Statistical Issues and Reliability of Eyewitness Identification as a Forensic Tool

Karen Kafadar

University of Virginia

kkafadar@virginia.edu

`http://www.stat.virginia.edu/KarenKafadar.shtm`

Outline

1. **Background**
2. **Charge to the Committee on Scientific Approaches to Understanding and Maximizing the Validity and Reliability of Eyewitness Identification in Law Enforcement and the Courts**
3. **Focus: How to compare reliability between *Sequential* vs *Simultaneous* Lineup**
4. **Data, Statistical Analysis, and Uncertainty**
5. **Final thoughts: Comparing two procedures**

ASA T-shirts for sale:

IN GOD WE TRUST . . . .

All others must bring ***data***

Statistics means never

having to say you're certain

1. Background

- Eyewitness testimony can be very useful and incredibly powerful in the courtroom
- But ... the memory can play tricks, hence not always accurate nor reliable
- *Picking Cotton* by Ronald Cotton (mistakenly accused, 10+ years in prison) and Jennifer Thompson (victim: “I think” at lineup → “absolutely sure” at trial)
- Innocence Project: 330 exonerations since 1989 from post-conviction DNA testing; 236 (72%) involved mistaken eyewitness identification (`http://innocenceproject.org/know`)
- What is involved in eyewitness identification (EWI)?
- Which aspects of EWI lead to accurate identifications?

2. Charge to the Committee (*NRC Report, p.1*)

1. critically assess the existing body of scientific research as it relates to eyewitness identification;
2. identify gaps in existing literature, suggest appropriate research questions to pursue that will further understanding of eyewitness identification and offer additional insight into law enforcement and courtroom practice;
3. provide an assessment of what can be learned from research fields outside of eyewitness identification;
4. offer recommendations for best practices in the handling of eyewitness identifications by law enforcement

(and three others)

Situational aspects of EWI (*Estimator variables*):
Beyond the control of the criminal justice system

1. Eyewitness' level of stress or trauma at incident

2. Conditions affecting visibility

3. Distance between witness and perpetrator

4. Presence/absence of threat (e.g., weapon)

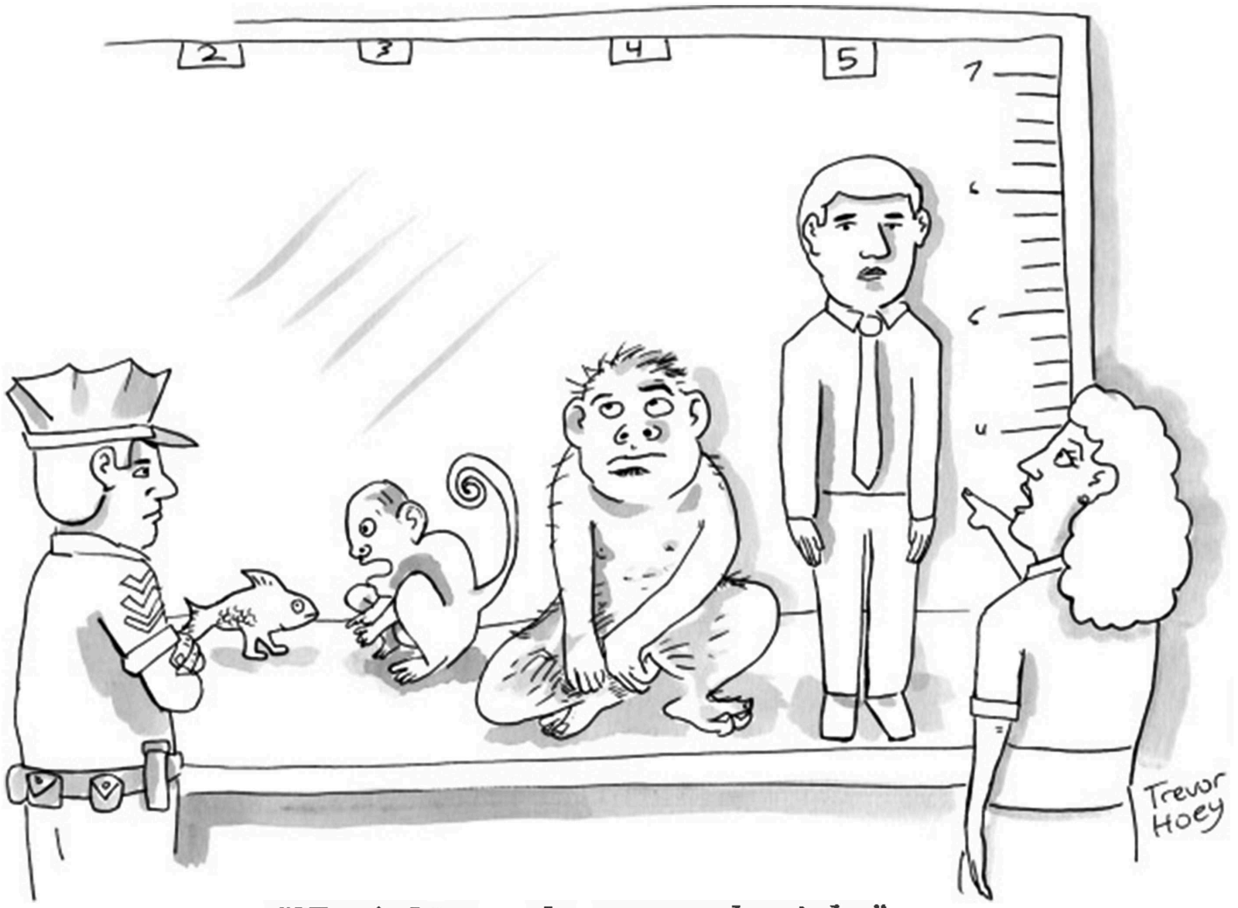
etc.

