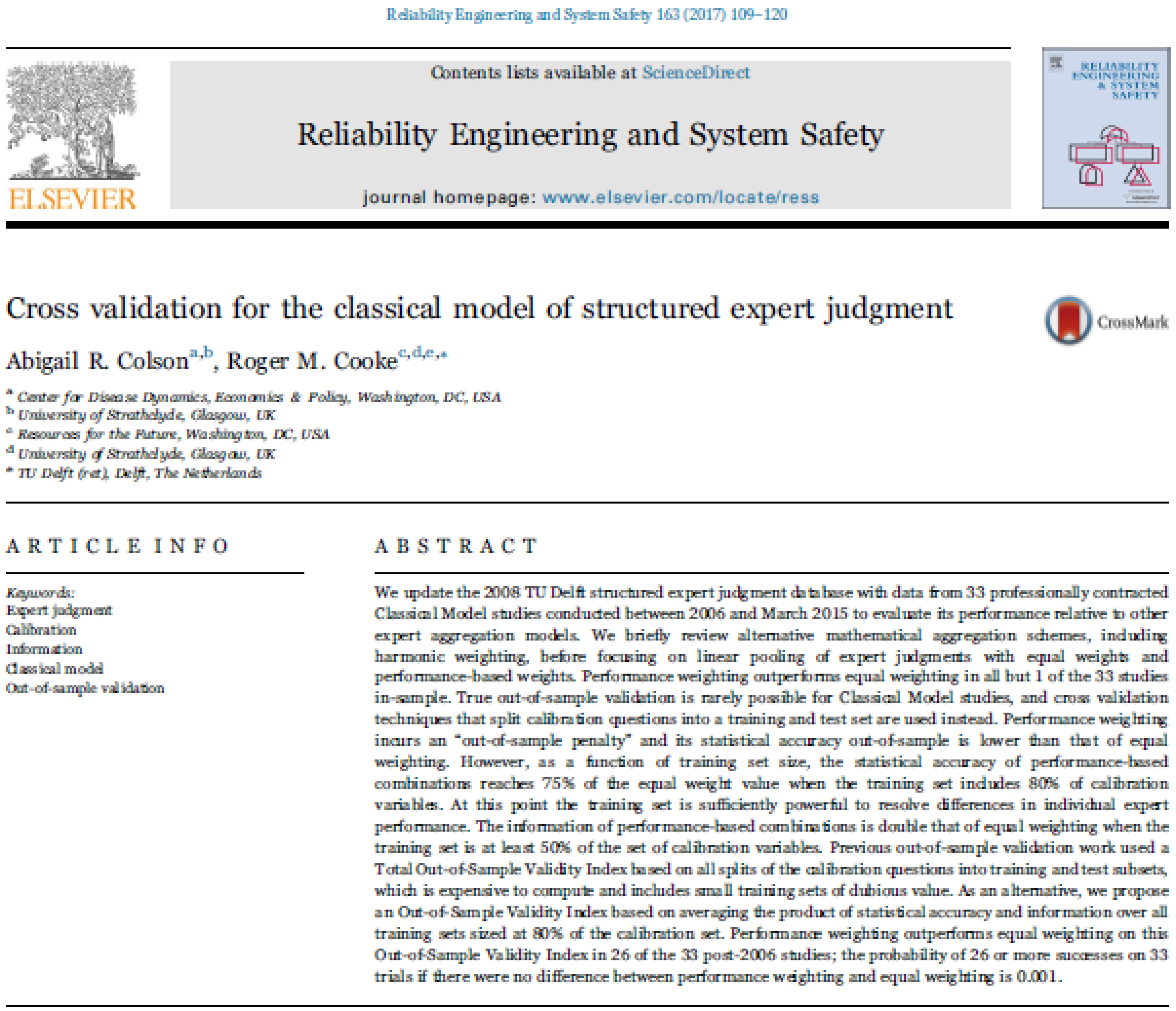
Validating expert judgments and the Classical Model

Abigail Colson, Department of Management Science
4 July, 2017
TU Delft COST Meeting

Roger Cooke, TU Delft, Resources for the Future

This talk comes from 2 papers.

- Colson, Abigail R., and Roger M. Cooke. 2017. 'Cross Validation for the Classical Model of Structured Expert Judgment'. *Reliability Engineering & System Safety* 163 (July): 109–20. doi:10.1016/j.ress.2017.02.003.
- Colson, Abigail R., and Roger M. Cooke. 2017. 'Validating Experts' Judgments with the Classical Model'. *Review of Environmental Economics and Policy*. Forthcoming.



