Statistical Issues and Reliability of Eyewitness Identification as a Forensic Tool

JSM'2016, Chicago, August 2016

Karen Kafadar

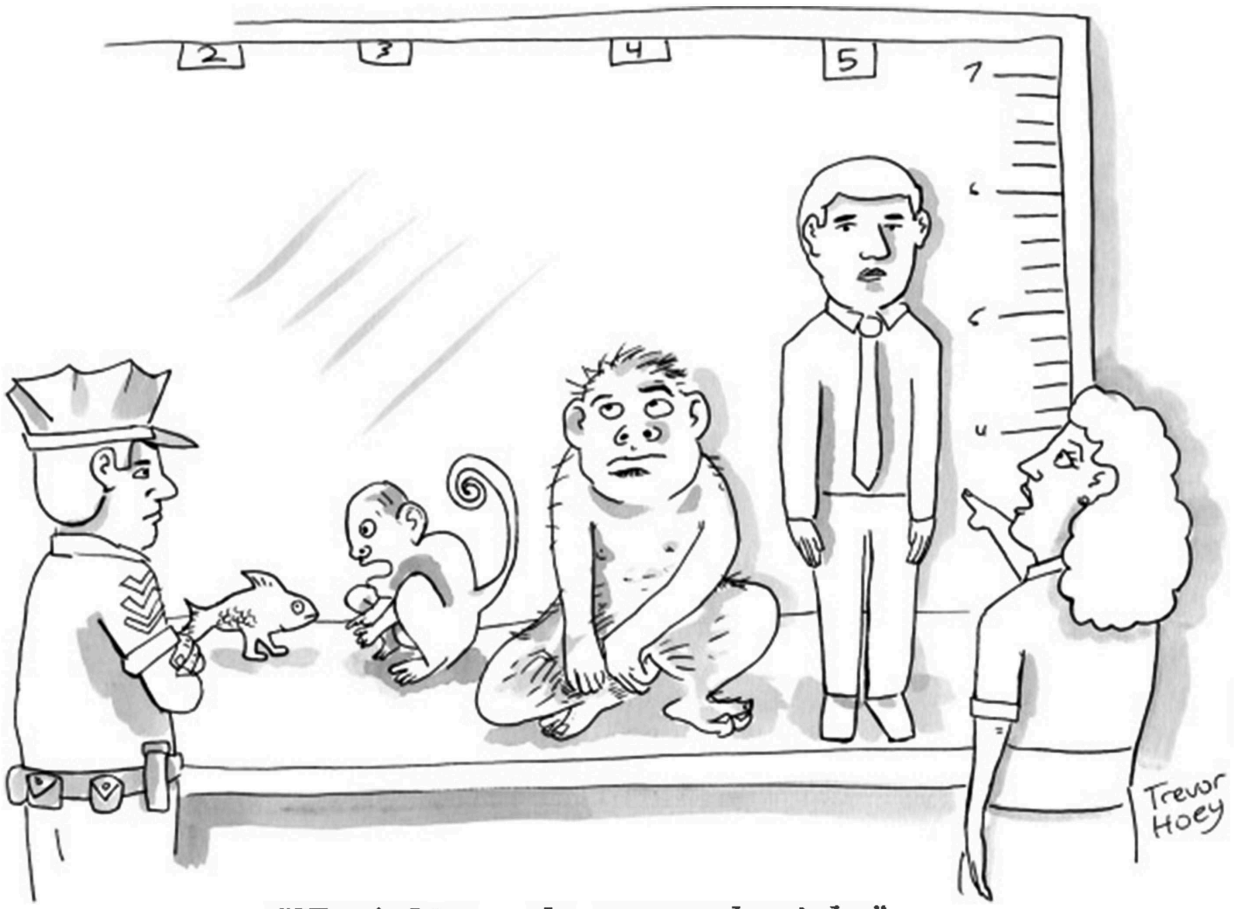
University of Virginia

kkafadar@virginia.edu

`http://statistics.as.virginia.edu/faculty-staff/profile/kk3ab`

Outline

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



"That's him—the one on the right."

From THE NEW YORKER, March 7, 2011

The task: Identify person in the incident (assualt, robbery, ...)

Binary decision, binary outcome

		Witness Classification	
		"Guilty"	"Innocent"
True Status	Guilty	True +	False −
of Suspect	Innocent	False +	True −

Ronald Cotton & Jennifer Thompson: *Picking Cotton*

- 1984 rape of Jennifer Thompson (college student in NC)
- Police sketch → Ronald Cotton
- 6 photos; Jennifer reluctantly chooses 2, then 1:
  "I think this is the guy."
- Detective: "You're sure?" — "Positive. Did I do OK?"
- Live lineup: Only Cotton was repeated from photo lineup
- Thompson selects Cotton: "looks the most like him"
- Courtroom: "100% sure. That's the guy who raped me."
- Convicted to life in prison + 54 years
- 1995: Cotton exonerated by DNA; Police arrest Bobby Poole.

NAS report, p10

John Jerome White

- Victim states: Attacker was "well built", "round face"
- 5-person live lineup: Selects White (middle)
- Courtroom: "*Do you see a person in the courtroom here today that was the person who came in your apartment that night?*"
- Victim: "*That's him (indicating).*"
- White convicted; 22+ years in prison; DNA exoneration 2007

B.L. Garrett, `http://harvardpress.typepad.com/hup_publicity/2011/03/understanding-eyewitness-misidentifications.html`

How do such mis-identifications occur?

- Theory: Memory is like a photograph, can be recollected
- Reality: Memory is fallible, influenced by many factors
- Each stage of memory is subject to degradation
  1. **Witnessed Event**: Data Acquisition Stage

     Accuracy depends upon:
     - acuity/fidelity of sensation
     - perception
     - memory storage
     - environmental conditions

2. **Memory of Event**: Data Handling Stage
   "Imprint" affected by:
   - Passage of time
   - Suggestion
   - Confidence inflators
3. **Retrieval of Event**: Data Presentation Stage (eg lineup)
   Accuracy depends upon:
   - acuity/fidelity of sensation
   - perception (e.g., of people in lineup)
   - quality of memory retrieval
   - environmental conditions (e.g., stress, pressures)

Passage of time $\Rightarrow$ further degradation

- Eyewitness testimony can be very useful and incredibly powerful in the courtroom
- Memory can play tricks: can be inaccurate, unreliable
- Innocence Project: 330 exonerations since 1989 from post-conviction DNA testing; 238 (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. Committee Charge (*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)

Eyewitness identification is affected by many variables

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)
5. Presence/absence of distinctive feature (e.g. scar)
6. Presence/absence of other distractions (e.g. people)
7. Common/Different race or ethnicity
8. Time between incident & report (*retention interval*)
9. Age of witness

Procedural aspects of EWI (*System variables*):

1. Conditions & protocols for **lineups**
2. Degree of similarities between fillers and suspect
3. Numbers and types of fillers
4. Declared number in lineup ("backloading": Horry & Palmer)
5. Nature of instructions (oral or written, short or long, ...)
6. Presence/absence of feedback
7. Other administrative behaviors (e.g. "blind")

etc.

