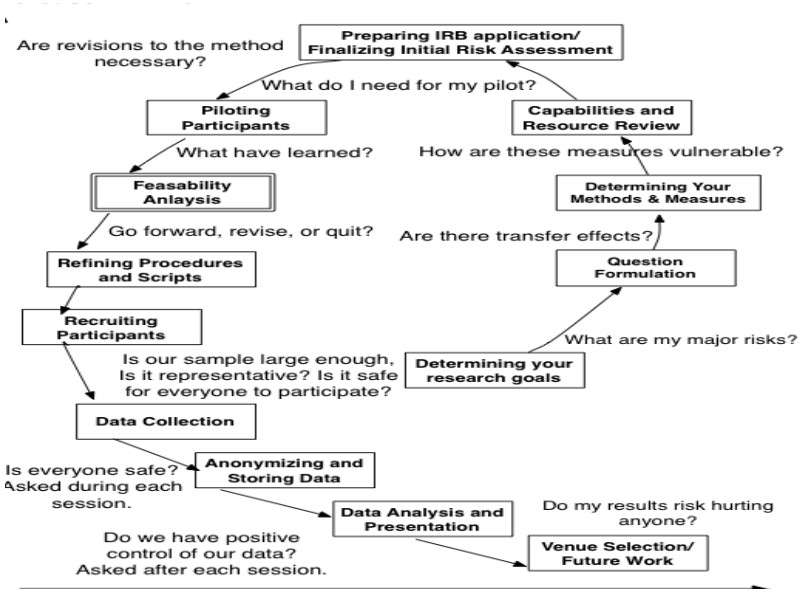
A Risk-Driven Approach to Experimental Design and Practice

Frank E. Ritter BSEE, PhD C. Psychol.
frank.ritter@psu.edu
ACS Lab, The College of IST, Penn State

Jonathan H. Morgan PhD (slides and book)
Duke Sociology
jhm18 @duke.edu

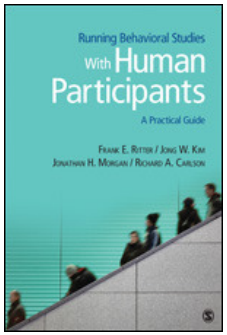
Jong W. Kim PhD and Richard Carlson PhD (book)
ACS Lab, Penn State || Psychology, Penn State



Supported by ONR, Contracts N00014-11-1-0275, N00014-10-1-0401, & W911QY-07-01-0004

If you will teach this or want to know more....

- Full book available from Sage
  [paper & PDF]



- Slides available as ppt or pdf
  (email us, or see RBS site)
- Tech report available as shorter version
  http://acs.ist.psu.edu/reports/ritterKM09.pdf
- Workbook and other resources available at
  http://frankritter.com/rbs

Overview

📖 acs.ist.psu.edu/papers

1200-1215 (0) 3 Orientation

1215-1250 (1) 4-10 An overview of risk-driven experimental design

1250-1310 (2) 11-14 Preparation for running an experiment

1310-1330 [20 min. break]

1330-1350 (3) 15-21 Ethical challenges in the experimental process

1350-1415 (4) 22 Risks to validity, with class participation

1415-1435 [20 min. break]

1435-1450 (5) 23-24 Running a session

1450-1505 (6) 25-27 Concluding an experiment, reporting results

1505-1540 (7) 28 Concluding comments and questions

1540-1560 Slack (running online? A recent update)

What to get out of this Tutorial

1) Some feeling for how to run a study
   - Cognitive science may be modeling + data
     So, to use data you have to know how it was gathered
   - Modeling is slow, so data publication helps modelers
   - If you are a computer scientist/engineer, you won't have taste in this area
     => Help you develop a green thumb for HCI/HF usability studies
   - Not how to *design* a study, but related
2) Some tools to help you set up a study
3) Materials

   Book (Sage) and 🕮draft book (tech report)
   [please let me know if you use it for a class!]

   Example problems, slides, book has exam questions
4) A greater appreciation for mistakes to avoid and a theory of how to avoid them, and thus problems in studies

Who are you?

1) We will put you into breakout rooms in pairs for intro, learning Zoom, and later exercise (use breakout room for 3 min.),
   any questions first?
2) Talk about: Name, organization, background, number of studies, what you want to get from this
3) 20 sec. Intro (timed): name, where, interest, question (or chat us the question)
4) We will use stack process for chat questions

Summary 1 of tutorial:

(Re)Looking at failure: What constitutes a failure/risk?

- Someone got hurt
- After committing significant resources, the study was never completed
- We have learned nothing new because our data is not repeatable or generalizable
- We have failed to communicate our results or their significance to anyone else

Ch 1. Overview
Some Terms used (09:10)

A study, varying an Independent variable (IV, e.g., amount of practice), to see the effect on a dependent variable (DV), such as response time

- Worth reading a methods book(s) (e.g., 🕮Ray; Calfe)

Subjects (Ss) or Participants (Ps), Users, learners, students, Experimenters (Es)

See *APA manual* and also 🕮Roediger (2006) for arguments for S and P and E/L/S

Example studies in the book:

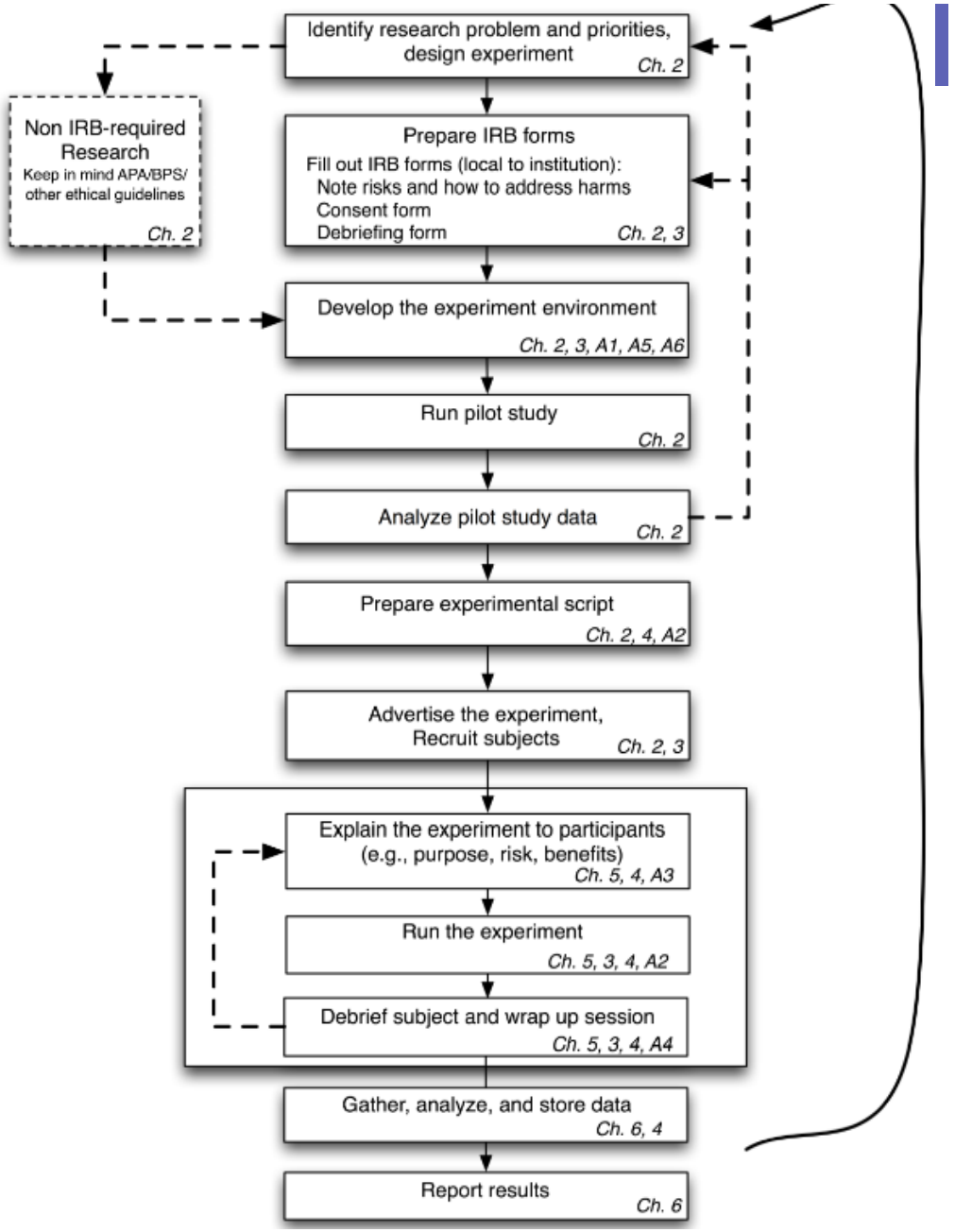
Multi-lingual fonts

Partially sighted and blind users

Human-robot interaction (HRI)

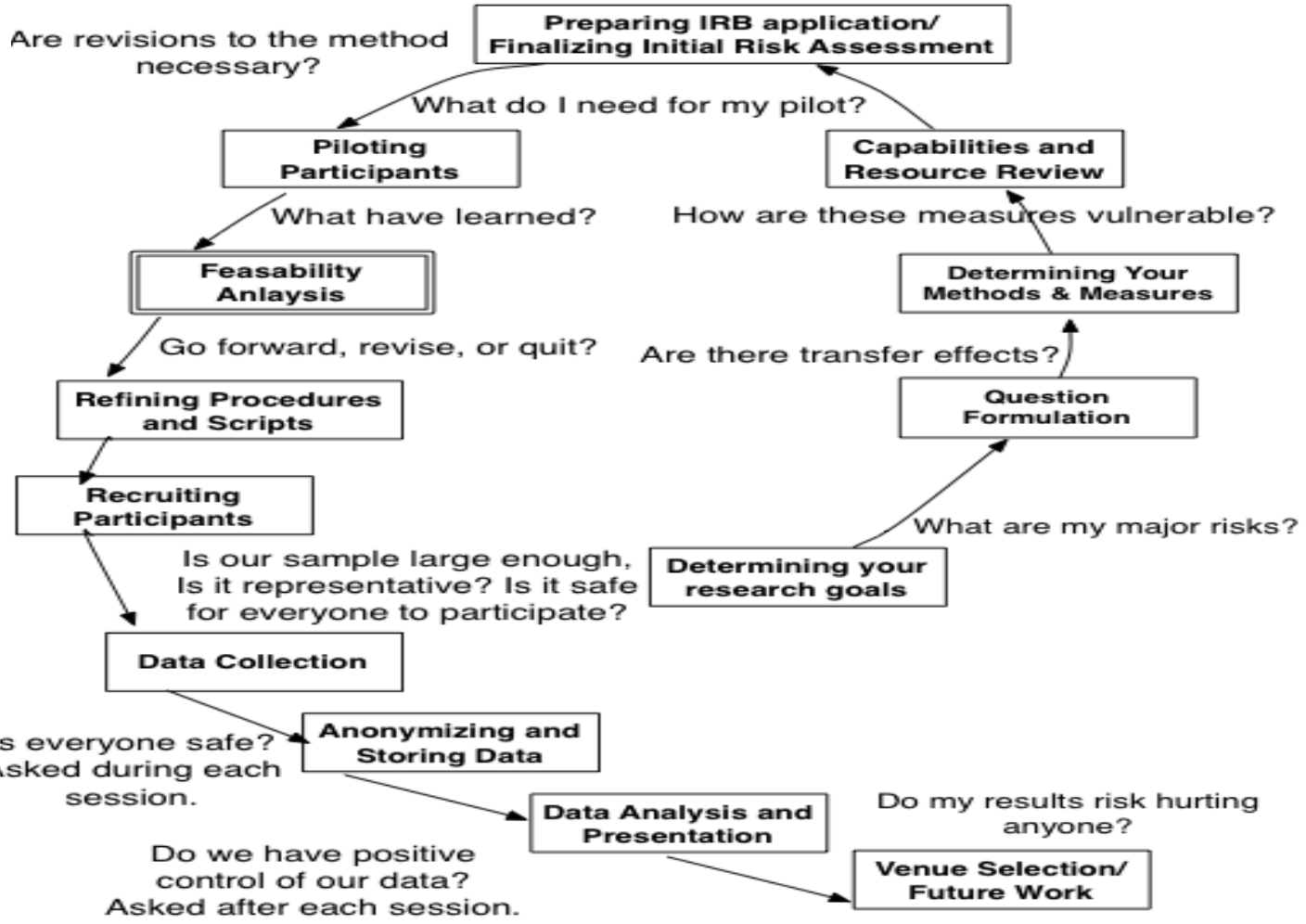
1. Experimental Process Overview, linear

