May the Force Be With You: The Role of Evidential Force in Empirical Software Engineering

Shari Lawrence Pfleeger
Senior Information Scientist
RAND
Pfleeger@rand.org

Overview

- From the part to the whole: examining the body of evidence
- Ignorance, uncertainty and doubt
- Evidential force
- Multi-legged arguments
- Example: What to do about ephedra
- Moving forward

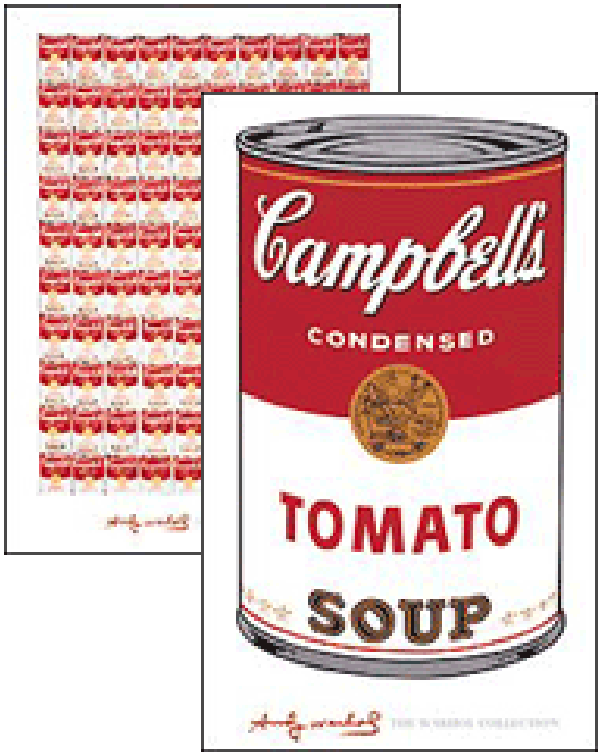
From the Part to the Whole: Examining the Body of Evidence

"Science is a particular way of knowing about the world. In science, explanations are limited to those based on observations and experiments that can be substantiated by other scientists. **Explanations that cannot be based on empirical evidence are not a part of science**." Introduction to *Science and Creationism: A View from the National Academy of Sciences,* National Academies Press, 2000.

R

Soup or Art?

"**It appears to me that they who rely simply on the weight of authority to prove any assertion, without searching out the arguments to support it, act absurdly**. I wish to question freely and to answer freely without any sort of adulation. That well becomes any who are sincere in the search for truth."
Vincenzo (father of Galileo) Galilei, 1574



Terminology

- We make a **case** for something.
- The case has three parts:
  - One or more **claims** that properties are satisfied
  - A body of supporting **evidence** (from a variety of sources)
  - A set of **arguments** that link claims to evidence

Two Key Uses of Evidence

- Hypothesis generation
  - Theories about the way processes, products and resources work alone and in concert
- Hypothesis testing
  - Is what we believe confirmed by the evidence?

Key Questions for Empirical Software Engineering

- What do we mean when we say that a technology "works"?
- What kinds of evidence (and how much evidence) do we need to demonstrate that it works?
- Who provides the evidence, and who vets the evidence? (For instance, many of the claims about data mining are provided by the vendors.)
- If it works in one domain, does that tell us anything about other domains?
- How can evidence inform our thinking about the social, economic and political tradeoffs of using an imperfect technology?

Ignorance, Uncertainty and Doubt

"When a scientist doesn't know the answer to a problem, he is **ignorant**.

When he has a hunch as to what the result is, he is **uncertain**.

And when he is pretty darn sure of what the result is going to be, he is in some **doubt**." (Feynman 1999)

More on Ignorance, Uncertainty and Doubt

"We have found it of paramount importance that in order to progress we must **recognize the ignorance** and **leave room for doubt**.

Scientific knowledge is a body of statements of varying degrees of certainty—**some most unsure, some nearly sure, none absolutely certain**." (Feynman 1999)

Types of Evidence (Schum 94)

- Tangible evidence
  - Can be examined to see what it reveals
  - Examples: objects, documents, images, measurements, charts
- Testimonial evidence: Unequivocal
  - Received from another person
  - Examples: Direct observation, hearsay, opinion
- Testimonial evidence: Equivocal
  - Examples: Complete equivocation, probabilistic
- Missing evidence (tangible or testimonial)
- Accepted facts (authoritative records)

Evidential Credibility

Depends on

- Type of evidence
  - Documented?
  - Replicable?
  - Well-designed?
  - Measurable?
- Creator
- Conveyor
  - Refereed publication?
  - Trade journal?
  - Self-published?

