CBSA ASFC

PROTECTION • SERVICE • INTEGRITY PROTECTION • SERVICE • INTÉGRITÉ

Error, Confidence & (Un)certainty

Deconstructing Authorship Opinions using a Forced-call Testing Protocol

R. Brent Ostrum

Senior Forensic Document Examiner

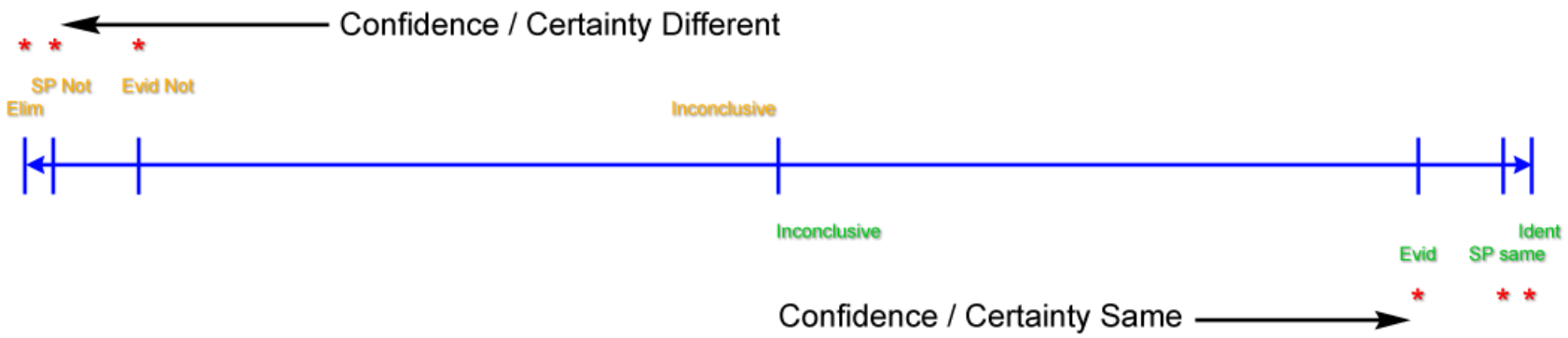
Study Purpose

- Conclusion scales have been a topic of discussion for many years in the FDE community

- Primary goals:
  - Attempt to validate the conclusion scale – explore how it 'works'
  - Explore the use of a 'forced-call' decision protocol for signature assessments

‘Traditional’ Conclusions

- ASTM/SWGDOC: a set of up to 9 'standard' definitions

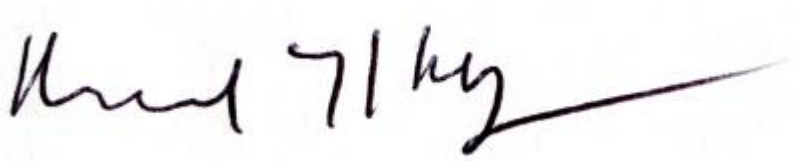


- ‘Levels’ intended to reflect confidence of the examiner
- Confidence, in this context, is intended to address concerns about ‘potential error’ in the conclusion

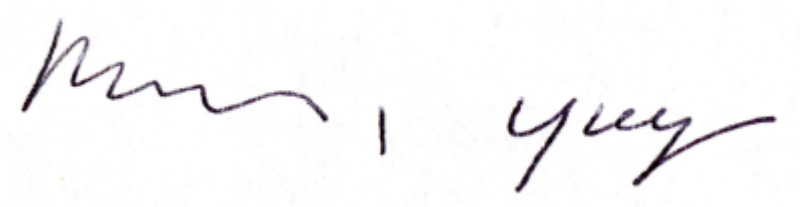
Study Design

- St2ar S-03 skill-task test
  - 18 specimen, 36 genuine Q, 5 disguise Q, 19 simulations Q

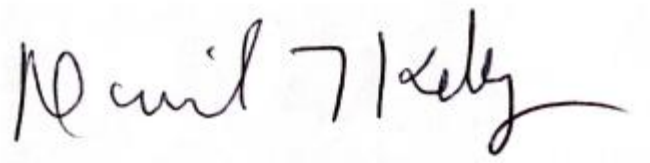
K-08 (18 Specimen)



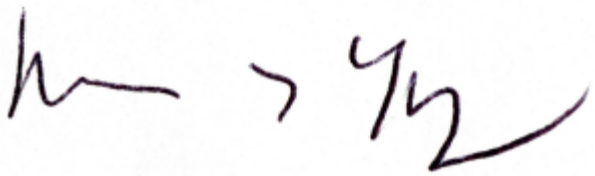
Q-02 (19 Simulation)



Q-27 (5 Disguise)

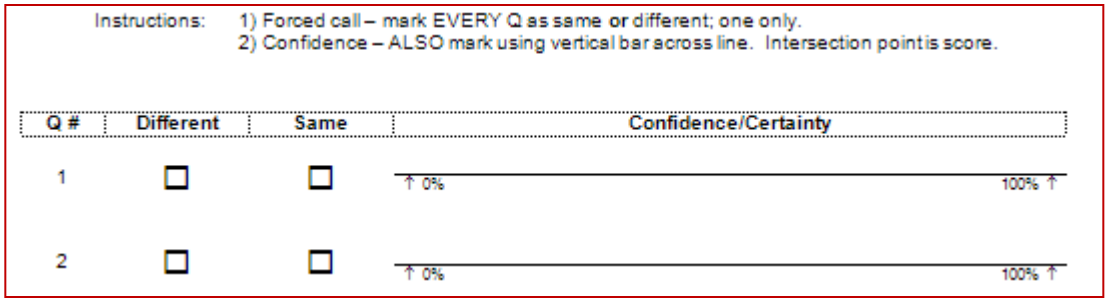


Q-09 (36 Genuine)



Instructions and 'Forced Call' Procedure

- Two-part process
  - 'Scoring'


Instructions: 1) Forced call – mark EVERY Q as same or different; one only.
2) Confidence – ALSO mark using vertical bar across line. Intersection point is score.