An iterative, and often over-lapping set of process steps



Experimental Process Overview

Risk Driven, more spiral



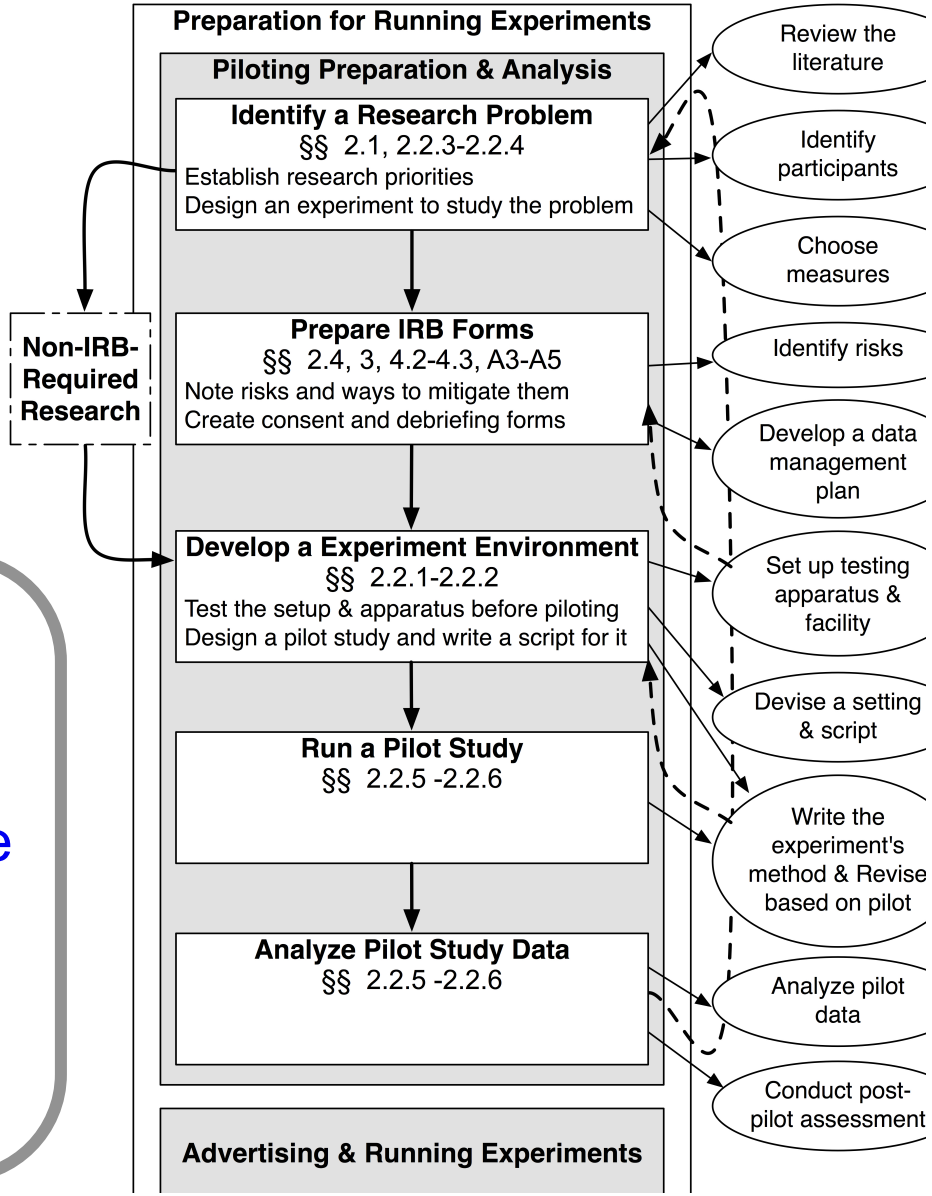
Summary: Lessons so Far

- More steps than I thought
- Iterative and risk-driven (if you pay attention)
- A process but not a set process
- Studies will overlap each other and inspire each other
- It is useful to have the RAs/Es pay attention
  - Ss suddenly ‘get it’
  - Ss don’t get some aspect
  - Ss comments
  - Ss ‘cheat’ somehow

2. Preparation for an experiment

(12:50-13:10)

Experiments are driven by their questions and shaped by the methods available to explore those questions and existing results/lessons in that area



What studies need IRB?

- In the US
  - If not publishing or class projects, no IRB (but, be careful)
  - If only authors are Ss, no IRB
  - If only published / publicly available data, no IRB
    but IRB has to ok this (!)
  - Else: IRB
  - Blood, sexual history, etc. are high-risk,=> full IRB
- Outside US
  - Depends, UK used to do IRB only on high-risk studies
  - Can you tell me?
- BUT, in all cases, worth having someone check your work

IRB Forms

- Used to check your work
- Can be a draft method section for papers
- May be worth being clear and concise
- Also check with example forms for language
- Draft for the Principle investigator (PI)

Summary: Piloting

- *New*: Check the literature further
- *New*: Check other data sources, e.g., your lab, respositories like Databrary.org
- Write out method
  - Used to check your work by you and others
- Use a script,
  Step 1, start program XXX on disk YYY, Step 2 “Welcome to…”
- Start local, e.g., You, and then officemate, and then move further and further away
- Mount a scratch monkey
- Check your apparatus and data gathering and use of data
- Consider/reconsider, number of Ss to run
  - Previous studies
  - Power analyses (Cohen for Ss; 🕮Ritter et al. (2011) for models)
  - Why not prefer large effects?
  - *New*: check your pilot data with analysis (linear, non-linear so take more points?)
  - *New:* and with target stat programs, see it loads, etc.