Tests for Testimonial Credibility

- Sensitivity
  - Sensory defects?
  - Conditions of observation?
  - Quality/duration of observation?
  - Expertise/allocation of attention
- Objectivity
  - Expectations
  - Bias
  - Memory-related errors
- Veracity

Putting It Together

How to combine evidence when there are pieces

- Of dubious credibility?
- Missing?
- Ambiguous?
- Conflicting?
- Not replicable?

R

R

R

R

R

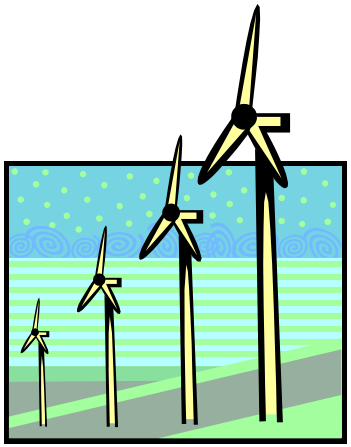
R

Examples

- Two conflicting studies of hormone replacement therapy (Kolata 2003)
  - Nurses' health survey: Long-term study indicating that HRT helps protect against heart disease
  - Women's Health Initiative: Recent study indicates that HRT increases risk of heart disease
- Curare study: confounding variable (natural vs. synthetic curare) discovered well after original study author embarrassed
- Conflicting studies of inspection teams
  - Some show team is useful, others don't

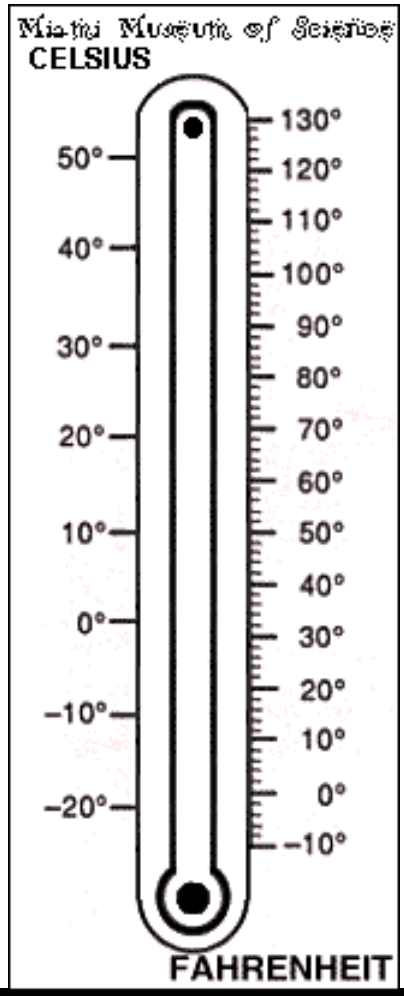
Evidential Force

- A body of evidence has evidential force, with each piece of evidence contributing to the whole.
- One piece of evidence can increase or diminish the evidential force.



Assessing Evidential Force

- Jeremy Bentham (1839) proposed a numerical scale.
- Range from –10 to +10
- Positive: gradations favoring H
- Negative: gradations against H
- Zero: no inferential force



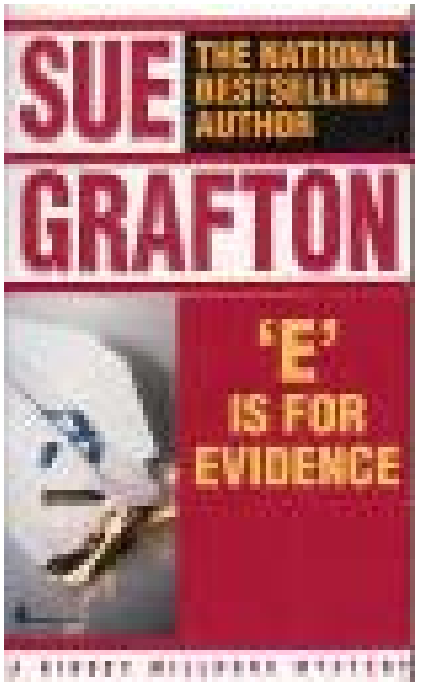
R

Bentham's Four Questions to Determine Evidential Force

- How confident is the witness in the truth of the event asserted?
- How conformable to general experience (that is, how rare) is the event asserted?
- Are there grounds for suspicion of the untrustworthiness of the witness?
- Is the testimony supported or doubted by other evidence?