Q #	Different	Same	Confidence/Certainty
1	☐	☐	↑ 0% ——— 100% ↑
2	☐	☐	↑ 0% ——— 100% ↑

- Key metrics:
  - Correct vs misleading (ER) calls – overall and by signature type
  - Confidence ratings
    - Overall pattern
    - Confidence vs elicited ER

- No formal statistical analyses or comparisons
  - More a test of feasibility and general results
  - Performance of FDE vs laypersons (latter is our baseline)

Test Subjects

- Forensic Document Examiners:
  - CBSA: 9 examiners
  - Training:
    - Several different programs but deemed equivalent
  - Experience:
    - From 3 to 25+ years
  - Certification:
    - Four examiners ABFDE
  - All conduct casework in this area

  - Other: 24 examiners
    - Limited biographic data available

- Laypersons:
  - CBSA laboratory employees: 14
  - Various positions in lab:
    - Admin, Math/data, Chemists/examiner
  - Education:
    - College, BSc, MSc, PhD
  - Self-rated knowledge of handwriting examination (0-10):
    - 10/14 self-rated 0
    - 2/14 self-rated 1
    - 1/14 self-rated 2
    - 1/14 self-rated 3

    - No observed performance difference by self-rating

'PILOT'

- Many significant limitations in the study
  - Small number of subjects
  - Novel approach for FDE subjects
  - Signature – only one Q writer
  - Non-random samples – 'convenience' and self-selected

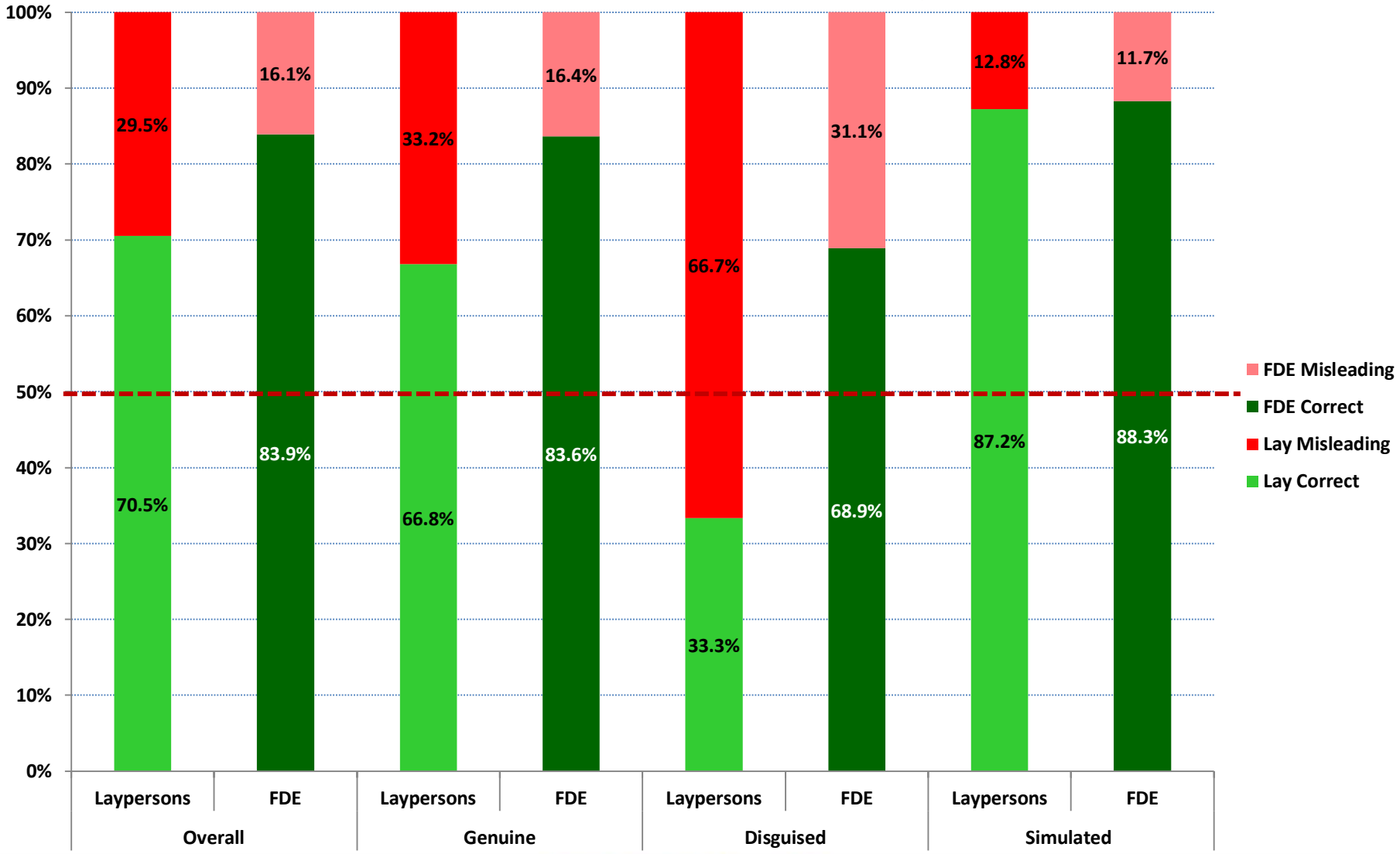
- Other issues
  - Compensation for laypersons – none

- Some notes:
  - Numbers shown here may change in any future presentation
  - Still collecting data and statistics will likely change