Procedural aspects of EWI (*System variables*):

1. Conditions & protocols for **lineups**
   (e.g., *sequential vs simultaneous*)
2. Nature of instructions (oral or written, short or long, ...)
3. Presence/absence of feedback
4. Number and similarities of fillers with “target”
5. Retention interval (longer ⇒ less reliable)

etc. *Which factors matter most to accuracy?*

Focus: Compare accuracy between two lineup procedures — but methods apply to comparing *any* two procedures



"That's him—the one on the right."

From THE NEW YORKER, March 7, 2011

3. Sequential vs Simultaneous?

- *Sequential*: Present each photograph, one at a time
- *Simultaneous*: Present all six photographs at once
- Early research: “Sequential is more accurate”
- Later research: “Metric for comparison is incomplete; Simulaneous is more accurate”
- Which was correct?

Lab tests and proposed metrics

Lab tests: Present participants (usually Psych 1 students) a scenario, followed by lineup (sequential or simultaneous); count proportions of correct IDs (*HR = hit rate*) and mistaken IDs (*FAR = false alarm rate*)

1. *Diagnosticity Ratio*: Collapse all participants, all scenarios:

   diagnosticity ratio = hit rate / false alarm rate
   = Sensitivity / (1 - Specificity)

2. Some participants express more *confidence* in their choices; *confidence* is related to *accuracy*; therefore, we should look at HR and FAR as functions of *levels of expressed confidence*.

Which approach is correct?

4. Data, Statistical Analysis, Uncertainty

- *Sensitivity*: When shown the *true* perpetrator, what is the probability that the "witness" identifies him/her?
- *Specificity*: When shown an *imposter*, what is the probability that the "witness" excludes him/her?
- *Sensitivity, Specificity* can be estimated only in studies *where truth is known* (by design)
- Real life: All you have is response:
  "Yes, that's the one" or "No, not that one"

- *Positive Predictive Value (PPV)*: If the claim is “Yes, that’s the one”, what is the probability that the identified person is the perpetrator?

- *Negative Predictive Value (NPV)*: If the claim is “No, not the one”, what is the probability that the excluded person is not the perpetrator?

- *PPV, NPV* are functions of *Sensitivity, Specificity, and odds that the suspect is the true perpetrator*