Schum's Approach

- Evidence marshalling
- Bayesian analysis
- Chains of reasoning
- Measures of likelihood: P(H|E)



R

Multi-legged Arguments

- Work done by Bloomfield and Littlewood.
- General idea: Two heads are better than one.
- Example: Use a process-based argument (e.g. review of practices) and a product-based one (e.g. static code analysis).
- Another example: UK Def Std 00-55
  - One leg is logical proof.
  - Another leg is probabilistic claim based on statistical analysis.

More on Multi-legged Arguments

- Easier to analyze than one comprehensive argument.
- Handles different types of evidence.
- Legs need not be independent.
- More confidence than in one leg alone (but does extra confidence justify extra cost?)

Criteria for Diversity

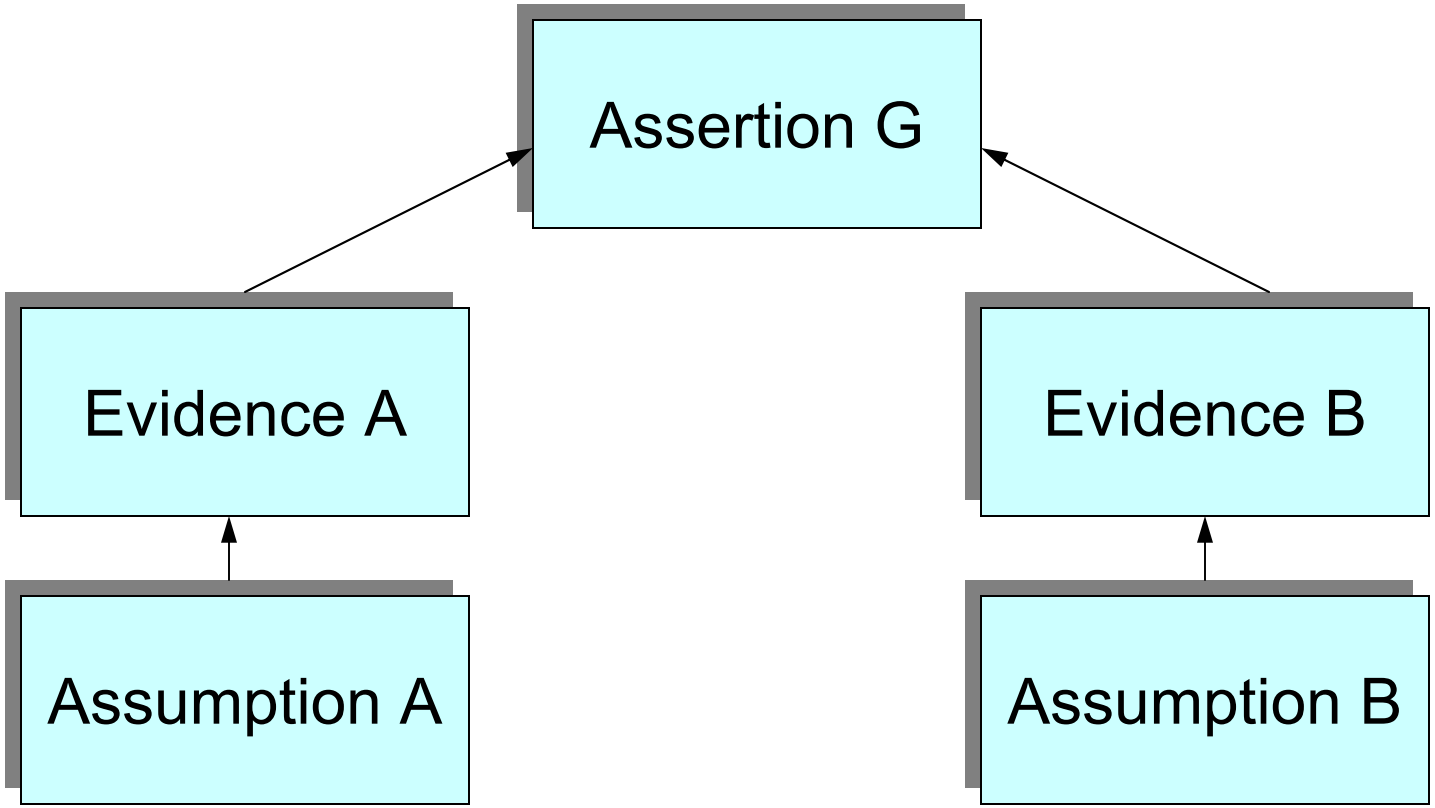
- Weaknesses in modeling assumptions
  - E.g. Is formal specification an accurate representation of higher-level requirements?
- Weaknesses in evidence
  - E.g. Is complete testing feasible?