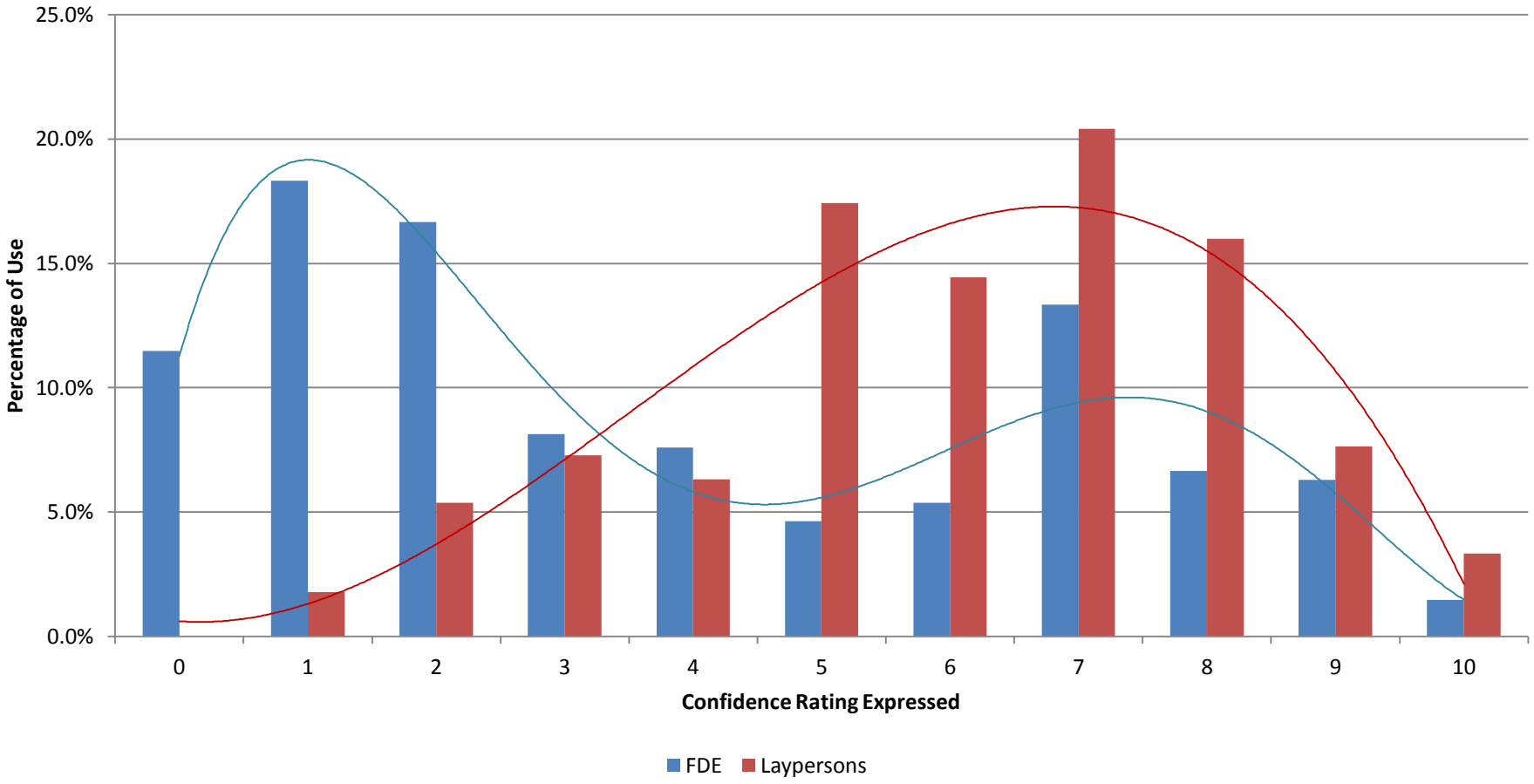
Hypotheses / Expectations

- Error rate (ER) – “elicited miscalls”
  - Protocol should result in higher ER
  - FDE ER < laypersons ER
  - ER will vary by signature type (for both groups)
- Confidence/certainty
  - Expressed uniformly? Any ‘preference’ or skew?
  - FDE confidence < laypersons confidence
- Relationships and calibration
  - Inverse relationship between Confidence and ER (for FDEs)

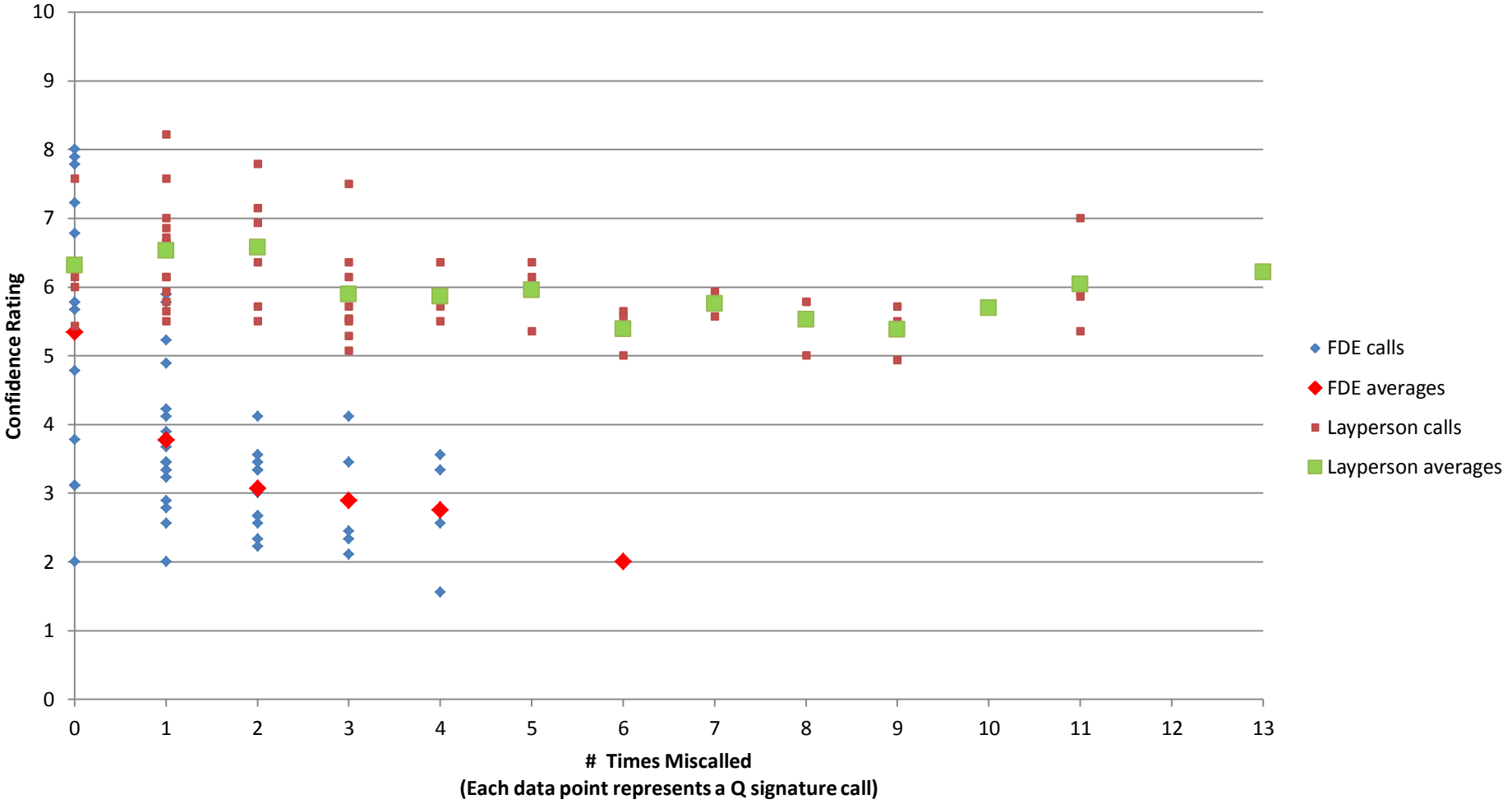
Correct vs Misleading calls - Forced Calls