Relative effects of multiple variables can be studied through well designed experiments Box, Hunter, Hunter 2005

- Variables not likely to operate independently
- Initial exploration: 2 levels on each of k variables
  (TP vs TA; Long vs Short instructions; Seq vs Sim; ...)
- 2^k conditions
- k large $\Rightarrow$ fractional factorial designs
- k super large $\Rightarrow$ super-saturated designs
- Taguchi (1980s): Choose levels of "system" (design) variables that are insensitive to settings of "estimator" (noise) variables

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

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)

   Conclusion: "Sequential > Simultaneous"

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*
   Conclusion: "Sequential < Simultaneous"

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 claim is "Yes, that's the one", PPV = probability that identified person is perpetrator
- *Negative Predictive Value (NPV)*: If claim is "No, not the one", NPV = probability that 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:

$$PPV = 1 \,/\, (1 + OR/DR) = 1 \,/\, (1 + OR/LR_{+})$$

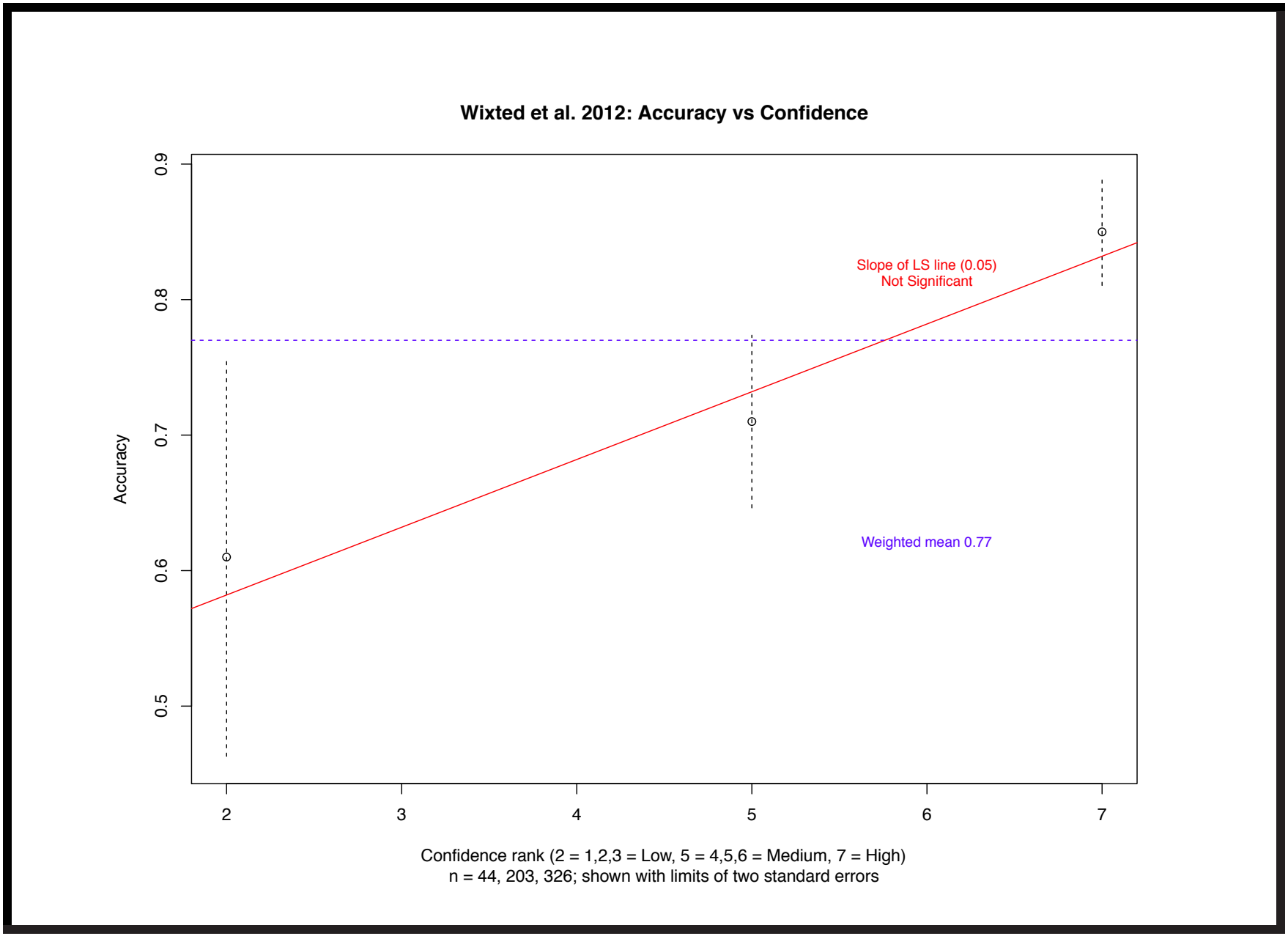
  where $OR = (1 - p)/p$, p = prevalence ("base rate")
- So *higher DR* (for same p) $\Rightarrow$ *higher PPV*

- Recall: $DR = LR_+ = Sens/(1 - Spec)$
  $=$ Pr{'+' | True +} / Pr{'+' | True −} (should be high)
- $PPV = 1\,/\,(1 + OR/DR) = 1\,/\,(1 + OR/LR_+)$
- $LR_- = (1 - Sens)/Spec =$ Pr{'-' | True +} / Pr{'-' | True −} (should be low)
- NPV relates to correct exclusions:

$$NPV = 1\,/\,(1 + LR_-/OR)$$

- So *smaller* LR_- (for same p) $\Rightarrow$ *higher* 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
0.5
0.6
0.7
0.8
0.9
2
3
4
5
6
7
Slope of LS line (0.05)
Not Significant
Weighted mean 0.77
Confidence rank (2 = 1,2,3 = Low, 5 = 4,5,6 = Medium, 7 = High)
n = 44, 203, 326; shown with limits of two standard errors

- 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)