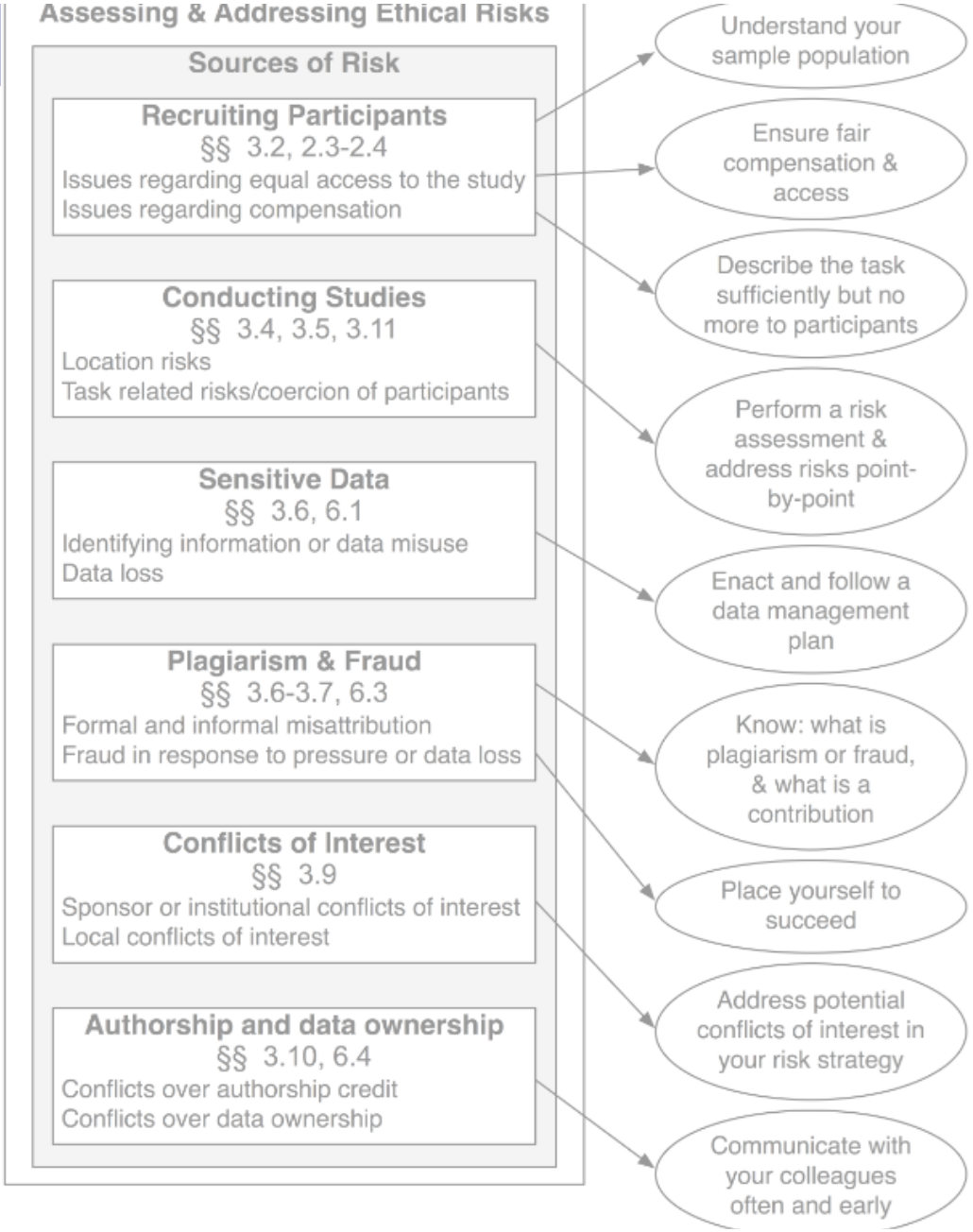
3. Ethical Challenges Associated with the Experimental Process

(13:10-13:30 break)

(13:30-1350)

Ethical problems can be decreased by deliberate proactive action.

A couple of bad examples and then a general view



The Monster Study:
Wendell Johnson's Stuttering Study (1939)

- Evaluated the effect of external valuations on stuttering
  - interupting vs. non-interupting conditions
- Studied 22 orphans ranging in age from 5-15 years old, grouping them into 5 fluency categories
- Resulted in long-term developmental and psychological harm, with $925,000 awarded to six of the participants in 2007
- Avoid manipulations that can harm people

Jesse Gelsinger (1981-1999)

- Included in a bio-medical intervention study to replace a missing participant despite testing positive for high ammonia levels
- The informed consent agreement failed to disclose either known adverse drug effects or the death of two monkeys in animal trials.
- A profound conflict-of-interest existed
- Avoid conflict of interests
- Cases like this give rise to the need for IRBs

