Human-Computer Interaction

Method-Design 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):** measures · scales · power
- **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 study design sound**, then you apply it as a reviewer on real designs: your classmates'.

Bring: the method your team will run. That is the artifact you crit.

Teaching block: *measures • scales • power*

(a) Choosing measures

Four kinds; match the **measure** to the **construct**:

1. **Objective:** against a standard (task success, time)
2. **Behavioral:** observed actions (edits, accepting AI suggestions, reliance)
3. **Subjective:** self-report (trust, satisfaction, ownership)
4. **Physiological:** from the body (eye tracking, GSR, EEG)

Example measures: *Tullis & Albert*[1]

Ch.	Metrics	Examples
4	Performance	task success, time on task, errors, efficiency, learnability
5	Issue-based	usability issues, severity ratings
6	Self-reported	rating scales, post-task and post-session ratings, SUS
7	Behavioral & physiological	verbal expressions, eye tracking, emotion, stress
8	Combined & comparative	single usability scores, scorecards, comparison to goals

[1] Tullis, T. & Albert, B. (2013). *Measuring the User Experience* (2nd ed.). Morgan Kaufmann. UW Library e-book

Measures must be *valid*

Are you measuring what you claim?

- **Construct:** captures the intended concept (convergent + discriminant)
- **Content:** covers all relevant facets
- **Criterion:** relates to or predicts an established measure
- **Ecological:** generalizes to the real setting

Most common crit finding: a measure that misses the construct.

THE STUDY
a 30-minute lab task, practice stakes

CONSTRUCT
can't be observed directly

Trust in the AI assistant

Performance
is it right?

Process
can I see how it works?

Purpose
is it meant to help me?

1 Construct
does the score track trust?

MEASURE
Trust questionnaire

- **Q1** It is accurate
- **Q2** It is reliable
- **Q3** It rarely errs
- **Q4** I see how it decides

3 Criterion
does the score predict what people do?

BEHAVIOR
Reliance
follows it when it's right, overrides it when wrong

2 Content
do the items cover every facet?
No item asks about Purpose.

Another trust scale
should agree

Liking technology
shouldn't

1 convergent

1 discriminant

Trust facets: Lee & See (2004)

REAL USE
at work, real stakes, months of use

BEHAVIOR
Reliance at work
when a wrong answer costs something

4 Ecological
does what the study found still hold here?

(b) Scale development

A **scale** measures a construct through a set of **items**.

- **Construct:** the concept (trust, usability)
- **Item:** one question capturing part of it
- **Likert:** agreement with statements; combine items into one score
- Prefer an **existing validated scale** (SUS, UEQ, trust in automation[2]) over a home-grown one

[2] Jian, J.-Y., Bisantz, A. M. & Drury, C. G. (2000). Foundations for an empirically determined scale of trust in automated systems. *International Journal of Cognitive Ergonomics*, 4(1), 53–71.

TRUST IN AUTOMATED SYSTEMS · JIAN, BISANTZ & DRURY (2000)

Rate how much each statement describes your experience with the system.

	not at all 1	2	3	4	5	6	extremely 7
1 The system is deceptive	○	○	○	○	○	○	○
2 The system behaves in an underhanded manner	○	○	○	○	○	○	○
3 I am suspicious of the system's intent, action, or outputs	○	○	○	○	○	○	○
4 I am wary of the system	○	○	○	○	○	○	○
5 The system's actions will have a harmful or injurious outcome	○	○	○	○	○	○	○
6 I am confident in the system	○	○	○	○	○	○	○
7 The system provides security	○	○	○	○	○	○	○
8 The system has integrity	○	○	○	○	○	○	○
9 The system is dependable	○	○	○	○	○	○	○
10 The system is reliable	○	○	○	○	○	○	○
11 I can trust the system	○	○	○	○	○	○	○
12 I am familiar with the system	○	○	○	○	○	○	○

TRUST IN AUTOMATED SYSTEMS · JIAN, BISANTZ & DRURY (2000)

Rate how much each statement describes your experience with the system.