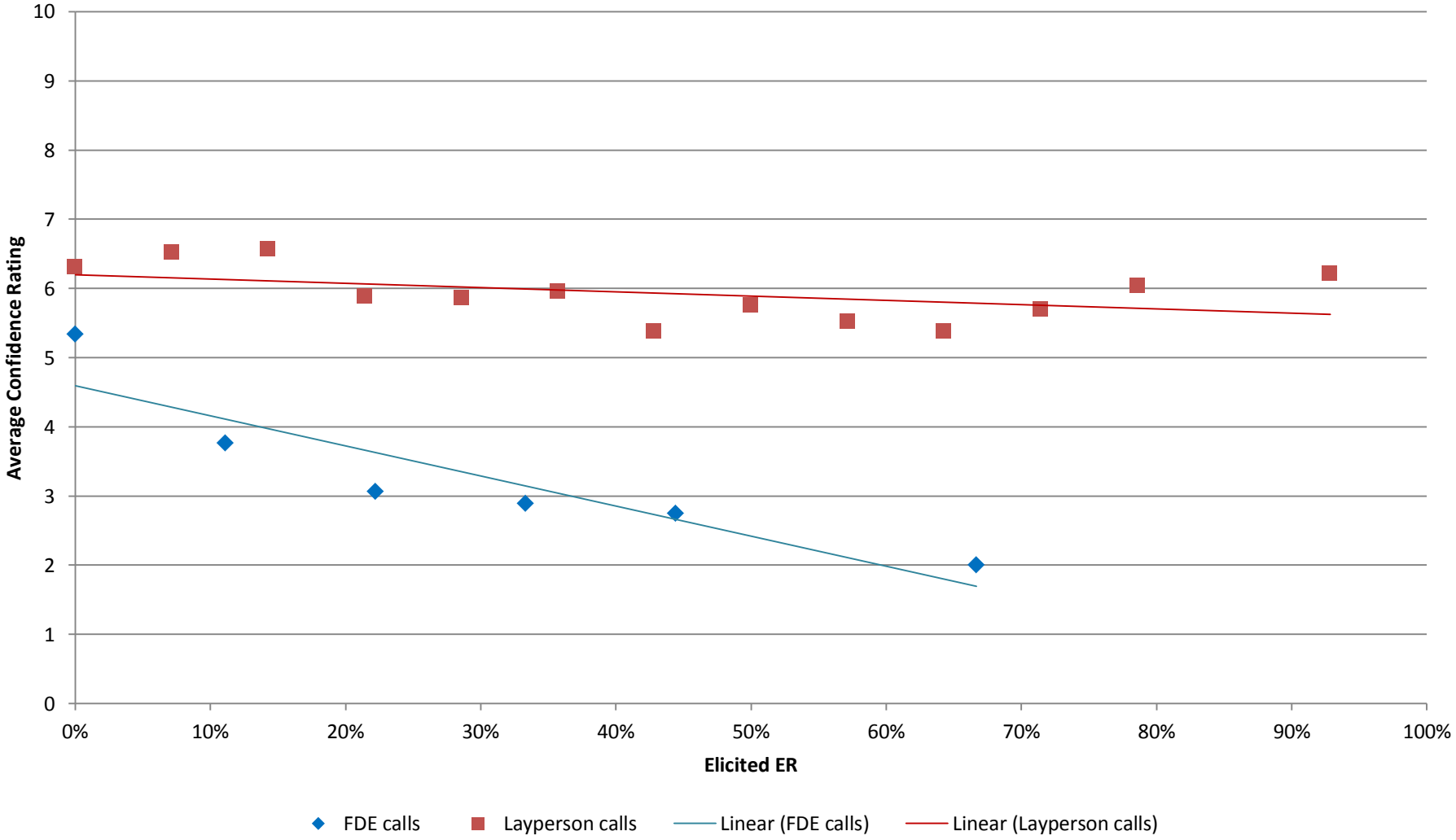
Confidence Ratings (proportion used)



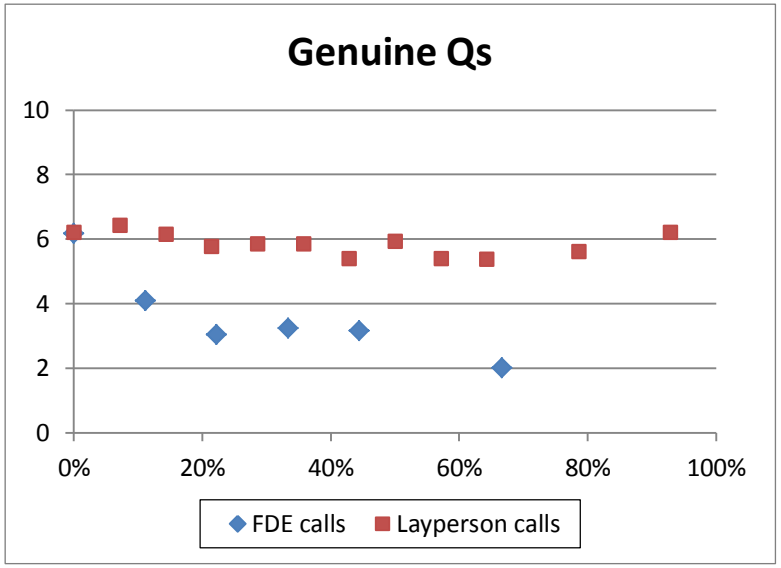
Confidence Rating vs Miscalls
(Individual Q signatures, Forced calls)

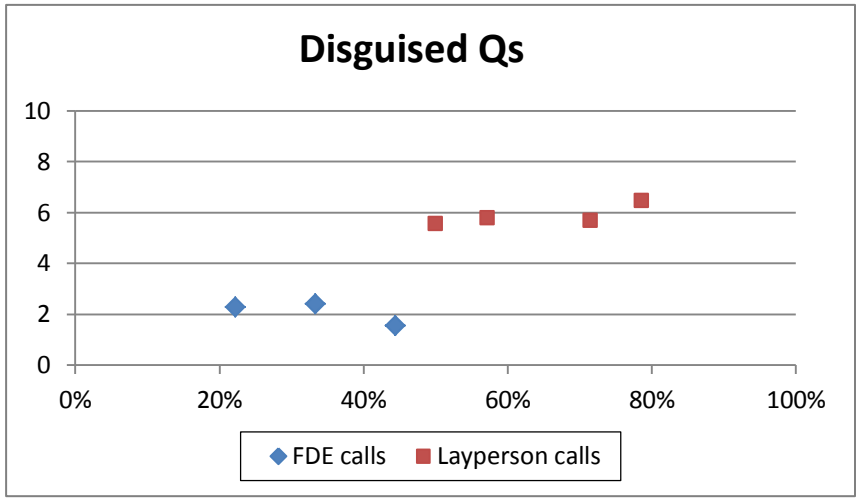


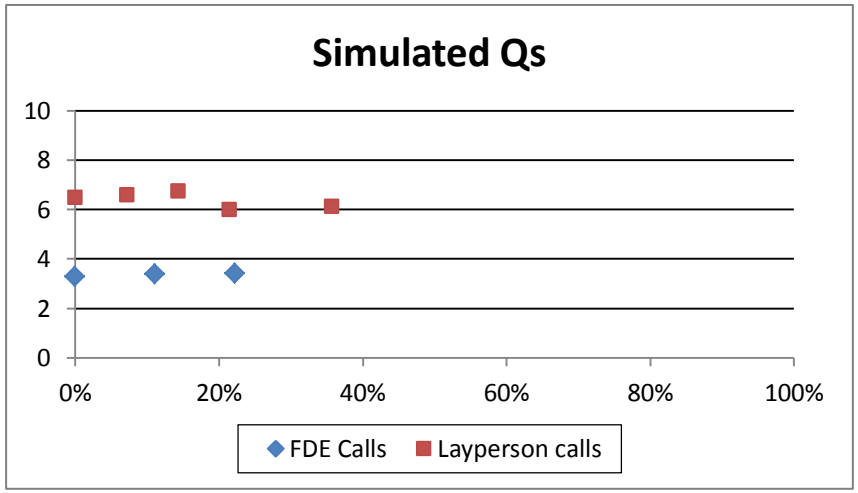
Average Confidence versus Error
(Overall - all signatures, Forced calls)