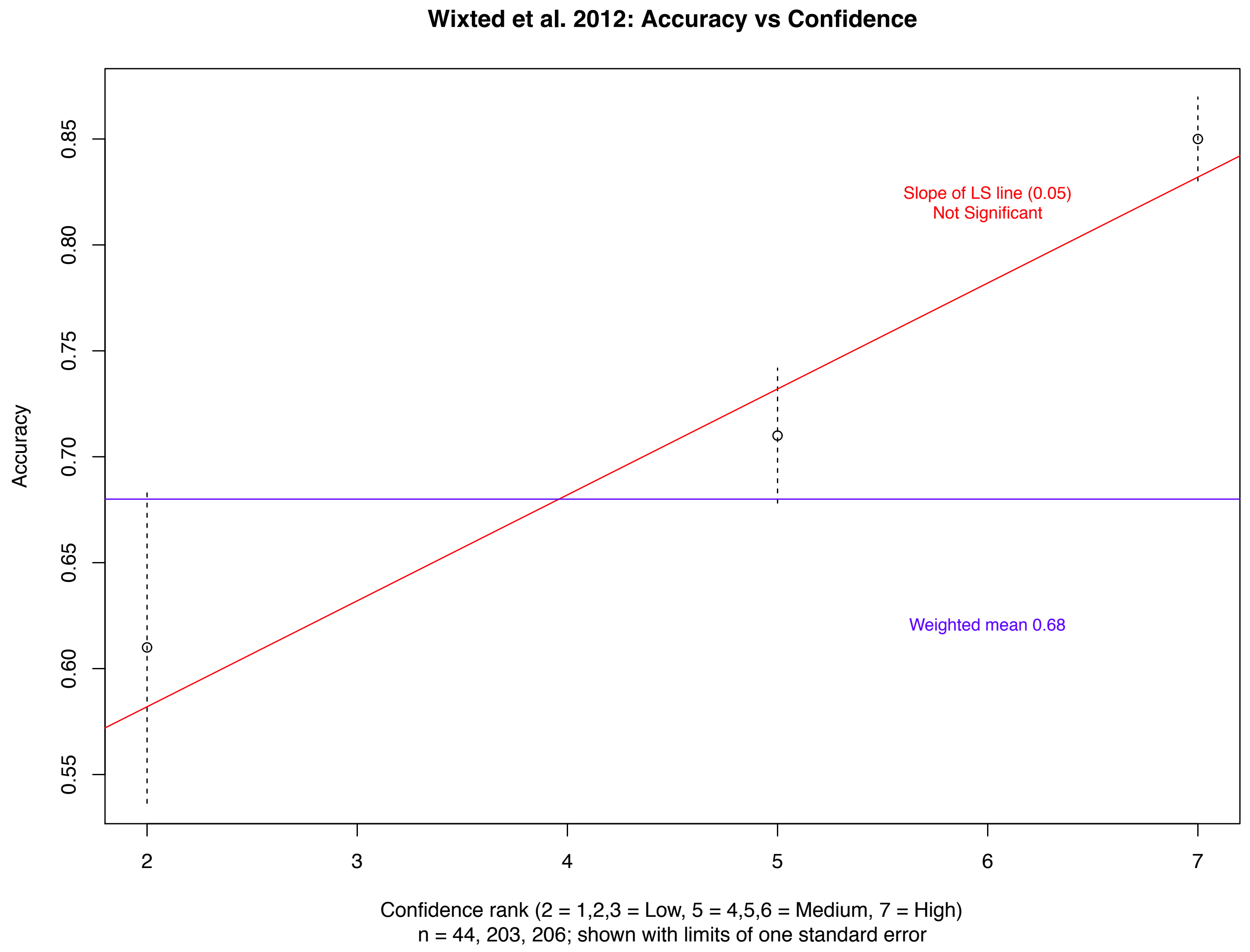
- *Diagnosticity Ratio* is related to PPV:

  $$PPV = 1\,/\,(1 + Odds/DR)$$

  so *higher DR* $\Rightarrow$ *higher PPV*

- What about correct exclusions, NPV?

- *Confidence-accuracy relationship*: Not clear that “accuracy” is related to “confidence”
- Ex: Wixted et al. (manuscript): “Confidence judgments are useful in eyewitness identifications: A new perspective”, p17:
  1. $n_1 = 44$ confidence ratings 1,2,3 (use C=2);
     Accuracy = 0.61 (0.07)
  2. $n_2 = 203$ confidence ratings 4,5,6 (use C=5);
     Accuracy = 0.70 (0.03)
  3. $n_3 = 326$ confidence ratings 7 (use C=7);
     Accuracy = 0.85 (0.02)


Wixted et al. 2012: Accuracy vs Confidence
Accuracy
0.55
0.60
0.65
0.70
0.75
0.80
0.85
2
3
4
5
6
7
Slope of LS line (0.05)
Not Significant
Weighted mean 0.68
Confidence rank (2 = 1,2,3 = Low, 5 = 4,5,6 = Medium, 7 = High)
n = 44, 203, 206; shown with limits of one standard error

- 3-point data: Weighted regression (A on C): Slope is “not significantly different from zero” (only 3 data points!)
- Other studies (more levels of confidence, larger lab studies) suggest perhaps slight relationship
- Field practice: Mixed opinions
- Reality: *accuracy* is a function of many variables (system, estimator, study design)

Using Confidence-Accuracy Relationship

If you believe confidence is related to accuracy:

- consider calculating $DR = HR/FAR$ as a function of *Expressed Confidence Level (ECL)*
- Split the sample participants into categories of ECL (those who expressed 10%, ..., 90% confidence); calculate DR for each ECL category
- even better: Plot HR vs FAR for different ECLs
- ROC curve = Receiver Operating Characteristic
- Used in quality control and comparing medical diagnostic procedures

Problem: Data points (HR, FAR) have uncertainty!