	not at all 1	2	3	4	5	6	extremely 7
1 The system is deceptive	○	○	○	○	○	○	○
2 The system behaves in an underhanded manner	○	○	○	○	○	○	○
3 I am suspicious of the system's intent, action, or outputs	○	○	○	○	○	○	○
4 I am wary of the system	○	○	○	○	○	○	○
5 The system's actions will have a harmful or injurious outcome	○	○	○	○	○	○	○
6 I am confident in the system	○	○	○	○	○	○	○
7 The system provides security	○	○	○	○	○	○	○
8 The system has integrity	○	○	○	○	○	○	○
9 The system is dependable	○	○	○	○	○	○	○
10 The system is reliable	○	○	○	○	○	○	○
11 I can trust the system	○	○	○	○	○	○	○
12 I am familiar with the system	○	○	○	○	○	○	○

Response format
7-point Likert-type rating

Subscale: distrust
items 1–5, negatively worded: reverse-key them (8 − x)

Subscale: trust
items 6–12

Item
one statement, one rating

Scale
all 12 items → one trust score (mean, after reverse-keying)

Scales must be *reliable*

- **Cronbach's alpha:** do the items move together? Runs from 0 to 1
- Reliable isn't valid: consistent yet wrong
- **Building your own scale?**
  1. Define the construct and its facets
  2. Write items for each facet
  3. Factor analysis: do items group as intended?
  4. Alpha for each group: do its items agree?

READING CRONBACH'S ALPHA

.90–1.00 Excellent internal consistency	EXCELLENT
.80–.89 Good internal consistency	GOOD
.70–.79 Acceptable for research	ACCEPTABLE
.60–.69 Questionable: use with caution	QUESTIONABLE
below .60 Poor: revise the scale	POOR

Rule of thumb after George & Mallery (2003)

(c) Power & sample size

Can your design **detect** the effect it's looking for?

- **Effect size:** how big is the effect? (Cohen's *d*: .2 / .5 / .8)
- **Significance level, alpha:** the false-positive rate you accept (usually .05); not Cronbach's alpha
- **Power:** chance of detecting a real effect (target **.80**)
- Together they fix the **required sample size**

From effect size to *n*

1. Decide the **smallest effect that would matter**, in your measure's units, then convert it to *d* (difference ÷ SD)
2. Set **alpha = .05**, **power = .80**
3. Compute the **required sample size**, e.g., in $G^*\text{Power}$[3]
4. Check it against your **feasibility**

Need more people than you can recruit? Fix the design now.

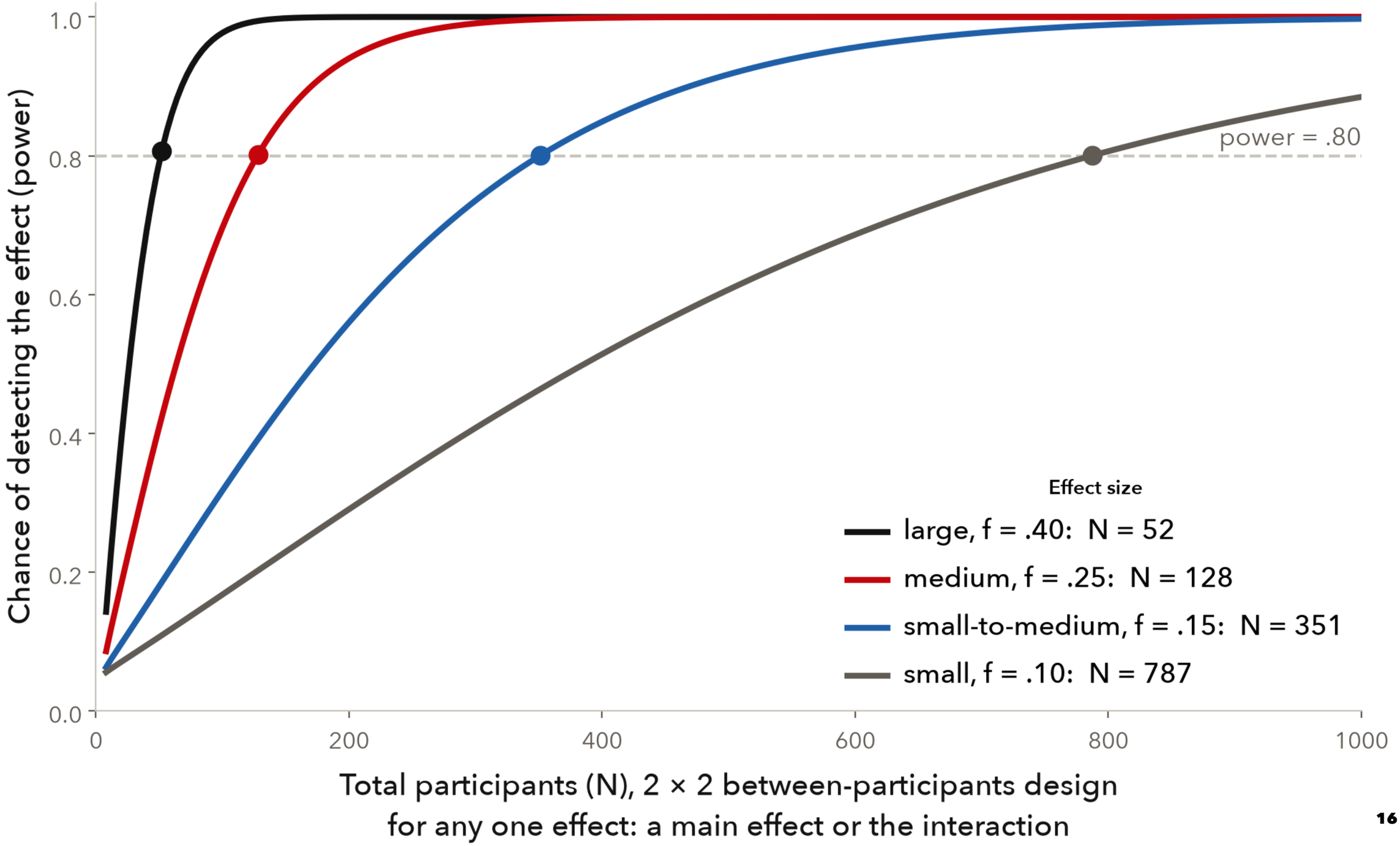
[3] Faul, Erdfelder, Lang & Buchner (2007), *Behavior Research Methods*, 39(2), 175–191. Free download