Relationship to Evidential Force

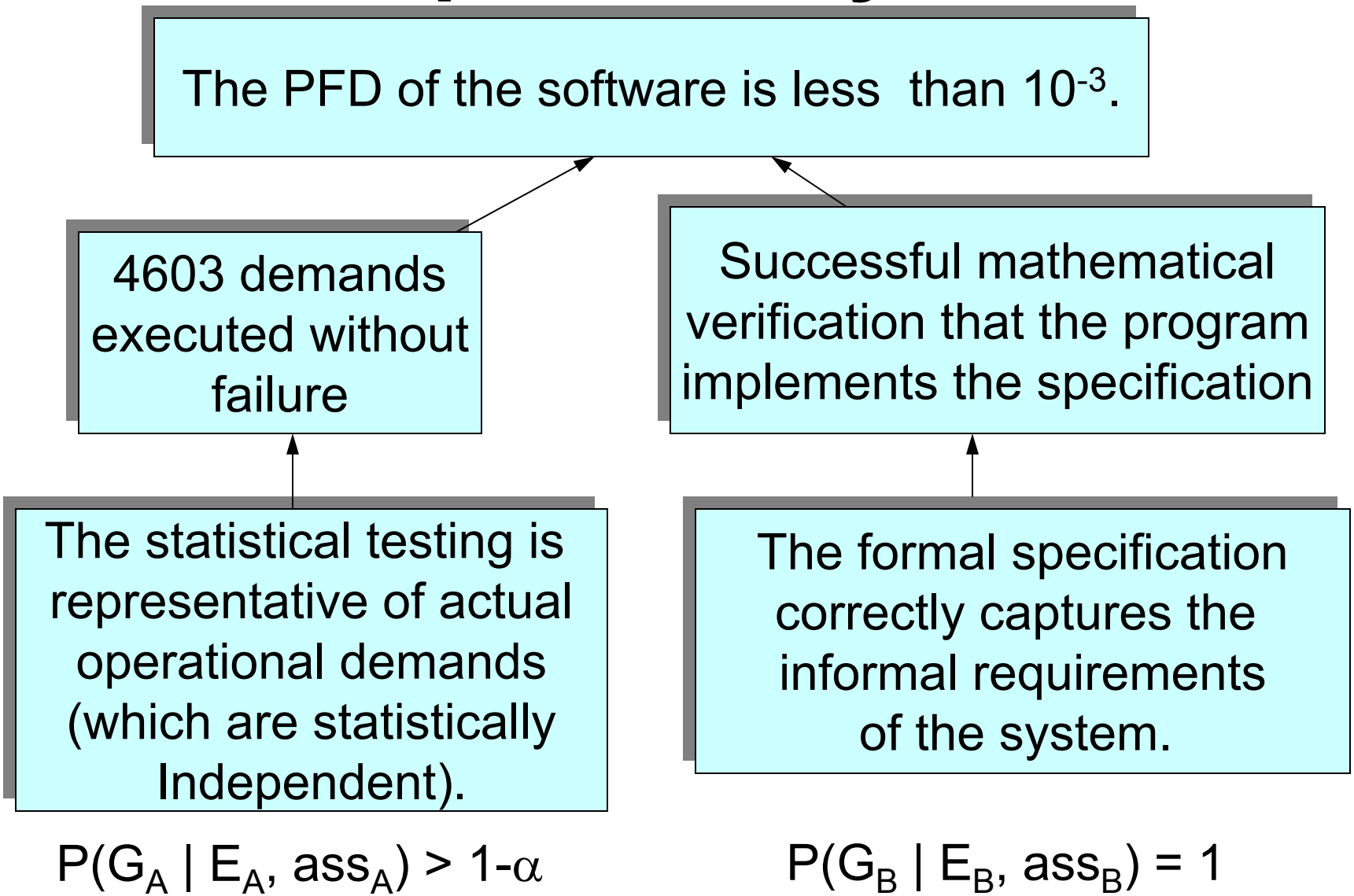
- Argument diversity increases confidence and thereby increases argument force.
- Example: “An argument that gives 99% confidence that the probability of failure on demand is smaller than 10^{-3} is stronger than one that gives only 95% confidence in the same claim.”