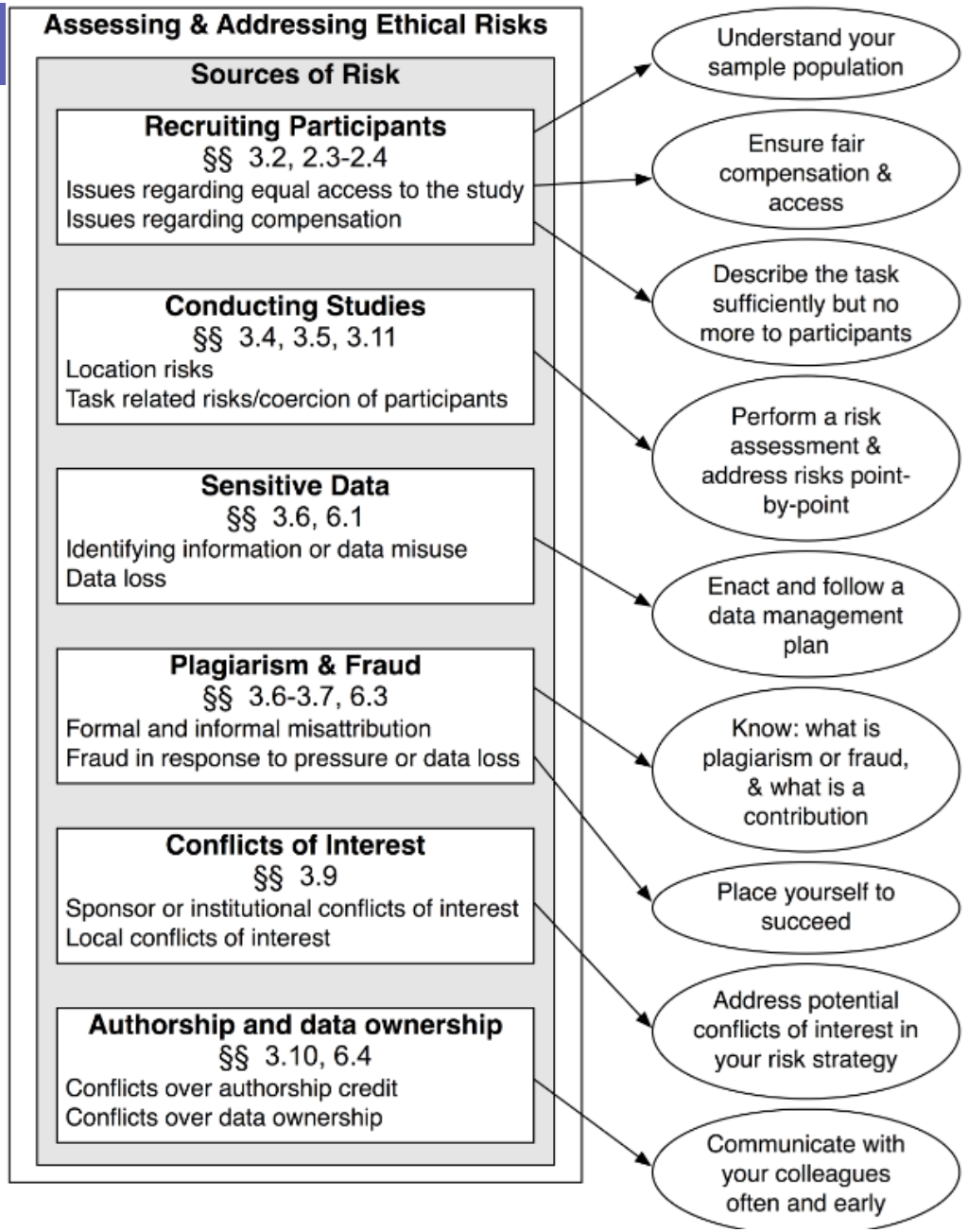
A HCI Study Gone Wrong (circa 2008)

- No informed consent
- No privacy grantees or data management plan
- "You have no friends." Yes, a student researcher felt compelled to inform a participant and the S's teachers and the Dean of this fact.

Ethical Challenges Associated with the Experimental Process

(13:30-1350)

- Ethical problems can be decreased by deliberate proactive action.



Exercise: Two ethical dilemmas
[iff time]

A. In screening candidates for a stress study, you discover one of your P's heart rate suggests a medical condition. (or, in any study situation, a subject arrives in an altered state.) Do you have an ethical obligation to report this to them?

B. In collaboration with Dept. of Veterans Affairs, you & your team are evaluating long term a learning theory and a tutor based on that theory where some learners have PTSD. As the study progresses, many of the learners experience significant personal hardship and prolonged unemployment.

Does this change in status present an ethical challenge with regards to the participants' freedom of consent? If so, does the veterans' right to participate and their self-felt obligation to help, and their increasing interest in the payments, outweigh this potential threat to consent? Also, what if the nature of the content knowledge (e.g., battlefield first-aid) interacts badly with their PTSD?

Summary: How to avoid ethical problems

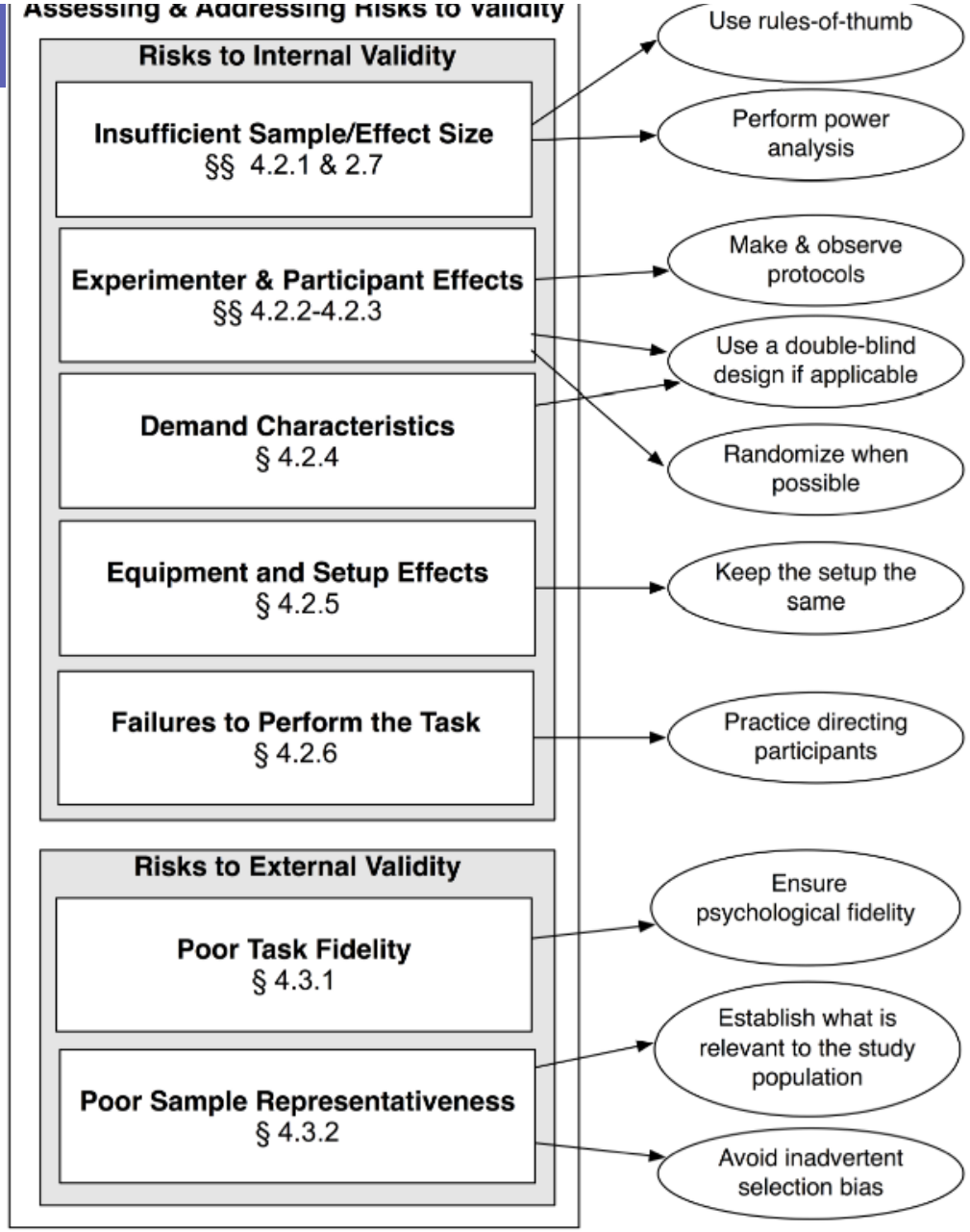
- Recruit fairly
- Look out for your Ss
- Anonymise data at the beginning of each session by using subject IDs, not names
  - *New*: Only take data you need
  - *New:* Be aware of deductive closure
- Have a plan for surprising data or situations (e.g., high BP, contact senior RA, PI, IRB Office, 911)
- Communicate early and relatively often about publication plans and data ownership
- Some argue that you have an obligation to use the data you gather