Reliability Engineering and System Safety 163 (2017) 109–120

Contents lists available at ScienceDirect

Reliability Engineering and System Safety

journal homepage: www.elsevier.com/locate/ress

Cross validation for the classical model of structured expert judgment

Abigail R. Colson[a,b], Roger M. Cooke[c,d,e,*]

[a] Center for Disease Dynamics, Economics & Policy, Washington, DC, USA
[b] University of Strathclyde, Glasgow, UK
[c] Resources for the Future, Washington, DC, USA
[d] University of Strathclyde, Glasgow, UK
[e] TU Delft (ret), Delft, The Netherlands

ARTICLE INFO

Keywords:
Expert judgment
Calibration
Information
Classical model
Out-of-sample validation

ABSTRACT

We update the 2008 TU Delft structured expert judgment database with data from 33 professionally contracted Classical Model studies conducted between 2006 and March 2015 to evaluate its performance relative to other expert aggregation models. We briefly review alternative mathematical aggregation schemes, including harmonic weighting, before focusing on linear pooling of expert judgments with equal weights and performance-based weights. Performance weighting outperforms equal weighting in all but 1 of the 33 studies in-sample. True out-of-sample validation is rarely possible for Classical Model studies, and cross validation techniques that split calibration questions into a training and test set are used instead. Performance weighting incurs an "out-of-sample penalty" and its statistical accuracy out-of-sample is lower than that of equal weighting. However, as a function of training set size, the statistical accuracy of performance-based combinations reaches 75% of the equal weight value when the training set includes 80% of calibration variables. At this point the training set is sufficiently powerful to resolve differences in individual expert performance. The information of performance-based combinations is double that of equal weighting when the training set is at least 50% of the set of calibration variables. Previous out-of-sample validation work used a Total Out-of-Sample Validity Index based on all splits of the calibration questions into training and test subsets, which is expensive to compute and includes small training sets of dubious value. As an alternative, we propose an Out-of-Sample Validity Index based on averaging the product of statistical accuracy and information over all training sets sized at 80% of the calibration set. Performance weighting outperforms equal weighting on this Out-of-Sample Validity Index in 26 of the 33 post-2006 studies; the probability of 26 or more successes on 33 trials if there were no difference between performance weighting and equal weighting is 0.001.

What is “The Classical Model”?

- A method to combine and validate experts’ quantifications of uncertainty
- It’s NOT a method to coerce agreement between the experts
- The method has been used by WHO, EU, EPA, NOAA, NASA, etc.

- In the classical model, experts answer 2 types of questions:
    - Calibration (aka “seed”) questions
    - Variables of interest
- With calibration variables, any expert (or combination of experts) can be treated like a statistical hypothesis.
- Experts’ assessments are weighted according to performance and combined.

An example question

In the United States in 2012, how many of the 4,104 tested *E. coli* isolates included in data from The Surveillance Network (TSN) were resistant to fluoroquinolones?				
____________	____________	____________	____________	____________
5%	25%	50%	75%	95%

An example question

In the United States in 2012, how many of the 4,104 tested *E. coli* isolates included in data from The Surveillance Network (TSN) were resistant to fluoroquinolones?				
____410___	____615___	____820_____	___1435_____	____2460___
5%	25%	50%	75%	95%

An example question

In the United States in 2012, how many of the 4,104 tested *E. coli* isolates included in data from The Surveillance Network (TSN) were resistant to fluoroquinolones?					
____410___	____615___	____820_____	X	___1435_____	____2460___
5%	25%	50%		75%	95%

True value: 1,230

Measuring expert performance

Statistical accuracy:

- Do the expert’s assessments capture the true values at the expected frequency?
- P-value of a statistical test of the expert’s hypotheses

Informativeness:

- How concentrated is the assessment, relative to a background measure?
- The background measure normally uniform with a 10% overshoot range.

One unique feature of the CM: DATA!