Dependence of Legs

Not a bad thing: It can increase confidence in overall assertion.



Example: Safety Goal



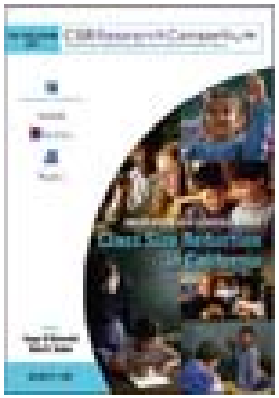
Things to Consider

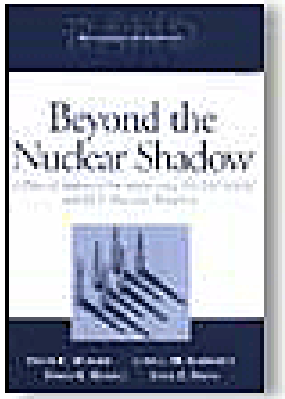
- Extensiveness of evidence
- Assumption confidence
- Difficulty of assigning numerical values
- Need for simplifying assumptions
- Contribution of each piece of evidence to the whole

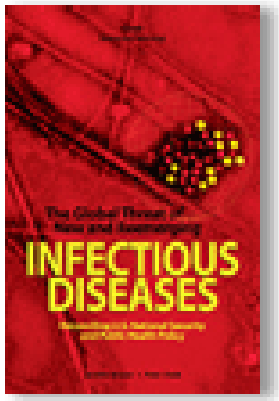
What We Do at RAND

"The RAND Corporation, America's original think tank, earns its money fishing truths out of murky political and social waters. The quantitative conscience of RAND [is] often the final arbiter of what constitutes the true story."

Bradley Efron,Stanford University







Example: What to Do About Ephedra

Ephedra is the herb (ma huang, as Chinese call it).

Ephedrine is the drug.

Claims About Ephedra

- Improves weight loss
- Enhances athletic performance
- Almost 18,000 reports of adverse effects (including death and illness)

What About the Evidence?

- Dietary supplements not subject to same rigorous standards as drugs; therefore no need to show evidence of safety.
- Therefore limited evidence on ephedra.
- FDA seeks evidence of “significant or unreasonable risk of illness or injury.”
- Safety of ephedra cannot be demonstrated with scientific certainty.

State of the Evidence

- 52 (published and unpublished) trials of ephedra or ephedrine for weight loss or athletic performance
  - Many had small numbers of people
  - Many had short periods of time
  - Other limitations, such as non-representative sample
- 1820 consumer complaints to FDA
- 71 reports in the medical literature
- 15,951 reports to Metabolife, a maker of ephedra-containing supplements

Step 1: Set Criteria for Confidence in Evidence

- For weight loss studies:
  - Select studies that assess ephedra, ephedrine, or ephedrine plus other compounds used for weight loss. (Yield: 44 of 52 studies)
  - Select studies with periods of at least 8 weeks. (Yield: 26 of 44 studies)
  - Select studies with no serious other limitations. (Yield: 20 of 26 studies)
- For athletic performance studies:
  - Select studies that assess ephedra, ephedrine, or ephedrine plus other compounds used for athletic performance. (Yield: no studies of ephedra, only 8 studies of ephedrine, all but one of which included caffeine)

Step 1: continued

- For safety studies:
  - Select studies with documentation of adverse event.
  - Select those confirming that ephedra or ephedrine had been consumed within 24 hours before adverse event OR with toxicological evidence of those substances in blood or urine.
  - Select those documenting exclusion of other possible causes.
- Yield: 284 possible events.