- John Tukey (in discussing uncertainty in rates at NCI): "*What has happened is history. What might happened is science and technology. So what you are really interested in is what might have happened if you could do it all over again.*"

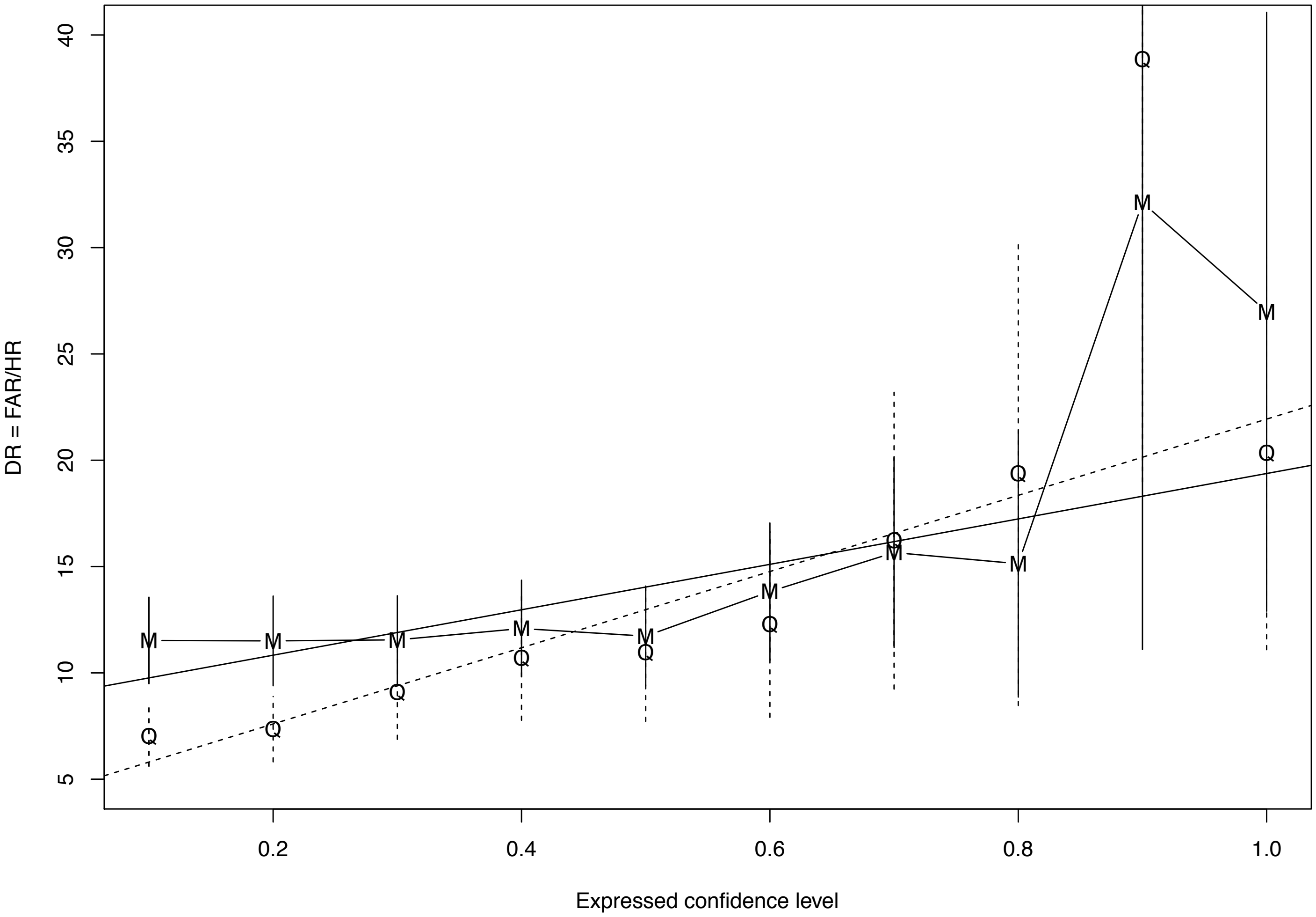
- Simulate what would happen if you calculated all the HRs and FARs (for different ECLs) *as if* you repeated the same experiment all over again

- DR vs ECL for Sequential and for Simultaneous:
  How different are they?

- How different do the two ROC curves look for Sim vs Seq?