ELSEVIER

Available online at www.sciencedirect.com
ScienceDirect

Reliability Engineering and System Safety 93 (2008) 657–674

RELIABILITY ENGINEERING & SYSTEM SAFETY

www.elsevier.com/locate/ress

TU Delft expert judgment data base

Roger M. Cooke[a,*], Louis L.H.J. Goossens[b]

[a] *Resources for the Future and Department of Mathematics, Delft University of Technology, Mekelweg 4, Delft, The Netherlands*
[b] *Department of Safety Science, Delft University of Technology, TU Delft, The Netherlands*

Available online 15 March 2007

Abstract

We review the applications of structured expert judgment uncertainty quantification using the "classical model" developed at the Delft University of Technology over the last 17 years [Cooke RM. Experts in uncertainty. Oxford: Oxford University Press; 1991; Expert judgment study on atmospheric dispersion and deposition. Report Faculty of Technical Mathematics and Informatics No.01-81, Delft University of Technology; 1991]. These involve 45 expert panels, performed under contract with problem owners who reviewed and approved the results. With a few exceptions, all these applications involved the use of seed variables; that is, variables from the experts' area of expertise for which the true values are available post hoc. Seed variables are used to (1) measure expert performance, (2) enable performance-based weighted combination of experts' distributions, and (3) evaluate and hopefully validate the resulting combination or "decision maker". This article reviews the classical model for structured expert judgment and the performance measures, reviews applications, comparing performance-based decision makers with "equal weight" decision makers, and collects some lessons learned.

Keywords: Expert judgment; Rational consensus; Calibration; Information; Subjective probability

One unique feature of the CM: DATA!

www.rogermcooke.net

Available online at www.sciencedirect.

ScienceDirect

ELSEVIER

Reliability Engineering and System Safety 93 (200

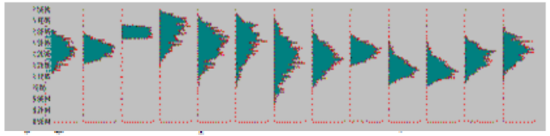
TU Delft expert judgmen

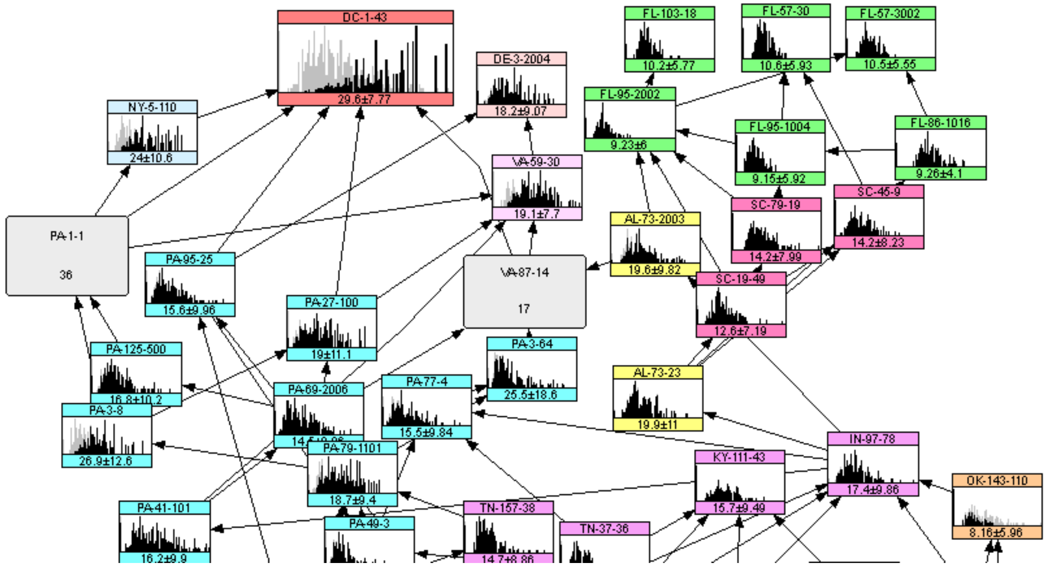
Roger M. Cooke[a,*], Louis L.H

[a]Resources for the Future and Department of Mathematics, Delft University of
[b]Department of Safety Science, Delft University of Technolo

Available online 15 March 200

Welcome to the Personal Website of Roger M Cooke (Mathematics and Philosophy)





Abstract

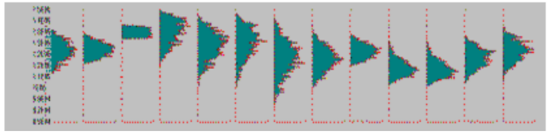
We review the applications of structured expert judgment uncertainty quantific
University of Technology over the last 17 years [Cooke RM. Experts in uncert
judgment study on atmospheric dispersion and deposition. Report Faculty of T
University of Technology; 1991]. These involve 45 expert panels, performed ur
approved the results. With a few exceptions, all these applications involved the u
area of expertise for which the true values are available post hoc. Seed variables
performance-based weighted combination of experts' distributions, and (3) evalu
"decision maker". This article reviews the classical model for structured expe
applications, comparing performance-based decision makers with "equal weight" decision makers, and collects some lessons learned.

