection
- **Open data and materials:** release what lets others reproduce your work
- **Open analysis code:** the pipeline from data to figure

Why pre-register?[4]

It separates **confirmatory** from **exploratory** analysis, the guard against *p*-hacking and HARKing.

- **Medicine** has required it for years: in one funder's trials, significant results fell from **57% to 8%** once it was mandatory[1]
- **Exploring is still allowed**, labeled: "In addition to the pre-registered analysis, we also ran…"[1]
- **A plan, not a prison:** report deviations

---

[4] Nosek, Ebersole, DeHaven & Mellor (2018). The Preregistration Revolution. *PNAS* 115(11), 2600–2606. doi:10.1073/pnas.1708274114

[1] Cockburn, Dragicevic, Besançon & Gutwin (2020). Threats of a Replication Crisis in Empirical Computer Science. *CACM* 63(8), 70–79. doi:10.1145/3360311

What you can commit to releasing

Name it now; it shapes how you run the study.

- **Data:** de-identified; does your consent form cover sharing?
- **Materials:** stimuli, instruments, prompts
- **Code:** analysis scripts, model configs
- If you *can't* release something, say **why** (privacy, IRB, licensing)

In class: Skeleton, Report, Release *(30 min)*

One copy per team (File → Make a copy):

1. **Skeleton:** one claim per section, and the figure that carries the paper
2. **Report:** one result in full, effect size + CI, checked against the conventions
3. **Release:** what you would pre-register, and what you can release

da.gd/qdVxGo



No result of your own?

Report one from your **practice set** (A or B) and compare it with the paper's. Pre-register as if you were about to run that study.

da.gd/d8BEE



Assignment: *Reporting & Writing Deliverable*

Produce, for your study:

1. **Paper skeleton:** full section structure, one-line claim per section
2. **Results reporting:** at least one result with effect size + CI, and the conventions you will follow
3. **Open-science plan:** pre-registration draft (or rationale) and a data / materials / code release statement

Deliverable: about 1–2 pp, individual, on Canvas. **Due Fri Dec 4, 9:30 am.**

What's Next?

- **Mon Nov 23 · Critical, Responsible & Sustainable AI:** read **Falco, Shneiderman et al.** (audits) and **Birhane et al.** (values); post your provocation by 9:30
- **Fri Nov 27:** Thanksgiving, no class
- **Mon Nov 30 · Studio Crit:** near-final results and the does-it-help-the-human gate
- **Fri Dec 4 · Writing & presentation clinic:** Reporting & Writing Deliverable due 9:30