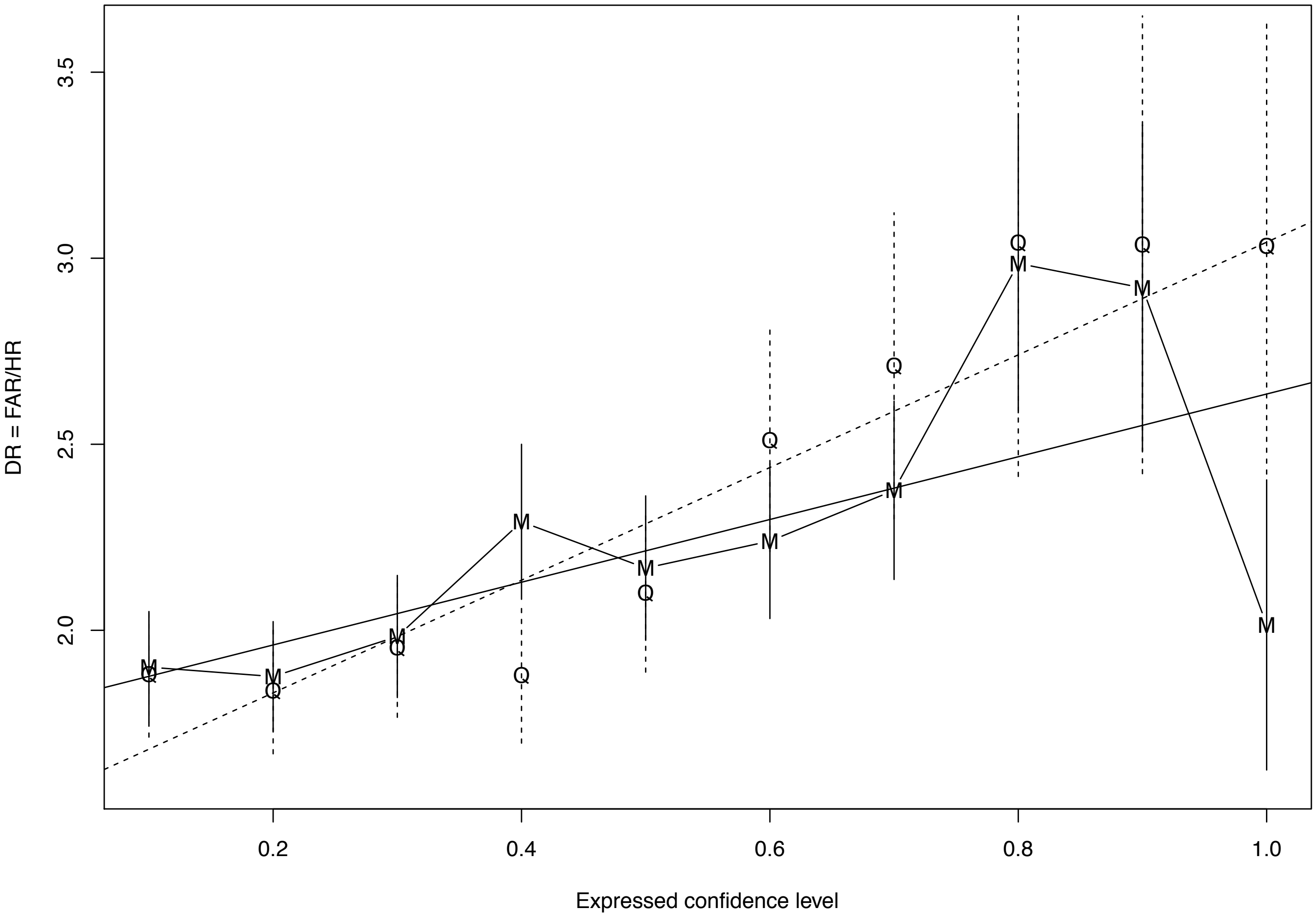
- Resulting uncertainty is underestimated, because ECLs can change (e.g., "40%" today; "20%" tomorrow)