Step 2: Categorize Treatments

- For weight loss studies:
  - Comparisons made in six categories
    - Ephedrine vs. placebo (5 trials)
    - Ephedrine and caffeine vs. placebo (12 trials)
    - Ephedrine and caffeine vs. ephedrine alone (3 trials)
    - Ephedrine and caffeine vs. another active pharmaceutical for weight loss (2 trials)
    - Ephedra vs. placebo (1 trial)
    - Ephedra with herbs containing caffeine vs. placebo (4 trials)

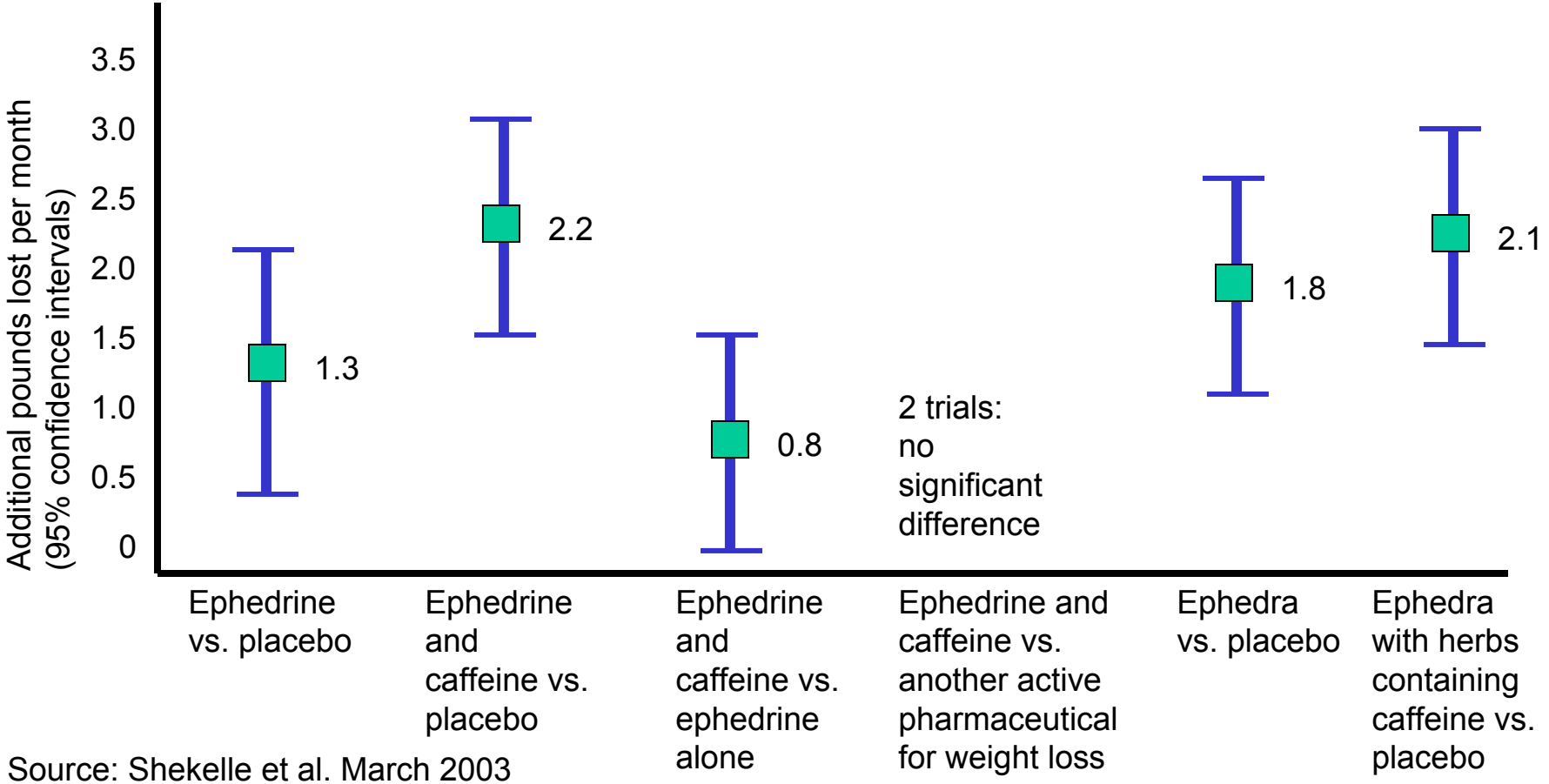
Step 2: continued

- For athletic performance studies:
  - Each trial involved a different kind of exercise, so each trial was assessed separately.