Confidence versus Error (Forced)







Summary FDE vs Laypersons

- Laymen are generally much more confident than FDEs
- Laymen are ‘always’ confident, even when in error

- FDE confidence relationship to error (by type)
  - Overall, there is an inverse relationship
  - Genuine Q show this relationship
  - Disguised and Simulated Q show little, if any, relationship
    - Both Disguised and Simulated Q expressed with low confidence

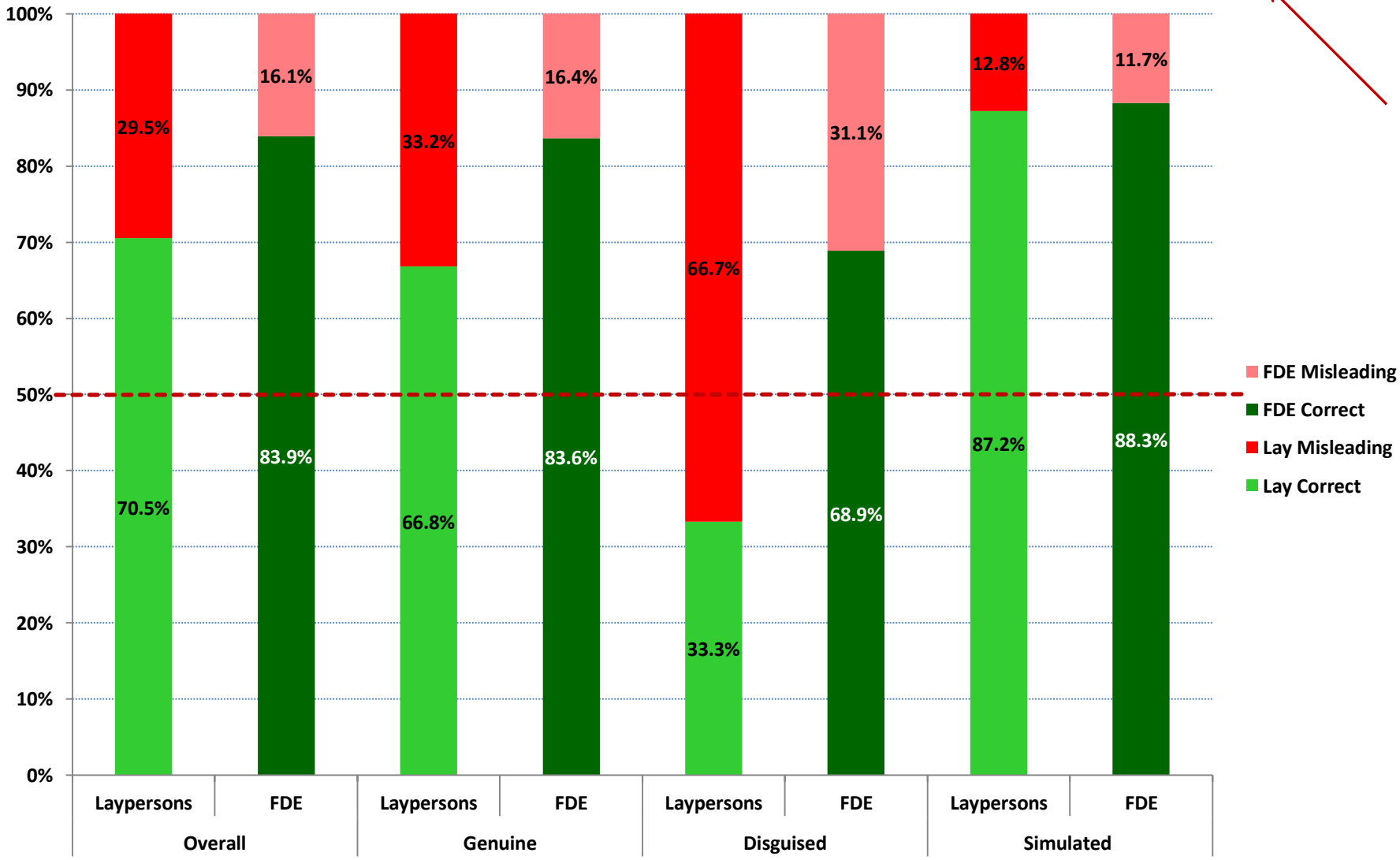
- FDE ER (forced) is higher than in other studies
- FDE ER is lower than layperson ER (even when forced)

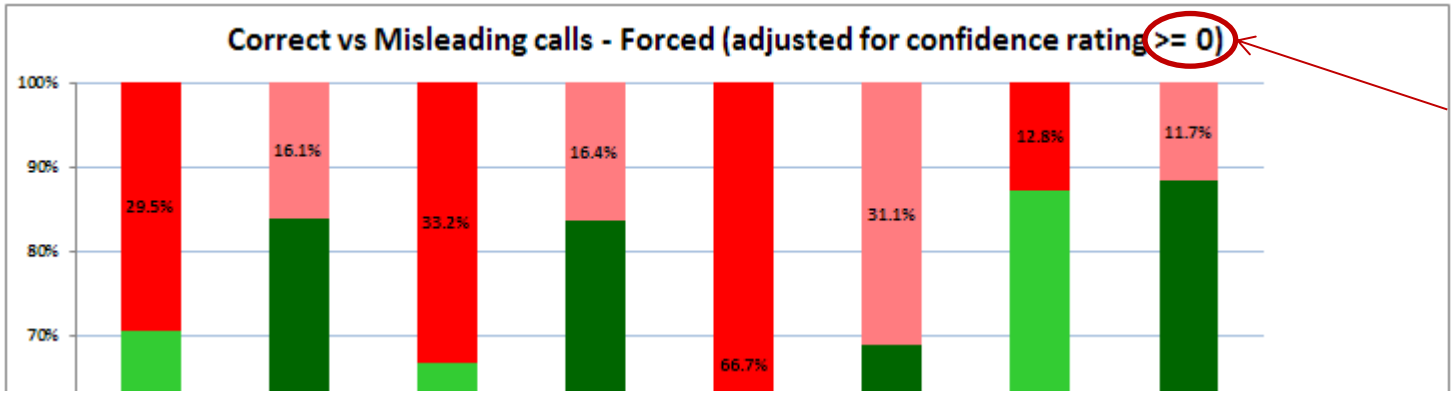
Inconclusive or No Authorship Opinion

- Inconclusive is **not** an ‘authorship’ call
  - Reflects ‘significant’ uncertainty in comparison process
  - Indicates a lack of confidence should any opinion be expressed
  - Perhaps best characterized as an assessment of quality