Diagnosticity Ratio vs Expressed Confidence Level



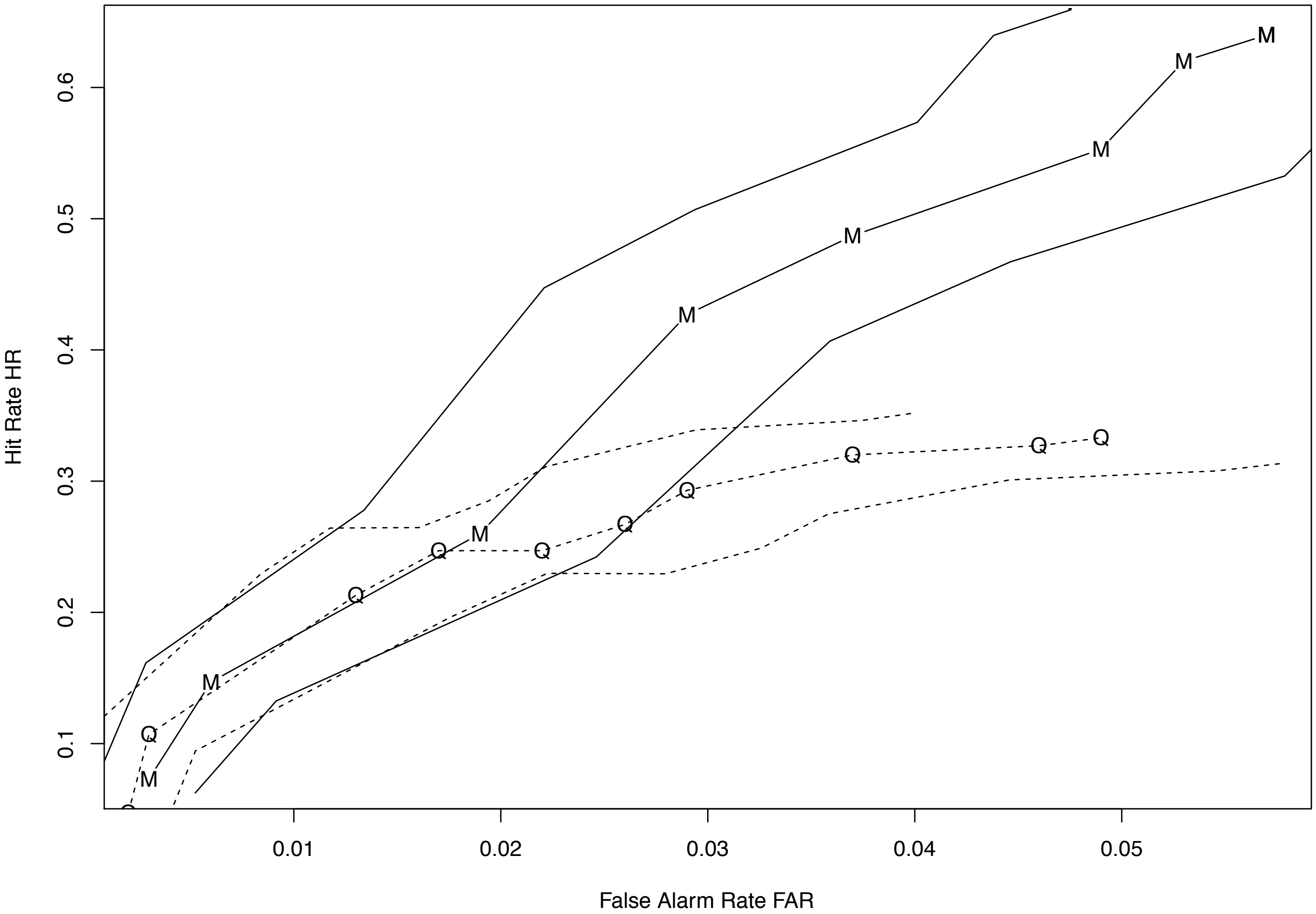
Data from MFW2012, p.372, Expt 1A: M=Simultaneous (solid), Q=Sequential (dash); limits of 1 standard error

Diagnosticity Ratio vs Expressed Confidence Level



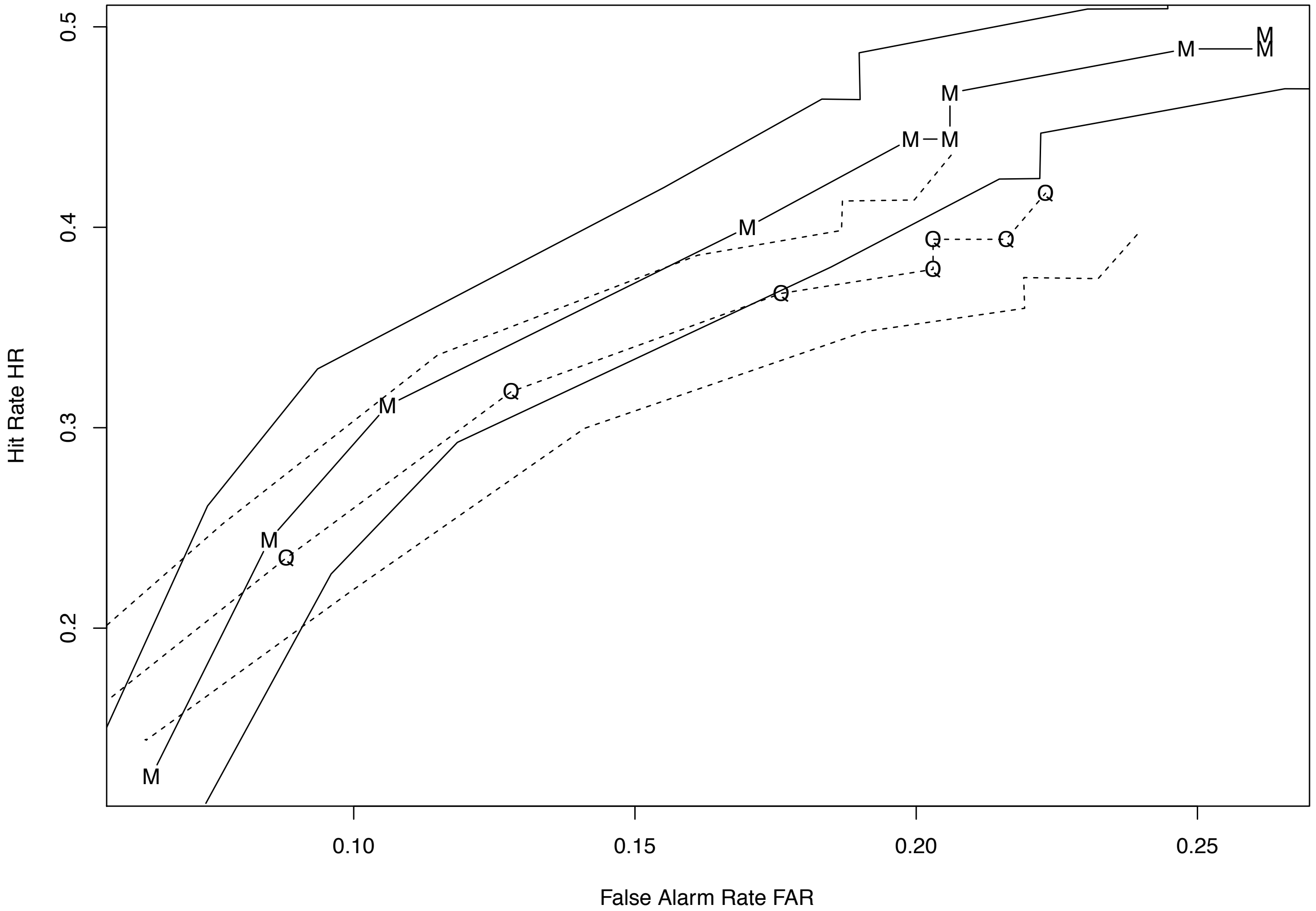
Data from MFW2012, p.372, Expt 2: M=Simultaneous (solid); Q=Sequential (dash); limits of 1 standard error