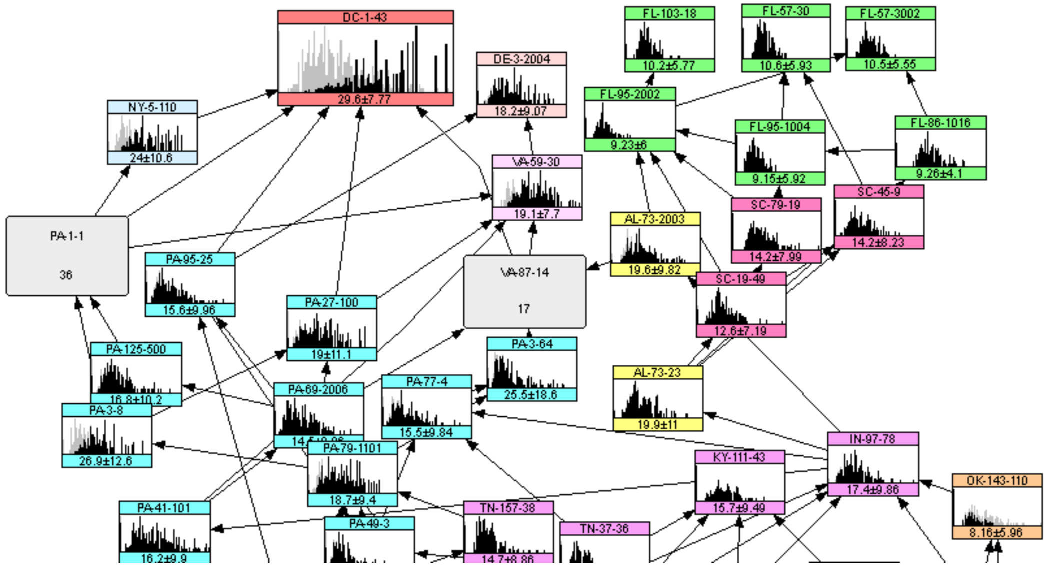
Keywords: Expert judgment; Rational consensus; Calibration; Information; Subjective probability

One unique feature of the CM: DATA!

www.rogermcooke.net

Welcome to the Personal Website of Roger M Cooke (Mathematics and Philosophy)