4. Risks to Validity:
Constraints on your study

(13:50-14:15)

Or: alternative hypothesis for results

Challenges to validity can be anticipated and mitigated.

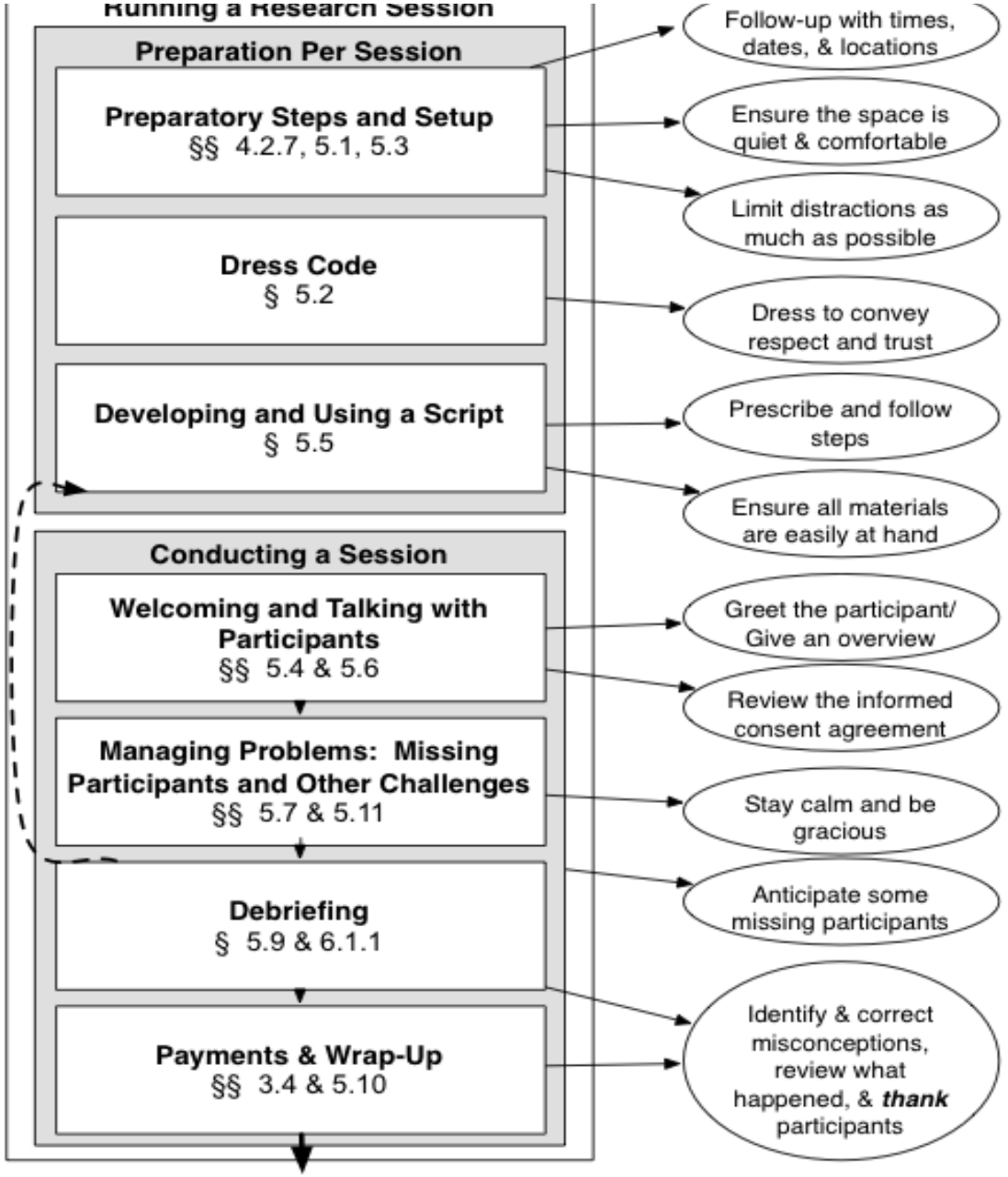


(5) Running a Session

(14:35-14:50)

(14:15-14:35 break)

