Human-Computer Interaction

Analysis/Results Crit

Professor Bilge Mutlu

Run-of-show (75 min)

- **0:00 · Frame:** today's method *is* today's crit lens
- **0:03 · Teaching (17 min):** what makes a result credible
- **0:20 · Set up:** tables, roles, the form, the timer
- **0:23 · Crit (42 min):** each team gets a full turn
- **1:05 · Debrief (10 min):** sharpest issues, patterns across tables

Today's method *is* today's crit lens

A focused 17 minutes on **what makes a result credible**, then you apply it as a reviewer on real results: your classmates'.

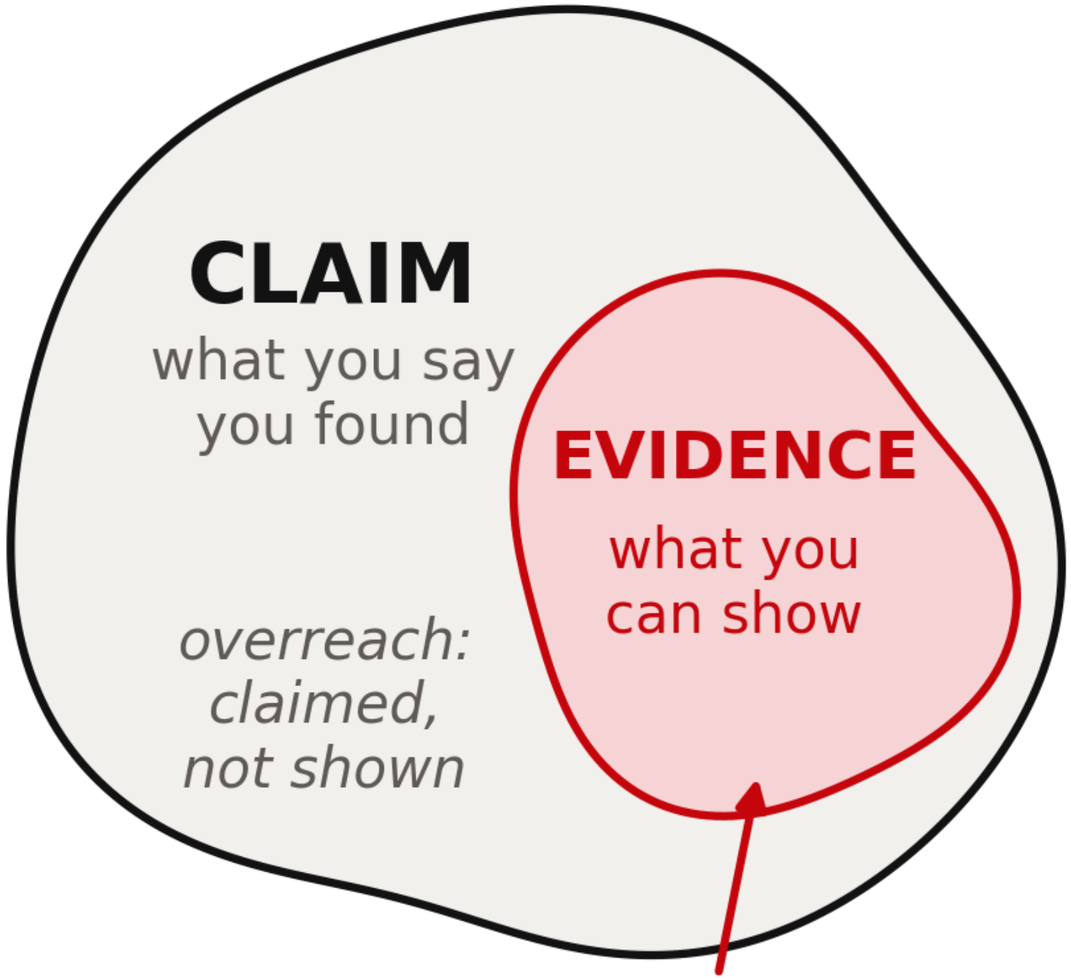
Bring: your team's results so far. That is the artifact you crit.

Teaching block: *what makes a result credible*

A result is a *claim* plus *evidence*

- **Claim:** what you say you found
- **Evidence:** what you can show for it
- **Credible:** the evidence supports *this* claim, and no more

Most common crit finding: a claim bigger than its evidence.



Credibility lives here: the claim the evidence covers

(a) Numbers

- **Right test** for the design and the variable types
- **Assumptions** checked, or a robust alternative
- **Effect size + CI** next to every p-value
- **Claim matches the test:** no cause from a correlation; no "no effect" from p > .05

Crit question: what would a reader need to believe this number?