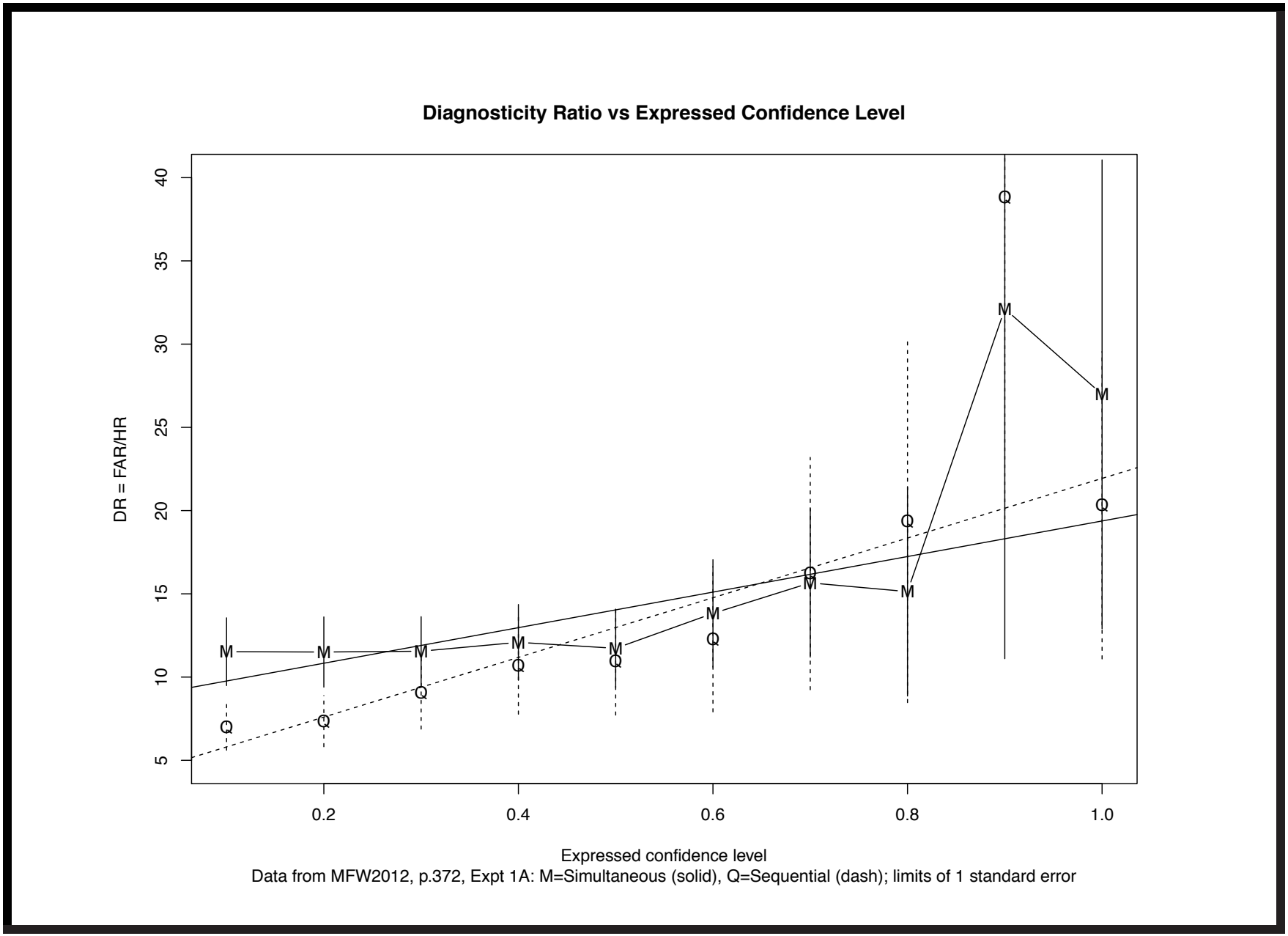
Using Confidence-Accuracy Relationship

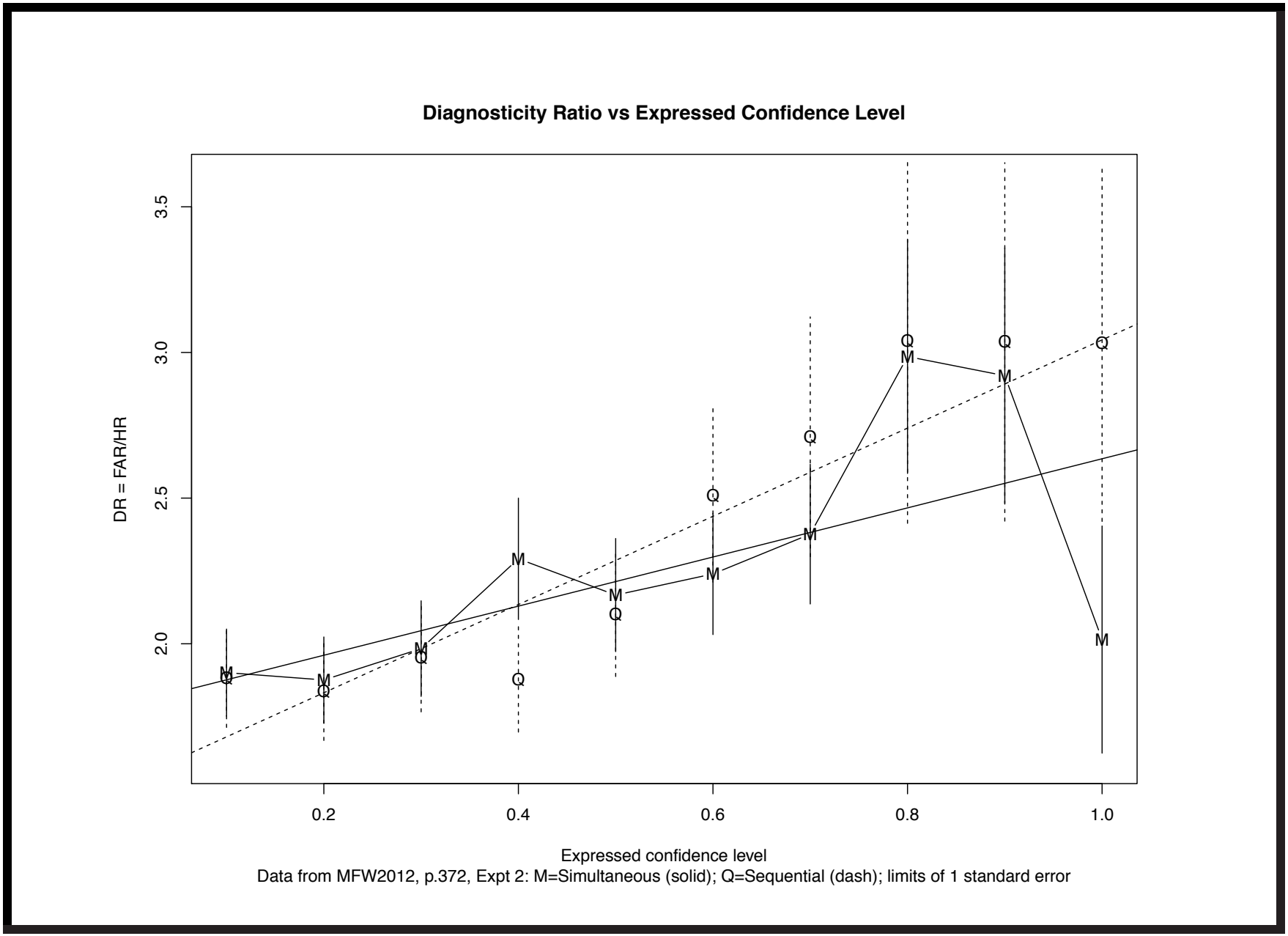
If you believe confidence is related to accuracy:

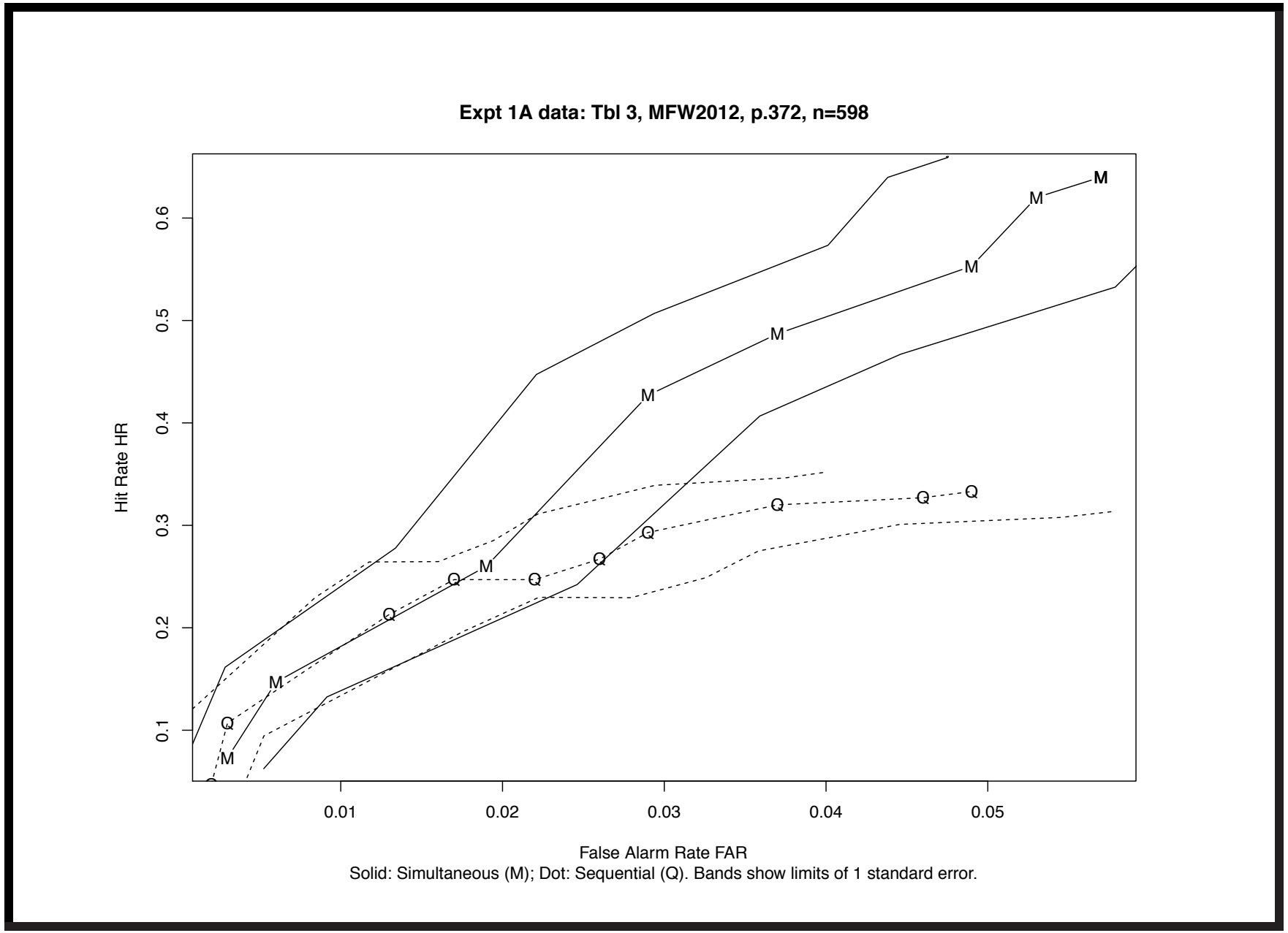
- consider calculating $DR = HR/FAR$ as a function of *Expressed Confidence Level (ECL)*
- Split the sample participants into categories of ECL (those who expressed 10%, ..., 90% confidence); calculate DR for each ECL category
- even better: Plot HR vs FAR for different ECLs
- ROC curve = Receiver Operating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

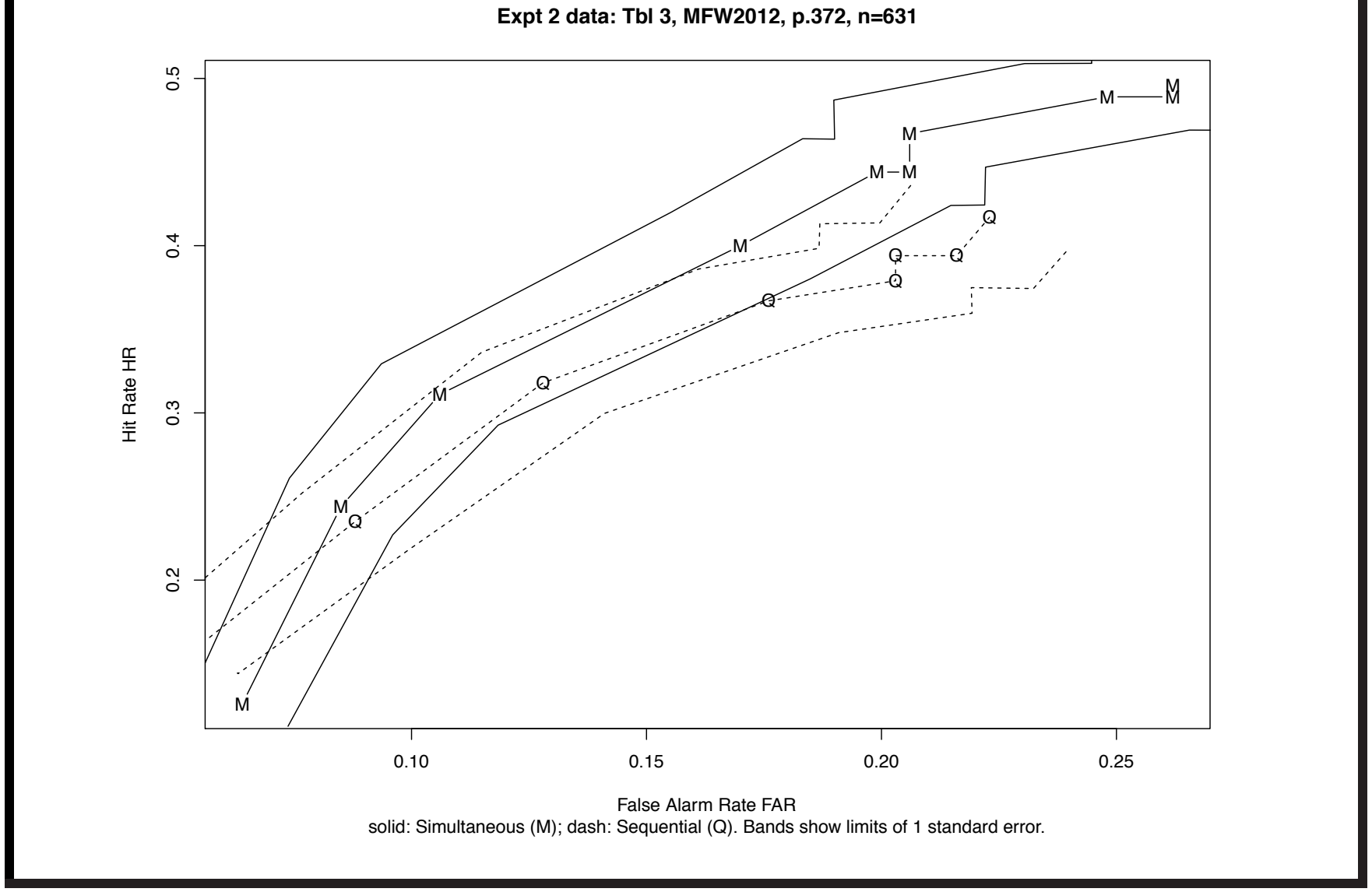
Problem: Data points (HR, FAR) have uncertainty!