Success in execution is directly correlated to careful preparation



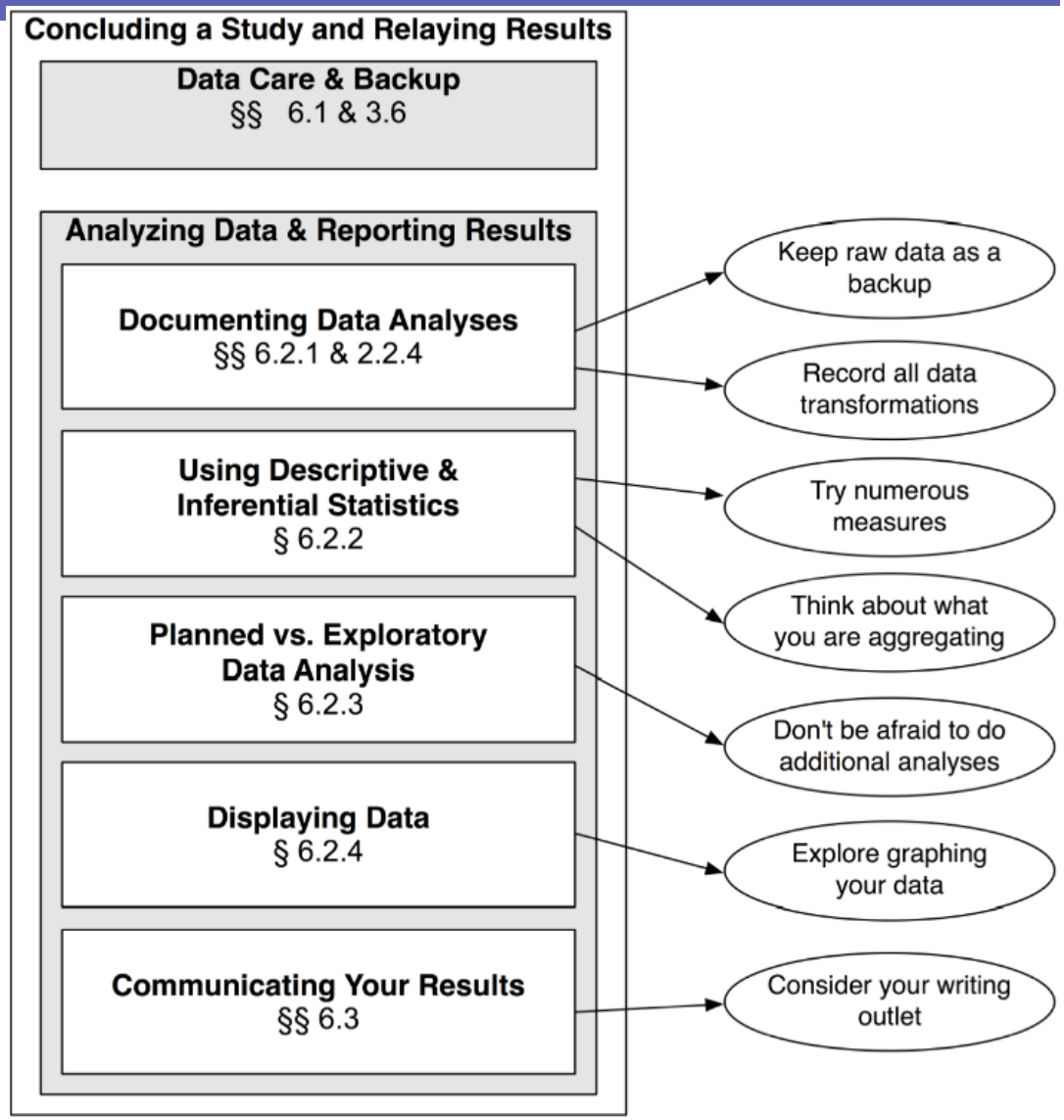
Summary: Running a session

- Use of piloting means no surprises (except for the data!)
- Script keeps treatment the same, it includes session set up
- *New*: Assignment needs to be random
  - Not a sample of convenience
  - Needs to not pick up: time of day, week, month, semester, cohort, age, etc.
  - Use cards, coins, SPSS, spreadsheet to randomize
  - We just don't use Ps who cannot make the long session
  - We assign PC and Mac equally to the 9 groups
- Keep eyes open while running for further insights
- Anonymise data as soon as possible

(6) Concluding an Experiment and Reporting Your Results

(14:50-15:05)

Debrief subjects,
Debrief each other,
Debrief potential audience!



Summary: Concluding an Experiment and Reporting Your Results

- Concluding a session
  - Finish with the subject (thank, debrief, check paperwork)
  - Check the data was collected and saved
  - Comment on the data if anomolies
- Data care, security and privacy
  - Anonymizing removes nearly all ills
- Back up data (daily, weekly)
- Data analysis
  - Not how, but note how (document and keep track of)
  - Know your data if you are the RA that analyses
  - Save the analyses, time is not important, space is not important, the insights and results are important
  - Aside: we prefer regression
  - Aside: we prefer individual analysis

Ch 6.5 Communicating your results

- Start with a target in mind
  (if you can)
- Work to larger publications
  (workshop, conf, journal, book)
- Rewrite, rewrite, rewrite
  (the book was draft 49 [mar12], final 53)

Running KRK Online

Sarah Ricupero & Frank E. Ritter

29jul2020

- IRB/protocol changes, cover online condition (keep conditions balanced)
- Scheduling the 2-5 sessions was with email, XL, box (shared cloud document)
- Tool choice varies based on Subject numbers, number of potential recruits, length of sessions, length of task, complexity of conditions, repeated sessions and their regularity
  - hroot (that's what we use), orsee (inspiration for hroot)
  - hroot: https://www.wiso.uni-hamburg.de/en/forschung/forschungslabor/experimentallabor/hroot.html
  - orsee: http://www.orsee.org/web/
- Recruitment
  - No: handouts, speaking in class
  - Yes: "Who can we email?" course instructors, other labs, social organizations, grad-student-list@psu.edu, facebook, linkedIn
- Payment
  - No: Cash doesn't work
  - Yes: gift cards (Amazon), University cash
- Running the study
  - Required: computer, camera, microphone, paper, pencil, email, internet
  - Zoom (provided by university, Adobe Connect, MS Teams), webcam, microphone
  - Changes: modify S's browser to turn-off autocomplete, ask them to clear their space
  - Use email to arrange
  - Use a web-based apparatus, or download and then delete
  - Does not work if ms timing required, unless they download keystroke logger (e.g., RUI)
  - Short tasks, try Amazon MT (https://www.prolific.co/)
  - Should include: displays (matters), type of machine (matters), type of mouse (matters)
  - Also see: http://sll.stanford.edu/docs/Webinar_materials_v2.pdf
  - It helps that people are at home and sheltering-in-place, essentially, less general movement less activity

7 Concluding Comments/Afterword

15:05-15:40

- Appropriate behavior with subjects
- Insights
- Repeatability
- Reportability

- Questions/Discussion

Summary 1 of tutorial:

Relooking at failure: What constitutes a failure?