(b) Codes and themes

- **Traceable:** each theme backed by excerpts from more than one participant
- **Agreement:** two coders on a shared segment; κ, or how disagreements were resolved
- **Negative cases:** where the theme does *not* hold
- **Counts with care:** "most participants" is a number; give it

Crit question: could a reader find this theme in your data?

(c) Design findings

- **Observed, not intended:** what people did, not what the design was meant to do
- **How many** sessions, and with whom
- **Fidelity matches the claim:** a Wizard-of-Oz study tests the *experience*, not a system that doesn't exist yet
- **What it teaches** beyond this prototype

Crit question: is this a finding, or the design rationale restated?

(d) When the result is an AI system's output

- **Human evaluation:** rating protocol fixed *before* looking at outputs; raters blind to condition
- **Agreement, not averages:** Cohen's κ (two raters) · Krippendorff's α (more raters, missing data)
- **Subgroups:** results by type of user or input (e.g., novices vs. experts, easy vs. hard tasks), not one overall number

Crit question: who rated, on what, and how well did they agree?

Does it help the human?

A claim of "help" needs a comparison:

- **Three arms:** human alone · AI alone · human + AI
- **After assistance:** unaided performance once the AI is removed
- **Outcome and process:** task success **plus** effort, trust calibration, agency

If human + AI underperforms a solo arm, that's the finding.

The results crit card

Today's teaching in four lines; it's on your crit form:

- **Evidence ↔ claim:** does the evidence support the claim as stated?
- **Analysis:** done right for this kind of data
- **Credibility checks:** agreement, negative cases, subgroups, baselines
- **Next step:** what's left, and doable before the Nov 30 Studio Crit

One card, three kinds of study

	Experiment / survey	Field / qualitative	Design-led
Evidence ↔ claim	Claim matches the test	Theme matches the excerpts	Finding matches what was observed
Analysis	Right test · assumptions · effect size + CI	Codebook · themes traceable to data	How sessions became findings
Credibility checks	Baseline arm · subgroups	Coder agreement · negative cases	Wizard consistency · fidelity vs. claim
Next step	Data or tests still to run	Coding still to do	Sessions or analysis still to run

Set up the crit

Present your results *so far*

Your crit question: **the claim you're least sure of**

Experiment / survey

Preliminary results, or one member's **Inferential Analysis Plan** if data are still coming in

Field / qual

Your **codes and emerging themes**

Design-led

What your **sessions** showed, with the evidence for each finding

Tables, roles, timer

- **Your table is your crit group:** your team and the other team at it
- **A team on its own** joins a neighboring table, making three
- **Roles** (rotate from Oct 16):
  - **Facilitator:** runs the round-robin, makes sure everyone speaks, reports at debrief
  - **Timekeeper:** backstops the timer
  - **Scribe:** captures feedback on the form

The per-team cycle

Phase	2-team table	3-team table	What happens
Present	6 min	4 min	Your results + the claim you want crit on
Silent writing	3 min	2 min	Each reviewer writes on the form
Round-robin	9 min	6 min	One strength + one concern each
Respond & commit	3 min	2 min	Name **one change** you'll make

Two turns of 21 min, or three of 14. Both end at the same time.

Feedback etiquette

Giving: describe, don't judge ("this is hard to interpret because…") · ask questions ("what made you choose that test?") · build on ideas · be kind and specific · focus on the research goal, not preference.

Receiving: listen without defensiveness · ask follow-ups · it's about the work, not you.

Capture it on the form

The written record is the deliverable, not feedback that evaporates.

- **One form per table:** File → Make a copy
- One section per presenting team, ending in its **committed revision**
- Each team submits its section on Canvas **by 11:59 pm tonight**

da.gd/LKWWGg



Debrief (10 min)

Debrief

- **Facilitators** surface the sharpest issue their table saw
- Patterns across tables, tied back to the teaching block
- Usual suspects: *claims bigger than the evidence · p-values without effect sizes · themes without excerpts · findings that restate the design*

What's Next?

- **Tonight, 11:59 pm:** your team's crit section + committed revision, on Canvas
- **Mon Nov 16 · Privacy, Security & Surveillance:** read Nissenbaum (2004), Mireshghallah et al. (2024); post your provocation by 9:30
- **Fri Nov 20 · Reporting & Writing:** team status by 9:30
- **Mon Nov 30 · Studio Crit:** near-final results and the does-it-help-the-human gate