- John Tukey:
  "*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 "Simultaneous" versus "Sequential"?


Diagnosticity Ratio vs Expressed Confidence Level
DR = FAR/HR
5
10
15
20
25
30
35
40
0.2
0.4
0.6
0.8
1.0
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
DR = FAR/HR
2.0
2.5
3.0
3.5
0.2
0.4
0.6
0.8
1.0
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
Hit Rate HR
0.1
0.2
0.3
0.4
0.5
0.6
0.01
0.02
0.03
0.04
0.05
False Alarm Rate FAR
Solid: Simultaneous (M); Dot: Sequential (Q). Bands show limits of 1 standard error.


Expt 2 data: Tbl 3, MFW2012, p.372, n=631
Hit Rate HR
0.2
0.3
0.4
0.5
0.10
0.15
0.20
0.25
False Alarm Rate FAR
solid: Simultaneous (M); dash: Sequential (Q). Bands show limits of 1 standard error.

Almost surely, resulting uncertainty is underestimated, because

1. ECL can change (e.g., person says "20%" now; "70%" later)
2. Person who says "50%" shows up in the calculations for "at least 10%", ..., "at least 40%"
3. Responses are not independent (especially if same "eyewitness" is shown more than one scenario)
4. Likely more variability in the proportions than simple binomial variation

Gronlund & Neuschatz (*J Appl Res Mem Cognition* 2014, p55): "*computing d' (and related measures) relies on underlying assumptions (e.g., normal evidence distributions), which usually are not met in an eyewitness experiment.*"

Use of ECL in ROC

The recognition of a variable as influencing DR is a step forward. But the use of a variable such as ECL raises many issues:

1. **Highly variable**: For same condition, an EW might say "somewhat confident" one day and "very confident" another day (cf. fingerprint studies by Itiel Dror).
2. ECL responses in lab experiment likely to be much different in **real-life, highly stressful conditions**, very difficult (if not impossible) to replicate in an academic setting.
3. "50%" may mean something different under one procedure (e.g., "sequential") vs another (e.g., "simultaneous"): the variable (ECL) on which the ROC is based could depend on the procedures themselves which the ROC curves are designed to compare.
4. More than ECL is likely to affect DR.

5. Real-life: One cannot ask EW to quantify "ECL" as 10%, 20%, ...; translation of verbal response can differ among agents.

6. ECL may, or may not, be an accurate measure of "response bias" (Lampinen).

7. ECL may, or may not, be related to accuracy (Roediger et al).

8. Lack of independence in data if lab experiments use same individuals in multiple ECL categories or in multiple tasks.

9. Wells et al.: ROC designed to compare only 2 levels (e.g., seq/sim) but comparison may involve 3+ (if target is present or absent). Address criticism by including influential variable in model ("target present/absent"; cf. A. Luby SAMSI poster, 8/30/15).

Another measure of confidence?

Committee Report: Consider alternatives to ROC for comparing EWI procedures

cf. Appendix: Consider alternative analyses (p150)

"*If a study is sufficiently large, one could develop a performance metric for each participant in the study corresponding to each of these conditions [such as] log(AUC) ... [or] log odds of a correct decision; e.g.,* $\log(HR/(1-HR))$ *or* $\log((1-FAR)/FAR)$"

Example using data from Carlson & Carlson 2014, *J Appl Res in Memory and Cognition*:

Data from Carlson & Carlson 2014:

- 12 conditions:
  - 3 Procedures (Sim, target #4; Seq, #2; Seq, #5)
  - 2 Weapon conditions (present, absent)
  - 2 Distinctive Feature conditions (present, absent)
- Compute ECL-based ROC for each condition
- Compare pAUC (but see justified criticisms in SD Walter *SIM* 2005 and Lampinen *JARMAC* MS 2015: "partial" arises here from data outcome, not from subject-specific concerns)

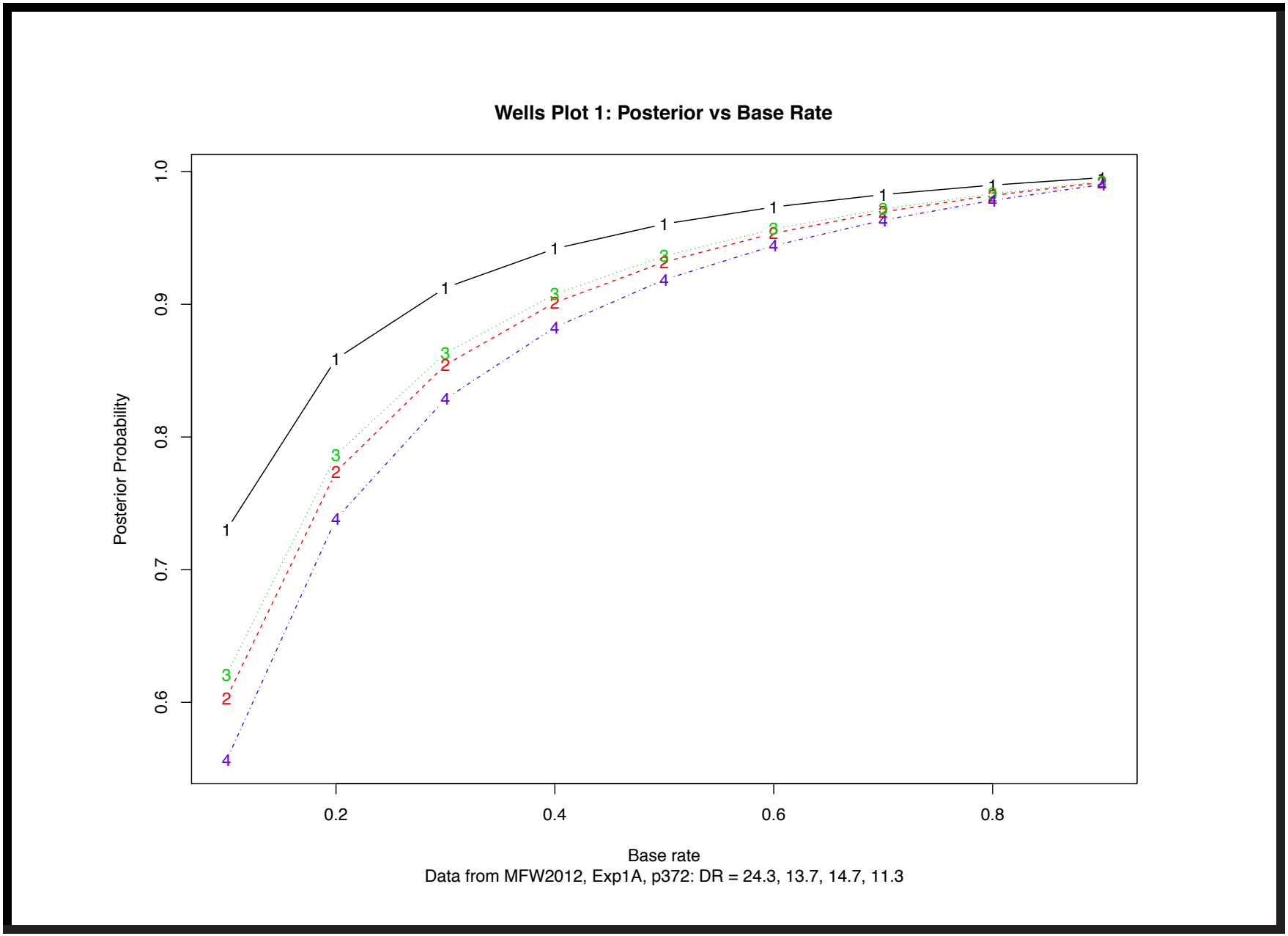
Model:

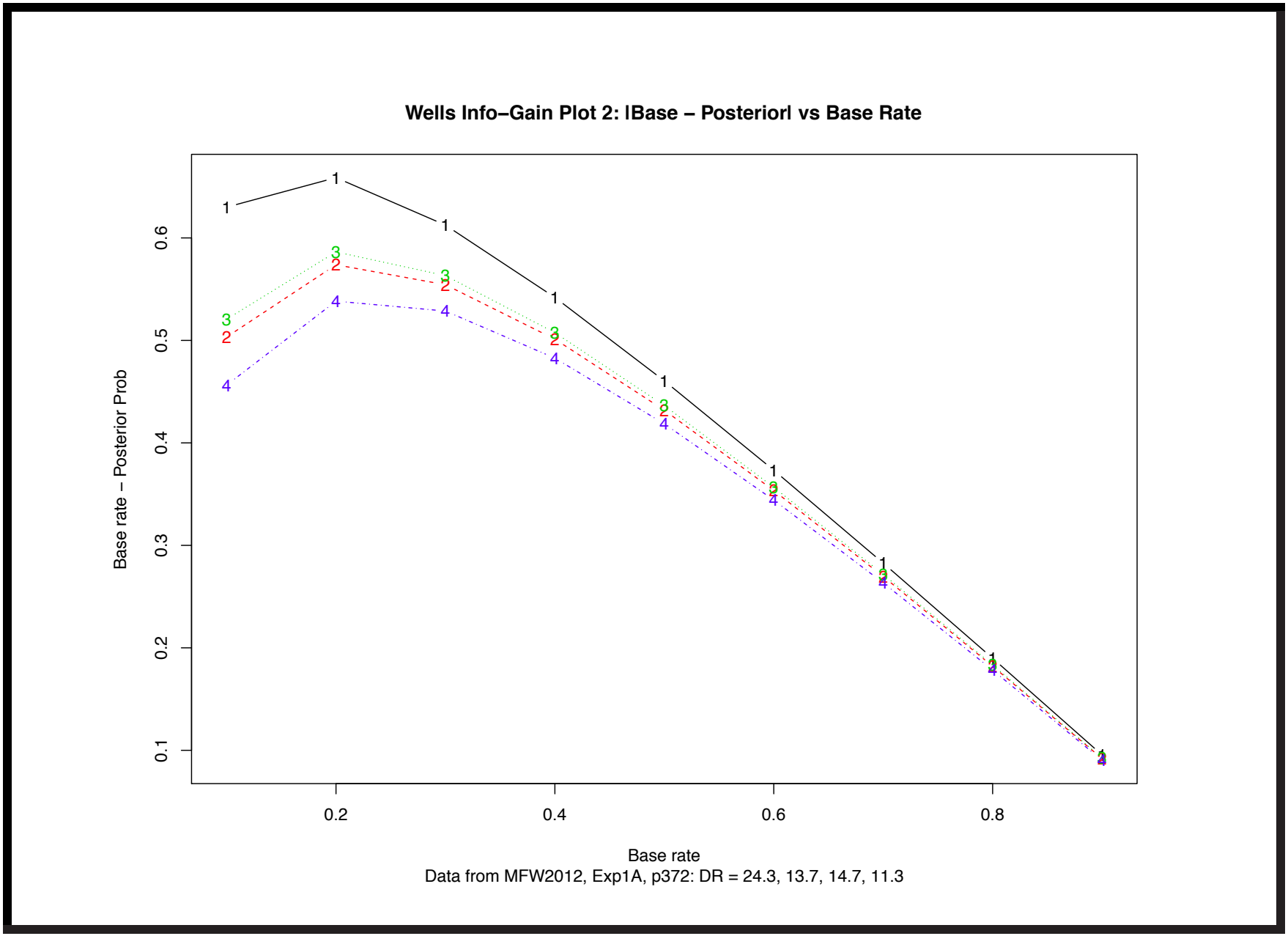
$$\log(pAUC) = \text{Proc Effect} + \text{Weapon Effect} + \text{Feature Effect} + (\text{all 3 pairwise interactions}) + \text{error}$$

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
Procedure×Weapon	2	0.59	0.30	0.083	0.923
Procedure×Feature	2	10.41	5.21	1.463	0.406
Weapon×Feature	1	34.80	34.80	9.780	0.089
Residuals	2	7.12	3.56		