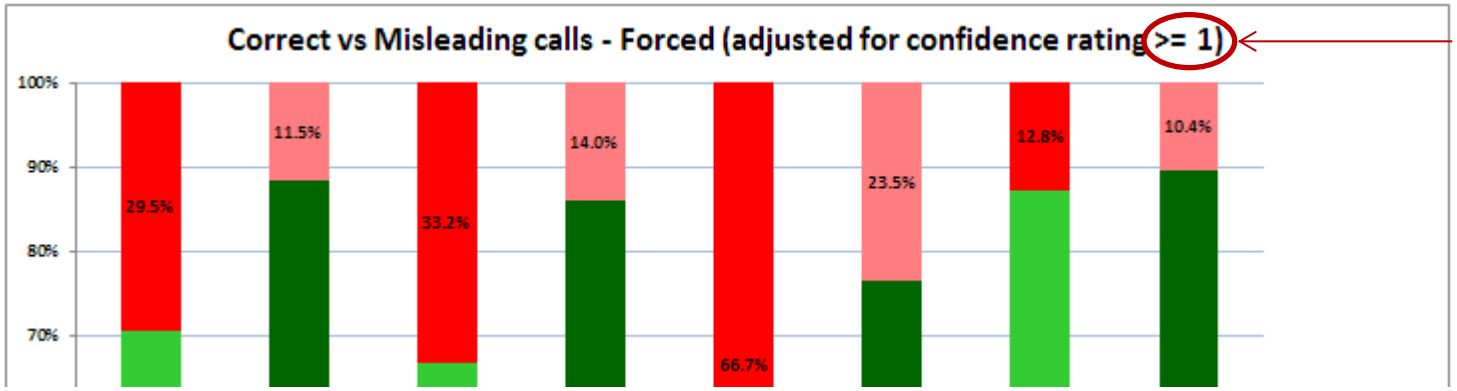
- Some questions arise:
  - What ‘confidence’ corresponds to inconclusive state?
  - If we ‘eliminate’ calls according to confidence, what is the effect?
  - Is the same effect seen for both FDEs and laypersons?

Correct vs Misleading calls - Forced (adjusted for confidence rating >= 0)

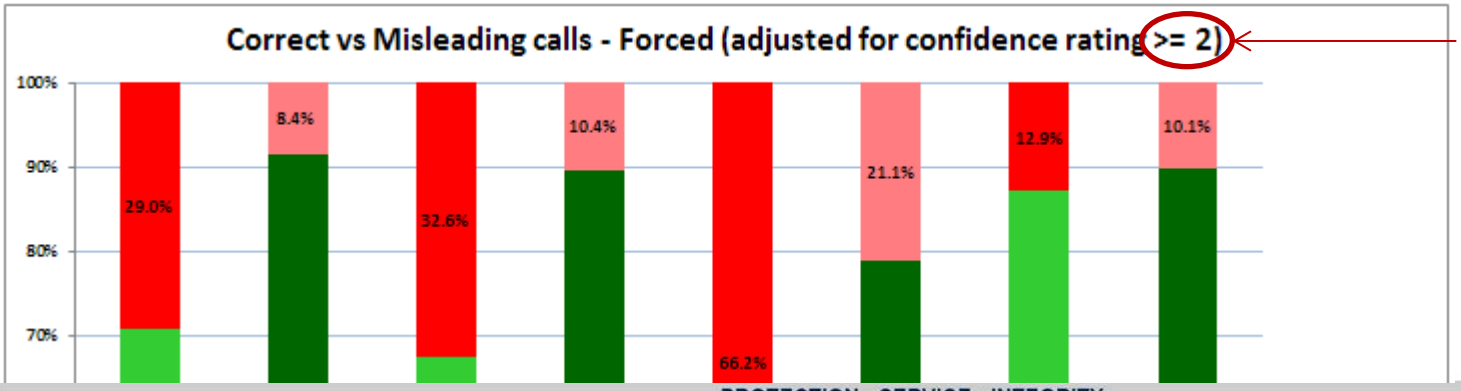




Includes all calls



Excludes 0 only



Excludes 0, 1

Effect of Confidence Rating adjustment vs ER

	OVERALL		Genuine Q		Disguised Q		Simulated Q	
	Lay	FDE	Lay	FDE	Lay	FDE	Lay	FDE
0 or >	29.5	16.1	33.2	16.4	66.7	31.1	12.8	11.7
1 or >	29.5	11.5	33.2	14.0	66.7	23.5	12.8	10.4
2 or >	29.0	8.4	32.6	10.4	66.2	21.1	12.9	10.1
3 or >	28.1	5.2	32.0	5.8	66.1	13.3	11.8	7.8
4 or >	27.3	4.9	30.8	5.4	70.9	7.7	11.3	1.5

Cell values = elicited ER when confidence ratings are ‘adjusted’ by removing calls as per column 1