- Someone got hurt.
- After committing significant resources, the study was never completed.
- We have learned nothing new because our data is not repeatable or generalizable.
- We have failed to communicate our results or their significance to anyone else.

Sources of Failure?

- Why did someone get hurt?
  - *We failed to do a risk assessment.*
  - *Being prepared for unanticipated problems.*
  - *We failed to screen participants properly.*
  - *We failed to either develop or follow procedures, either experimental procedures or data management procedures.*
  - *We did not anticipate or mitigate situational risks either in our experimental setting or outside of it that hurt our participants.*
  - *We ignored additional insights we could have learned from the participants through observation or debriefing.*
  - *New: deductive disclosure, enough to identify someone. Male, US citizen, IST professor (there are three in the world)*
  - Others?

Sources of Failure?

- Why we were unable to complete the study?
  - ➢ *We were overly ambitious*, perhaps because we failed to fit the research question or methods to the problem at hand.
  - ➢ *We ran out of time.*
  - ➢ *We ran out of resources or lacked them in the first place.*
  - ➢ *We lacked the people, either participants or staff, or trained staff.*

  (experiments appear to have less risk than modeling)

Sources of Failure?

- Why we were unable to reproduce our results or generalize them?
  - *We failed to use the same experimental procedures or test under the same conditions for each S.*
  - *We failed to achieve an adequate sample size or sufficient degree of representativeness in our sample.*
  - *Our task fidelity was poor. We failed to construct an experimental task that was analogous with respect to its key points.*

Sources of Failure?

- Why have we been unable to report our results or communicate their significance?
  - *We failed to properly catalog or backup our data.*
  - *We failed to write as we went. We no longer remember some of the critical, early details.*
  - *We made poor data analysis or display choices.*
  - *We failed to identify a venue early, or understand who we should consider our audience.*

How do we avoid failure?

- We recognize that running a study is an incremental risk-driven process, similar in some respects to spiral development of systems (Boehm & Hansen, 2001; Pew & Mavor, 2007).
- To be successful, we need to:
  - Formulate a research question that meets our research goals
  - Have a theory of transfer effects that minimizes risks associated with confounding variables, and enables us to conserve time and resources
  - Pilot studies and study components
  - Be candid in our risk assessments and be willing to adapt and refine

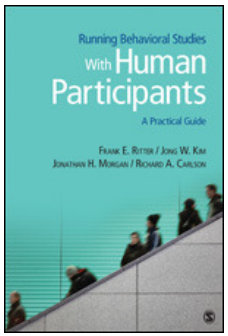
Summary 2 of Tutorial

- There are steps to running a study separate from design and analysis
- These steps include practical, hands-on, implicit and local knowledge
- They are informed by previous studies
- To be successful, we need to:
  - Formulate a research question that meets our research goals
  - Pilot studies and study components
  - Be candid in our risk assessments and be willing to adapt and refine
  - Be aware of alternative hypotheses, and avoid what we can and control what we cannot avoid
  - Plan for reporting results early

Adjustments for Models as Ss

- Document and freeze the model
- Document and annotate its trace/ predictions
- Run the model until you have predictions, not samples of predictions (📖Ritter et al., 2011)

If you will teach this or want to know more....



- Full book available from Sage [paper & PDF]
- Slides available as ppt or pdf (email us or see RBS site)
- Tech report available as shorter version

  http://acs.ist.psu.edu/reports/ritterKM09.pdf
- Workbook and other resources available http://frankritter.com/rbs

7 Concluding Comments/Afterword

15:05-15:40

- Appropriate behavior with subjects
- Insights
- Repeatability
- Reportability

- Questions/Discussion

References

📖 acs.ist.psu.edu/papers

Boehm, B., & Hansen, W. (2001). The Spiral Model as a tool for evolutionary acquisition. *Crosstalk: The Journal of Defense Software Engineering, 14*(5), 4-11.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*, 155-159.

Calfee, R. C. (1985). *Experimental methods in psychology*. New York, NY: Holt, Rinehart and Winston.

Pew, R. W., & Mavor, A. S. (Eds.). (2007). *Human-system integration in the system development process: A new look*. Washington, DC: National Academy Press. http://books.nap.edu/catalog.php?record_id=11893, checked March 2012.

Ray, W. J. (2008). *Methods toward a science of behavior and experience* (9th Edition ed.). Belmont, CA: Wadsworth Publishing.

📖 Ritter, F. E., Kim, J. W., & Morgan, J. H. (2010). *Running behavioral experiments with human participants: A practical guide* (Tech. Report No. ACS 2005-2). Applied Cognitive Science Lab, School of Information Sciences and Technology, Penn State. acs.ist.psu.edu/reports/ritterKM09.pdf.

Ritter, F. E., Kim, J. W., Morgan, J. H., & Carlson, R. A. (2013). *How to run experiments: A practical guide to research with human participants*. Thousand Oaks, CA: Sage. http://www.frankritter.com/rbs

📖 Ritter, F. E., Schoelles, M. J., Quigley, K. S., & Klein, L. C. (2011). Determining the number of model runs: Treating cognitive models as theories by not sampling their behavior. In L. Rothrock & S. Narayanan (Eds.), *Human-in-the-loop simulations: Methods and practice* (pp. 97-116). London: Springer-Verlag.

Roediger, R. (2004). What should they be called? *APS Observer, 17*(4), 5, 46-48. Online: www.psychologicalscience.org/observer/getArticle.cfm?id=1549.

Winston, A. S. (1990). Robert Sessions Woodworth and the "Columbia Bible": How the psychological experiment was redefined. *The American Journal of Psychology, 103*(3), 391-401.