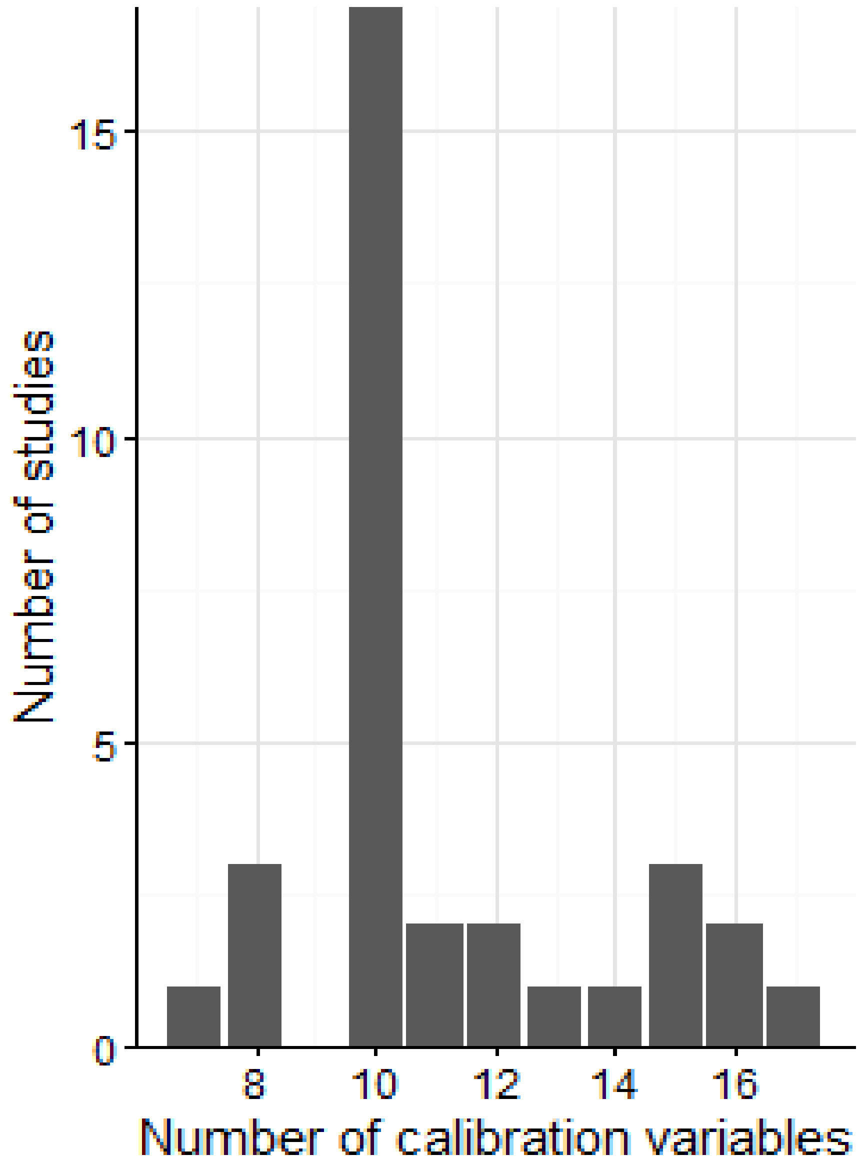
DATA from Structured Expert Judgment

53 SEJ studies pre 2009

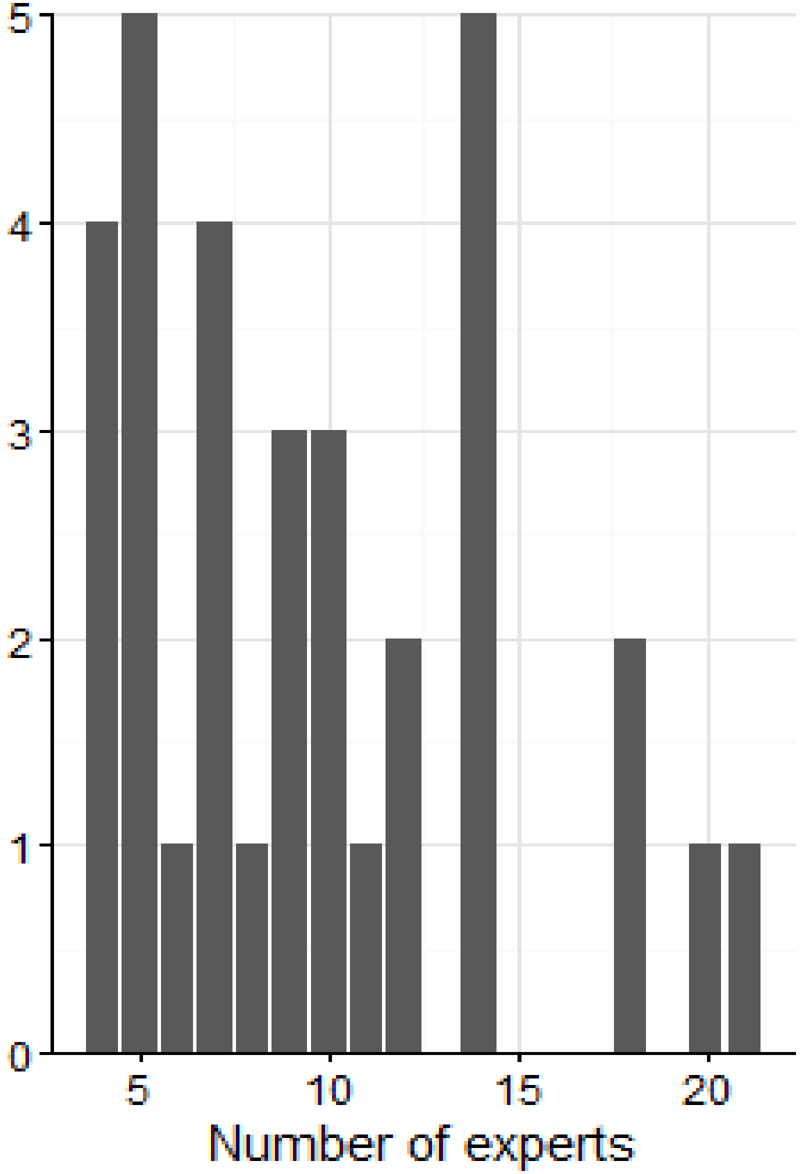
33 SEJ studies post 2006