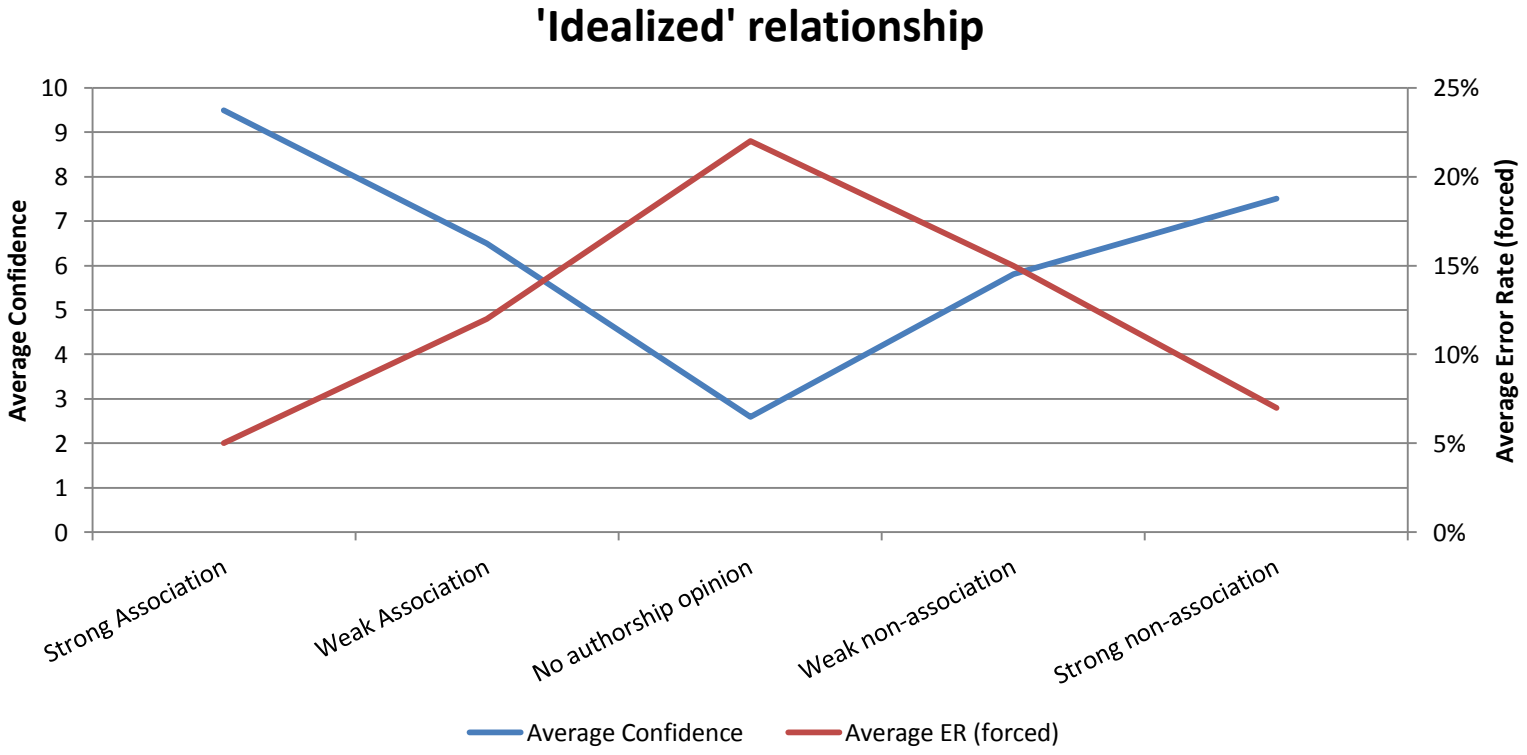
Forced Call vs 5-step scale

- Essentially, process of calibration
- Identify Q signatures that elicit a specific opinion
  - Using any desired scale (5, 7, 9, etc)
- Determine confidence rating and/or ER through forced call for those Q signatures
  - Complicated since any given Q may fall into different categories

- Expectation:
  - Inverse relationship between ER and Confidence seen earlier should be apparent using the normal scale

The 'Ideal' Relationship (Dummy Data)

- Conclusion scale is expression of examiner ‘confidence’
- Ideal relationship between ER and Confidence