Expt 1A data: Tbl 3, MFW2012, p.372, n=598



Solid: Simultaneous (M); Dot: Sequential (Q). Bands show limits of 1 standard error.

Expt 2 data: Tbl 3, MFW2012, p.372, n=631



solid: Simultaneous (M); dash: Sequential (Q). Bands show limits of 1 standard error.

A “more complicated model”: Data from Carlson & Carlson 2014, *J Appl Research in Memory and Cognition*:

- 12 conditions:
  - 3 Procedures (SIM, target in #4; SEQ, #2; SEQ, #5)
  - 2 Weapon conditions (present, absent)
  - 2 Distinctive Feature conditions (present, absent)
- Compute confidence-based ROC for each condition
- Compare “Partial Area under ROC curve” (bigger = better)

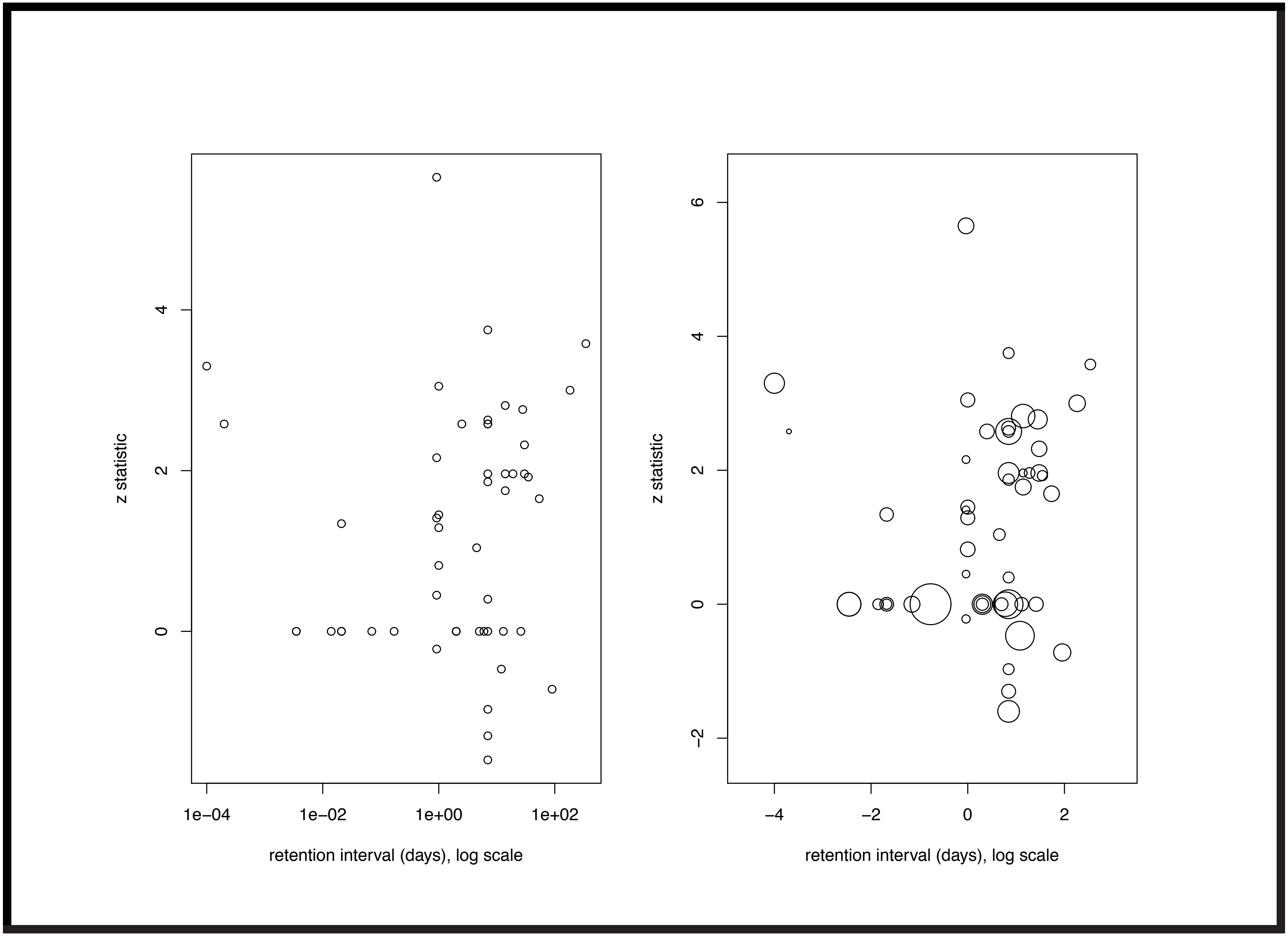
Model:

$$\log(pAUC) = \text{Proc Effect} + \text{Weapon Effect} + \text{Feature Effect} + \\ \text{(all 3 pairwise interactions)} + \text{error}$$

Source	df	SS	MS	F-stat	p-value
Procedure	2	8.04	4.02	1.129	0.470
Weapon	1	2.94	2.94	0.826	0.460
Feature	1	14.72	14.72	4.138	0.179
Proc × Weapon	2	0.59	0.30	0.083	0.923
Proc × Feature	2	10.41	5.21	1.463	0.406
Weapon × Feature	1	34.80	34.80	9.780	0.0.089
Residuals	2	7.12	3.56		

Other statistical approaches

1. Other models for comparing two procedures
2. Combining studies on effect of *retention interval*:
   - Deffenbacher et al. 2008: “Forgetting the once-seen face”
   - 39 studies (“long” vs “short’ retention interval)
   - “*compared the longest and shortest retention intervals in each study to determine effect size, we selected z scores for a difference between proportions as the primary dependent measure.*”
   - Plot “significance” of study vs. retention interval


z statistic
0
2
4
1e−04
1e−02
1e+00
1e+02
retention interval (days), log scale
z statistic
−2
0
2
4
6
−4
−2
0
2
retention interval (days), log scale

5. Final thoughts: Comparing two procedures

- *Accuracy* is likely to be related to *many* variables, both procedural (*system*) and situational (*estimator*)
  — and maybe *expressed confidence*

- Comparing two procedures should consider not just *diagnosticity ratio (PPV)* but also ratio related to accuracy of exclusions (NPV)

- More complicated statistical models may be needed; e.g., *Accuracy* (or AUC) = function of system/estimator variables

(Final thoughts, continued)

- "Eyewitness" can be thought of as a "binary classifier": Given true perpetrator or imposter, what is the proportion of correct / incorrect calls?

Person	"Yes, that's the one"	"No, not the one"
True Target	Correct	Incorrect
Imposter	Incorrect	Correct

- Huge literature on measuring performance, accuracy, reliability of binary classifiers, some may be relevant to EWI
- Much more research is needed

Some References

Identifying the Culprit: Assessing Eyewitness Identification, National Academies Press, 2014

Deffenbacher et al: Forgetting the Once-Seen Face: Estimating the Strength of an Eyewitness's Memory Representation, *J Exp'l Psych* 14(2): 139-150, 2008

Hastie T, Tibshirani R, Friedman JH, *The Elements of Statistical Learning*, 2nd ed, 2009

Kafadar K: Statistical Issues in Assessing Forensic Evidence, *International Statistical Review* 83(1):111–134, 2015

Wang F, Gatsonis C: Hierarchical models for ROC summary measures: Design and analysis of multi-reader, multi-modality studies of medical tests, *Statistics in Medicine* 27:243-256, 2008 (doi: 10.1002/sim.2828); model for log(AUC)

Wixted et al. (various papers)