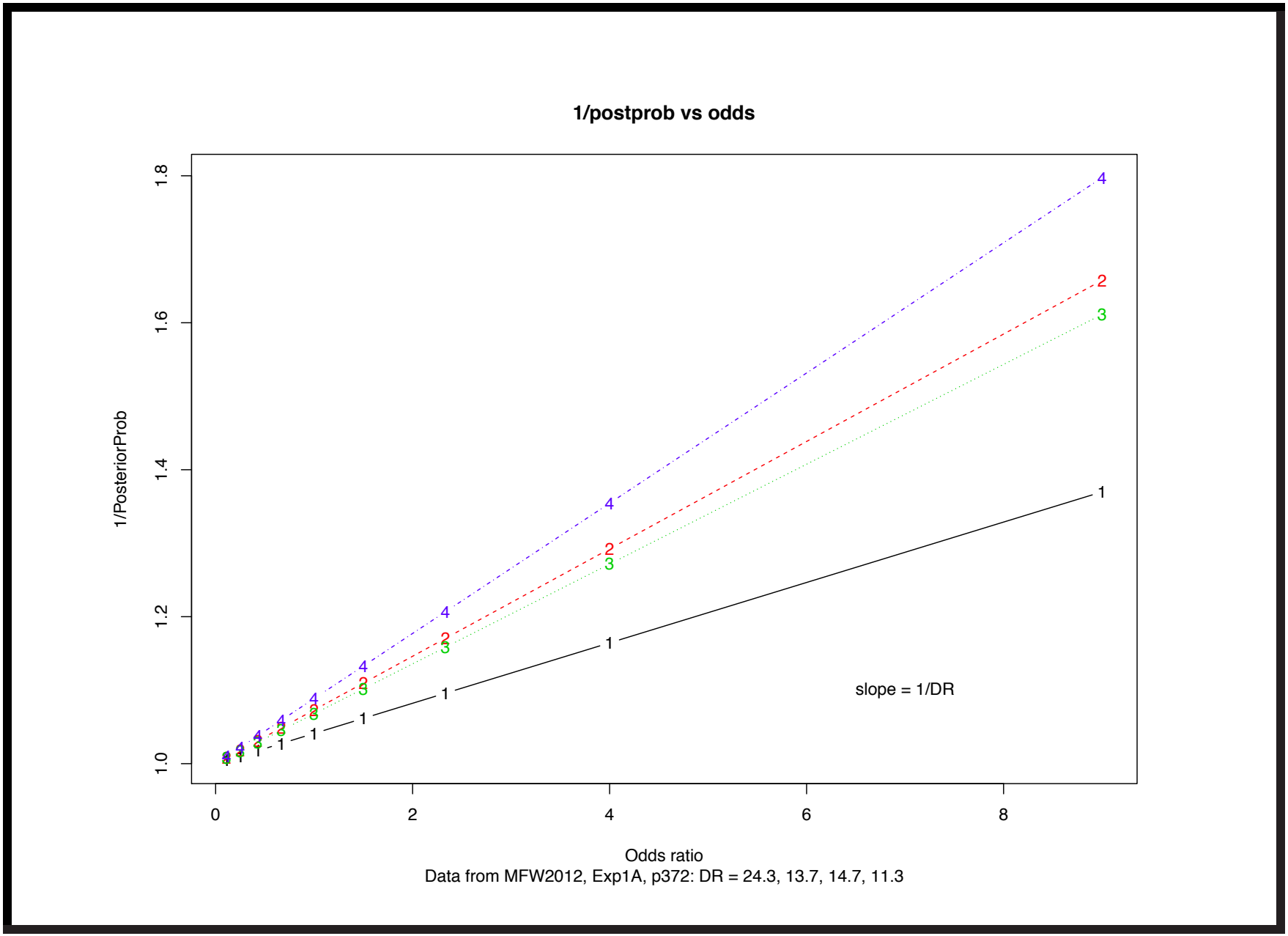
We need other plots for comparing procedures

- Wells, Yang, Smalarz 2015 *Law & Human Behavior*: Plot posterior probability vs "base rate" p
- WYS2015 also suggest Info-Gain plot: $|p - PostProb|$ vs p
- Recall: posterior probability $= PostProb =$
  $Sens \cdot p/(Sens \cdot p + (1 - Spec) \cdot (1 - p))$
  $= 1/(1 + OR/DR)$, $DR = Sens/(1 - Spec)$, $OR = (1 - p)/p$
- Use data in Mickes, Flowe, Wixted 2012 (MFW2012): 10 DRs ($ECL =$ "10%", ..., "100%")


Wells Plot 1: Posterior vs Base Rate
Posterior Probability
1.0
0.9
0.8
0.7
0.6
0.2
0.4
0.6
0.8
Base rate
Data from MFW2012, Exp1A, p372: DR = 24.3, 13.7, 14.7, 11.3


Wells Info–Gain Plot 2: |Base – Posterior| vs Base Rate
Base rate – Posterior Prob
0.1
0.2
0.3
0.4
0.5
0.6
0.2
0.4
0.6
0.8
Base rate
Data from MFW2012, Exp1A, p372: DR = 24.3, 13.7, 14.7, 11.3

- Both plots involve curves. Curves are hard to interpret.
- Note: $1/PostProb = 1 + OR/DR =$ linear function of OR.
- Plot $1/PostProb$ vs $OR = (1-p)/p$; slope $= 1/DR$ (smaller slope is better); see difference in slopes as indications of effects of ECLs.


1/postprob vs odds
1/PosteriorProb
1.0
1.2
1.4
1.6
1.8
0
2
4
6
8
Odds ratio
slope = 1/DR
Data from MFW2012, Exp1A, p372: DR = 24.3, 13.7, 14.7, 11.3

We need other methods for comparing procedures

- Problem is one of *characterizing accuracy of a binary classifier.*
- Given a scenario, each "EW" is a binary classifier: the EW is presented either a target or a filler, says either "yes" or "no".
- Many methods have been proposed to evaluate binary classifiers; the most obvious is logistic regression:

  log(accuracy / (1-accuracy)) = function of many variables

- Alternatives: since (LR_+, LR_-) related to (PPV, NPV), consider bivariate linear model for

  $$(LR_+,\ 1/LR_-) \text{ or } (\log(DR),\ -\log(LR_-))$$

  or bivariate logistic regression models for

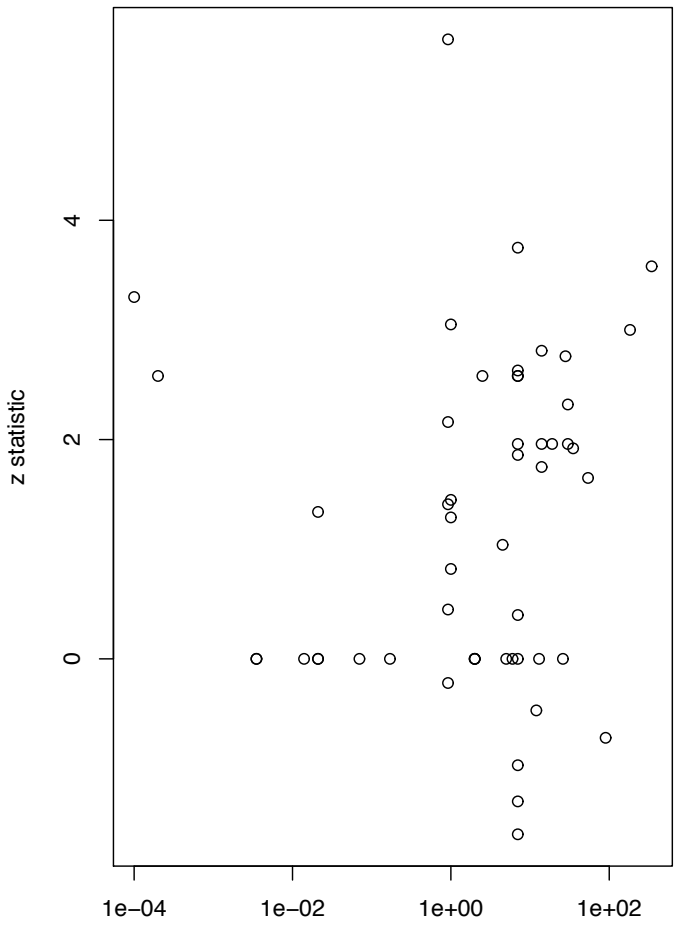
  log(HR / (1-HR)), log(FAR / (1-FAR))