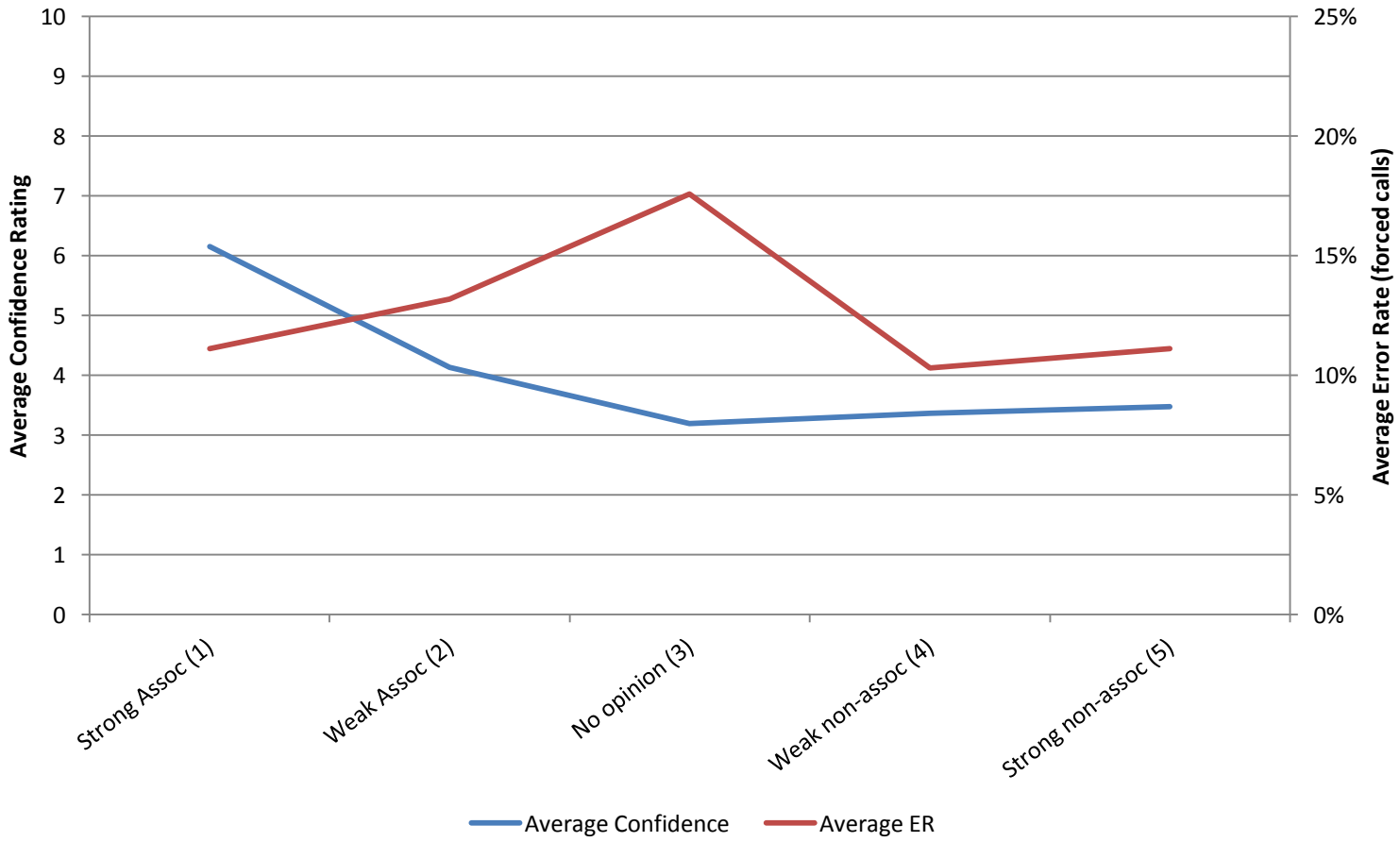


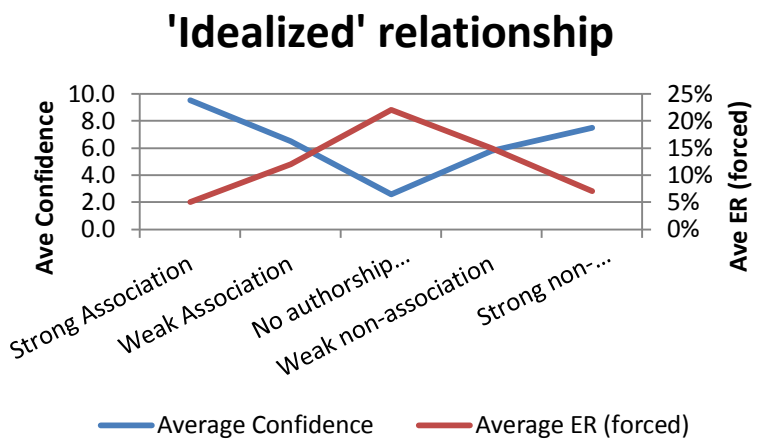
Calibrating the Conclusion Scale

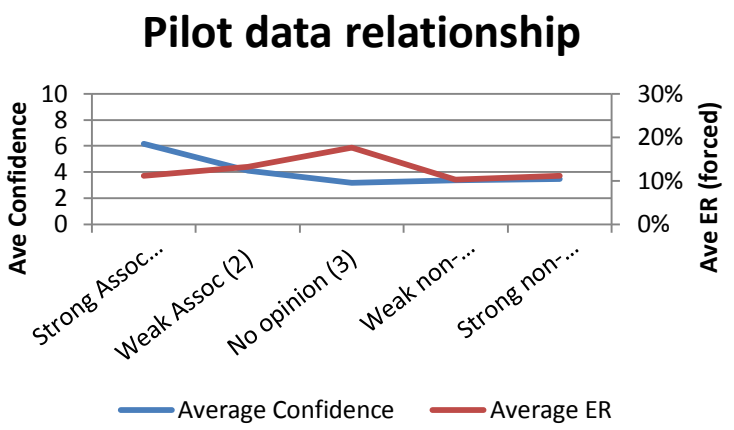
- Specifically 5-level scale used in S-03 (1-5)
- Quasi- ASTM

S-03 conclusions	# Q calls (total 240)	Weighted Confidence (forced calls)	Weighted Error Rate (forced calls)
Strong assoc. (1)	30	6.2	11%
Weak assoc. (2)	55	4.1	13%
Inconclusive (3)	134	3.2	18%
Weak non-assoc. (4)	18	3.4	10%
Strong non-assoc. (5)	3	3.5	11%

Conclusion Category vs Confidence/Error Rate (forced calls)



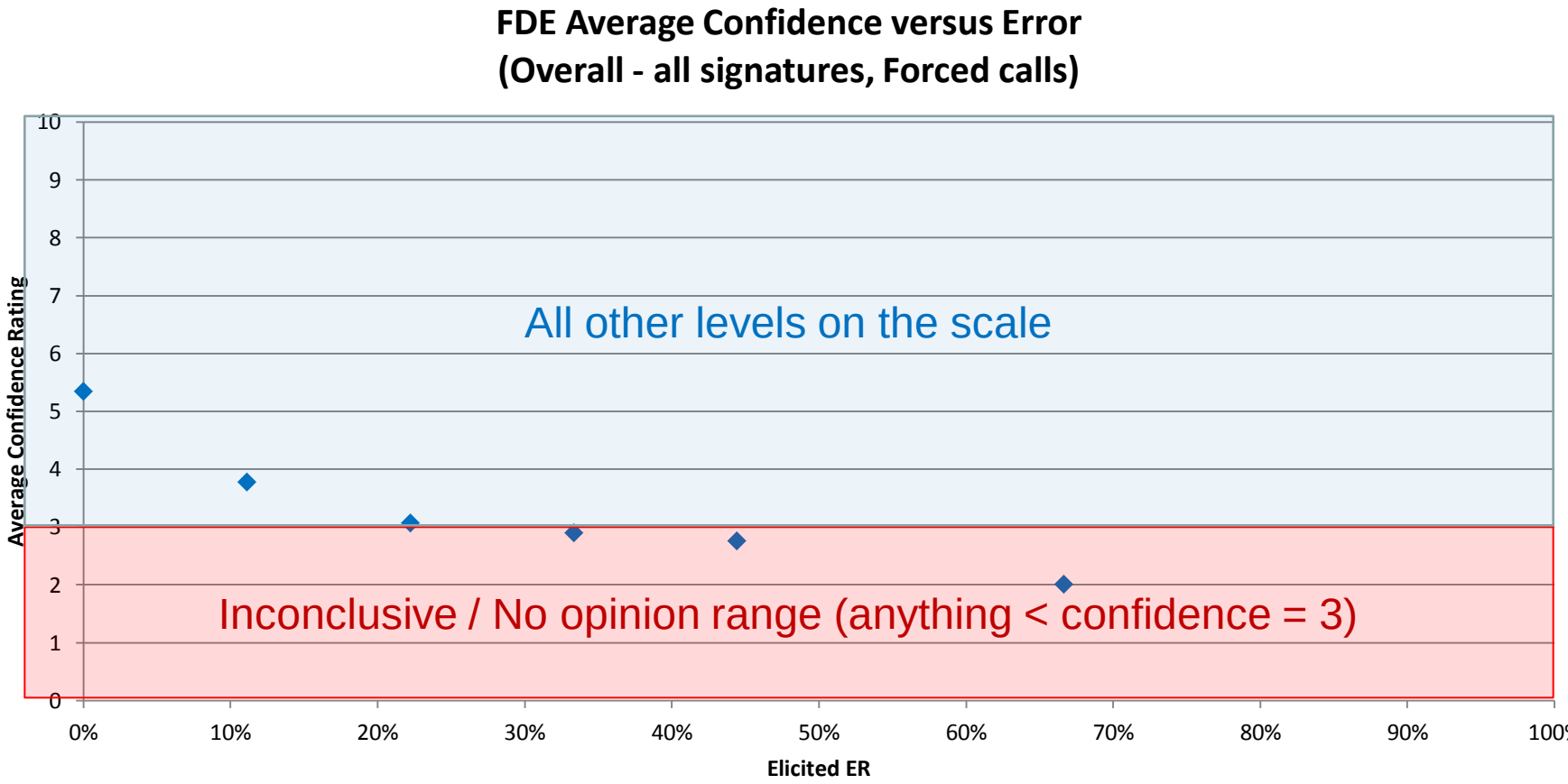




- Pilot data show the expected form but not perfect
  - Provides limited ‘validation’ of the scale
  - Potential issue might be use of CBSA-only data
- Shape may be due to restricted data set OR possibly representative of a ‘mis-calibration’ in the scale
  - More, and expanded, testing is required to sort out
- Valuable use of testing is QA and ‘re-calibration’

Steps on the Scale

- How many steps or levels are supported? 9?



The main St2ar S-03 Group

- CBSA group was used for above observations
- Can we use the main test group? 24 subjects (excl. CBSA)

- Unfortunately, some ‘issues’ observed – not yet resolved:
  - 8 FDEs expressed no inconclusive ‘3’ calls
  - 1 subject expressed all ‘1’ calls
  - 3 expressed either ‘1’ or ‘5’ (no 2, 3, or 4)
  - 2 expressed ‘1’, ‘3’ or ‘5’ (no 2 or 4)

- No clear pattern emerged for non-CBSA FDE group

Testing vs. the ‘Real World’

- Do test results like these help in the 'real world'?

- Generalization of results is very difficult
  - Group vs individual results – a **lot** of between-subject variability
  - No ‘control' over participants/subjects
  - Lack of random selection of subjects
  - Limitations in actual test design
    - % of disguised vs simulated signatures
    - Sample ‘variation’ (or lack thereof) – eg. single writer

Practical considerations

- Value of testing is not clear to some
- Interpretation of any derived or estimated ER is difficult
- Testing should be a QA/QC function first and foremost
  - In FDE, competency is a serious concern