Worked example: *AI timing* × *experience*

2 × 2 between participants · alpha = .05 · power = .80

- **Timing main effect**, medium (*f* = .25): **N = 128**, 32 per cell
- **Interaction**, smaller (*f* = .15): **N = 351**
- **Any effect**, small (*f* = .10): **N = 787**

Same calculation for each: 1 df, 4 groups; only *f* changes. Interactions are usually smaller, so they need far more people.


Chance of detecting the effect (power)
1.0
0.8
0.6
0.4
0.2
0.0
power = .80
Effect size
large, f = .40: N = 52
medium, f = .25: N = 128
small-to-medium, f = .15: N = 351
small, f = .10: N = 787
0
200
400
600
800
1000
Total participants (N), 2 × 2 between-participants design
for any one effect: a main effect or the interaction

The research crit card

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

- **RQ ↔ method fit:** does the design answer the question?
- **Validity:** internal (confounds) · external (generalization) · construct (measuring what you claim)
- **Measures & power:** right measures for the construct; enough *n* to detect the effect
- **Feasibility:** doable in the time and resources left

One card, three kinds of study

	Experiment	Field / qualitative	Design-led
RQ ↔ method fit	Does the design answer the RQ?	Do these data answer the RQ?	Does building and studying it answer the RQ, and teach something new?
Validity	Internal · external · construct	Credibility · transferability	Prototype embodies the idea · defensible seam
Measures & power	Right measures · enough *n*	Right data sources · why this sample is enough	What counts as data · *n* if comparing conditions
Feasibility	Time, people	+ access, IRB	+ build time, wizard, debrief

Set up the crit

Present the method you'll *run*

Your crit question: **where your drafts or plans differ**

Experiment	**Field / qual**	**Design-led**
One member's **Step-by-Step design**	Your **Qualitative Field-Method Plan**	Your **Design-led / Wizard-of-Oz Study Plan**

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 each crit):
  - **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 method + the question 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 design?") · 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/bosB9w



Debrief (10 min)

Debrief

- **Facilitators** surface the sharpest issue their table saw
- Patterns across tables, tied back to the teaching block
- Usual suspects: *underpowered designs · measures that don't match the construct · RQ ↔ method mismatch*

What's Next?

- **Tonight, 11:59 pm:** your team's crit section + committed revision, on Canvas
- **Mon Oct 19 · Ubiquitous, Mobile & Tangible:** read Weiser (1991), Dourish (2004); post your provocation by 9:30
- **Fri Oct 23 · Qualitative Analysis:** read Muller (2014), Xiao et al. (2023)
- **Project:** finalize your method by ~Oct 31