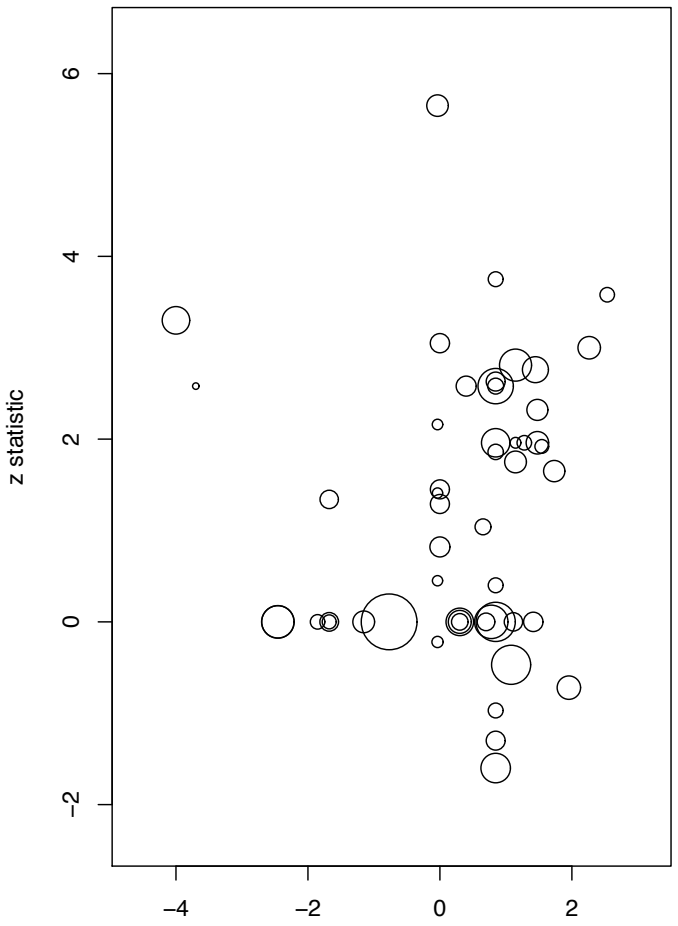
Other methods of comparing methods:

- Cross-validation (CV), bootstrap, ...
- Report acknowledges other methods but that "they have not been vetted" for EWI
- How distinctive is the EWI scenario that would render the decades of research on logistic regression, CV, bootstrap, ... irrelevant to this problem?
- Experimental designs that involve more factors
- Perhaps a model for the bivariate outcome that incorporates *both PPV and NPV*: $(\log(LR_+), -\log(LR_-))$
- etc.

Committee used meta analyses to suggest variables that may influence accuracy of EWI.

- Ex: Deffenbacher et al. 2008, "Forgetting the once-seen face"
- 39 studies ("long" vs "short" retention interval)
- "*compared longest & 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

Committee Findings

- Take EW statement as soon as possible after incident
- "Blind" lineup: one sets photos in envelopes, another administers lineup
- Record EW remarks following statement ("In your own words, how confident are you?")
- Administrators should provide no feedback
- Studies on effects of jury instructions
- No recommendation on "sequential" vs "simultaneous"
- Other performance metrics for comparing procedures

The Novel New Jersey Eyewitness Instruction Induces Skepticism but Not Sensitivity: Papailiou et al., *PLoS One*, 9 Dec 2015

- Effect of NJ's jury instructions to notify jurors of EWI limitations (memory, effects of police feedback, use of blinding)
- 335 'jurors' (Amazon Turk) watch 35-min murder trial video, "weak" or"strong" ID quality, NJ or standard instructions
- Metric: % convict
- Results:

	Std	NJ
Strong	26%	12%
Weak	23%	9%

- All CIs overlap; Odds Ratio 2.55 (1.37, 4.89) for Std/NJ

Yokum et al.

- About $335/4 = 84$ "jurors" per group
- 35-min video $\neq$ day-long trial
- Lower convict rate due to being more cautious, or better understanding of EWI limitations?
- Interesting study that bears repeating under other conditions ("live" trial, more pools of "jurors", ...)

"Perceptual expertise in forensic facial image comparison,"
D White et al, *Proc Royal Soc B* 282:20151292 (Sep 2015)

- "*Study Reveals Forensic Facial Examiners Can Be Near Perfect*"
  `http://www.nist.gov/itl/iad/20150928facial.cfm`
- Task: View 2 pictures: Same or different person?
- 27 facial ID examiners, 14 controls (all govt employees), 32 UNSW undergraduates (FISWG mtg attendees in May 2014)
- 3 tests to judge forensic examiners' performance:
  - Glasgow Face Matching Test (GFMT): 300 pairs of high-quality "mug-shot" images
  - Person Identification Challenge Test (PICT): "selected to have no computationally useful identity information in the face ... leading algorithms make 100% errors on this set"

  - Expertise in Facial Comparison Test (EFCT): "selected pairs of images for identity comparisons that were challenging for computers and untrained humans based on data from pilot work and previous evaluations of human and computer face matching performance": upright vs inverted, 2 vs 30 seconds view time

- Decision time unbounded
- Results: Examiners > Controls > Students; Accuracy > 90%; performance ↑ with more examiners
- Facial image comparison (decision immediately after seeing image paris) much different from EWI in crime scene (stress, threat, fright, weapon, lightening, etc.)

Committee Findings & Recommendations

1. Train LE officers in EWI: variables affectingvision & memory, non-leading questions, avoid suggestiveness, ...
2. Double-blind lineup & photo array procedures
3. Standardize witness instructions
4. Document witness confidence judgment at 1st ID
5. Videotape witness ID process
6. Conduct pre-trial judicial inquiry
7. Inform juries of prior identifications
8. Use scientists in expert testimony
9. Use clear instructions for juries
10. Establish educational research initiative on EWI

5. Summary: Characterizing EWI performance

- *Accuracy* is likely to be related to *many* variables, both procedural (*system*) and situational (*estimator*) — and maybe *expressed confidence*
- More complicated statistical models would be needed: *Accuracy* (or AUC) = function of system/estimator variables
- Comparing two procedures involves not just *diagnosticity ratio (PPV)* but also ratio related to accuracy of exclusions (*NPV*)
- "Eyewitness" can be thought of as a "binary classifier": Given true perpetrator or imposter, what is the proportion of correct (incorrect) calls?
- Relevant literature on characterizing reliability & accuracy of binary classifiers may apply to EWI performance

"***Statistics means never having to say you're certain***"

John W. Tukey (1915-2015):

"*Often a problem can benefit from more than one approach; try several*"

"*What we do tomorrow may differ from what we do today*"

"*Finding the question is often more important than finding the answer*" (cited by Brillinger 2002, p1571)

Some References

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

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)

John Wixted et al. (various papers)