Step 3: Set Outcome Measures

- For weight loss studies:
  - Clear indicators of weight loss
- For athletic performance studies:
  - Measures of exercise performance, such as
    - Oxygen consumption
    - Time to exhaustion
    - Carbon dioxide production
    - Muscle endurance
    - Reaction time
    - Etc.
- For safety studies:
  - Group symptoms into clinically similar categories

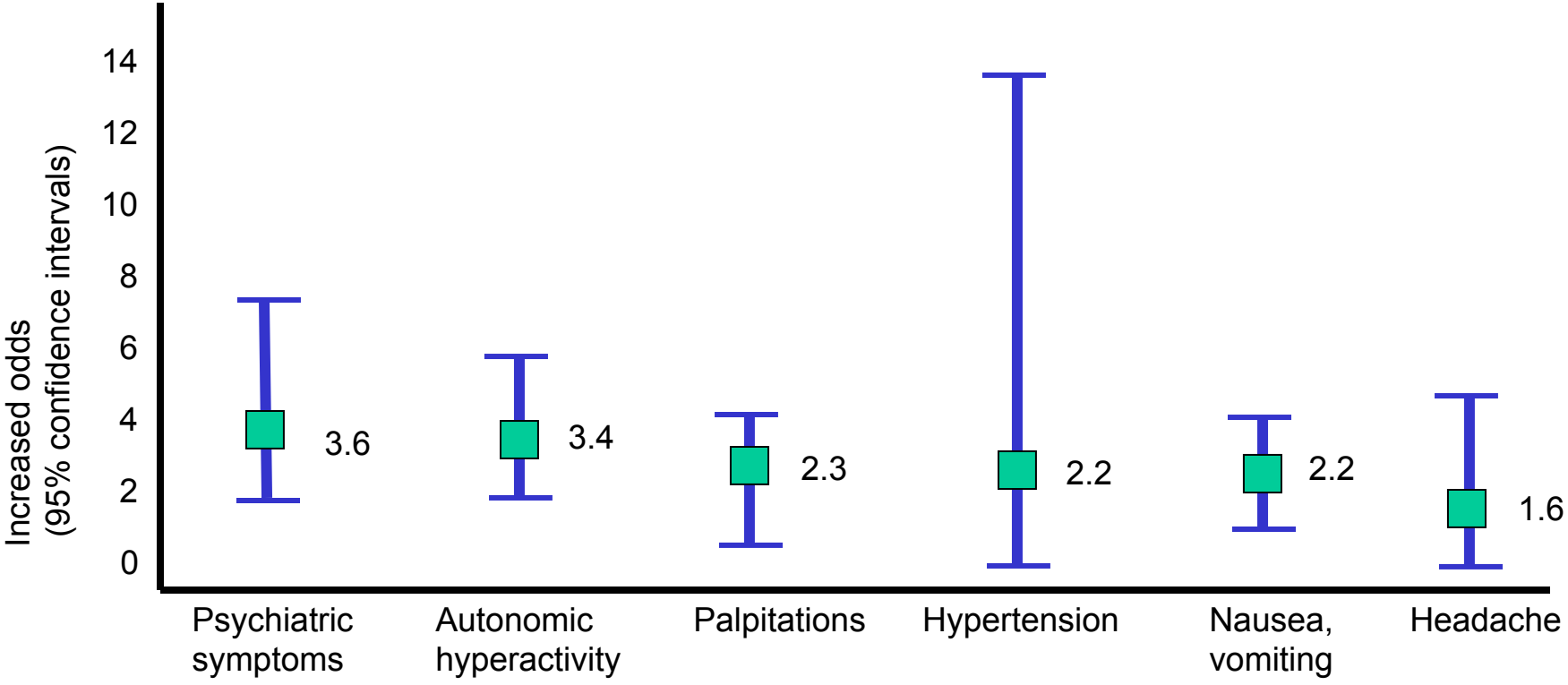
Taking Ephedrine or Ephedra Can Increase Weight Loss in the Short Term



Source: Shekelle et al. March 2003

R

Taking Ephedrine or Ephedra Increases the Odds of Suffering Adverse Events



Source: Shekelle et al. Spring 2003

Lessons to Learn

- There are techniques for combining data and results.
- Including uncertainty helps us evaluate risks.
- Empirical investigation is a process, not an end in itself.
- We can become more sophisticated—and knowledgeable—in software engineering by using these approaches.

Steps Toward Improvement in Software Engineering

- Design **families** of studies
  - Plan for imperfection
  - Deal appropriately with uncertainty
- Survey existing evidence
- Hypothesis generation or testing?
- Confidence in evidence?
- For each study
  - Set criteria for confidence in evidence
  - Categorize treatments
  - Set outcome measures
- Examples: Inspections and reviews? Cost models?

Moving Forward: The Obstacles

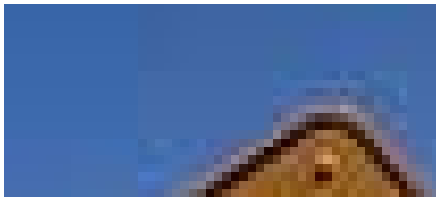
- We tend to focus on individual studies or small aspects of technology.
- We tend to look to the “hard” sciences, to statistics, and to experimental design for our models.
- We hope that evidence won’t be conflicting, rather than plan for when it is.

R

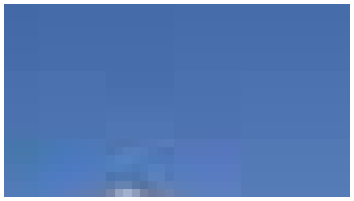
Moving Forward: The Rewards

- Results and methods used in other disciplines can help us transform the field of empirical software engineering from a disparate collection of interesting results to a discipline rich with theory and theory-testing.
- We see software engineering in a larger context.

Seeing the Big Picture

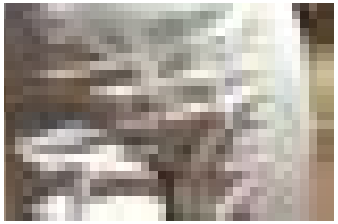


Seeing the Big Picture

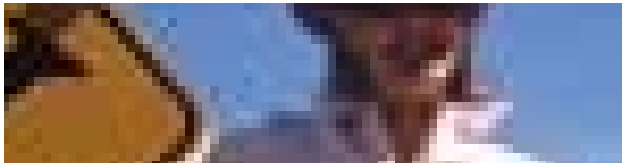


Seeing the Big Picture

Seeing the Big Picture



Seeing the Big Picture



Seeing the Big Picture

Seeing the Big Picture

Seeing the Big Picture

Questions?

References

- Jeremy Bentham, *The Rationale of Judicial Evidence*, Bowring ed., William Tate, Edinburgh, 1839.
- Robin Bloomfield and Bev Littlewood, "Multi-legged Arguments: The Impact of Diversity Upon Confidence in Dependability Arguments," *Proceedings of DSN03*, San Francisco, California, IEEE, 2003.
- Richard P. Feynman, *The Pleasure of Finding Things Out: The Best Short Works of Richard P. Feynman*, Helix Books/Perseus Books,1999.
- David A. Schum, *Evidential Foundations of Probabilistic Reasoning*, Wiley Series in Systems Engineering, New York, 1994.
- Gina Kolata, "Hormone Studies: What Went Wrong?", *New York Times*, April 22, 2003, http://www.nytimes.com/2003/04/22/health/womenshealth/22HORM.html
- Paul G. Shekelle, Mary L. Hardy, Sally C. Morton, Margaret Maglione, Walter A. Mojica, Marika J. Suttorp, Shannon L. Rhodes, Lara Jungvig and James Gagné, "Efficacy and Safety of Ephedra and Ephedrine for Weight Loss and Athletic Performance: A Meta-Analysis," *Journal of the American Medical Association*, 289(12), March 26, 2003, pp. 1537-1545.
- Paul G. Shekelle, Margaret Maglione and Sally C. Morton, "Preponderance of Evidence: Judging What to Do About Ephedra," *RAND Review*, Spring 2003, pp. 16-21