Curriculum vitae Roger M. Cooke

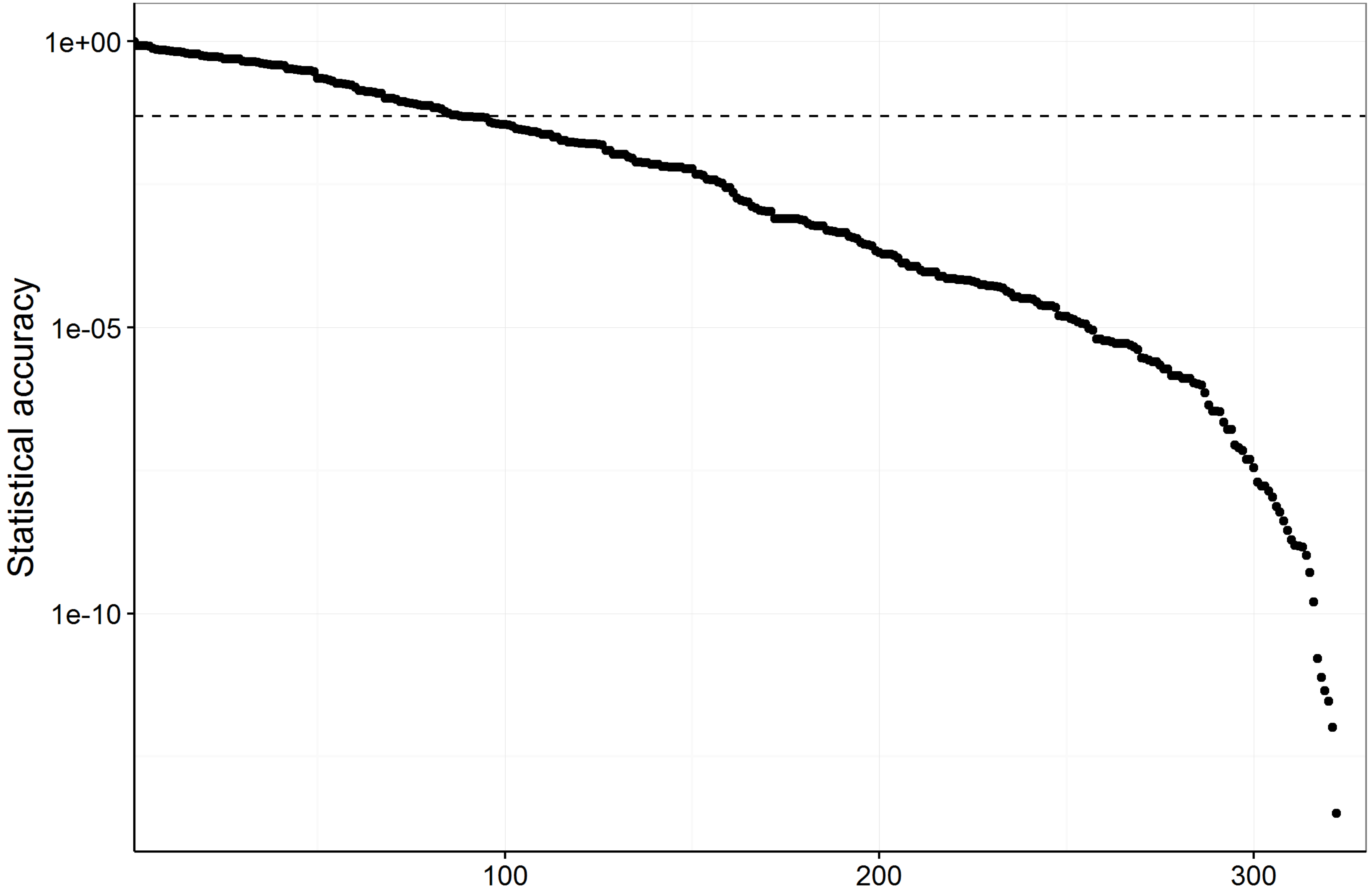
2/3 of the studies have 10 calibration ?s



Most studies have 5-14 experts



Statistical accuracy of 322 experts



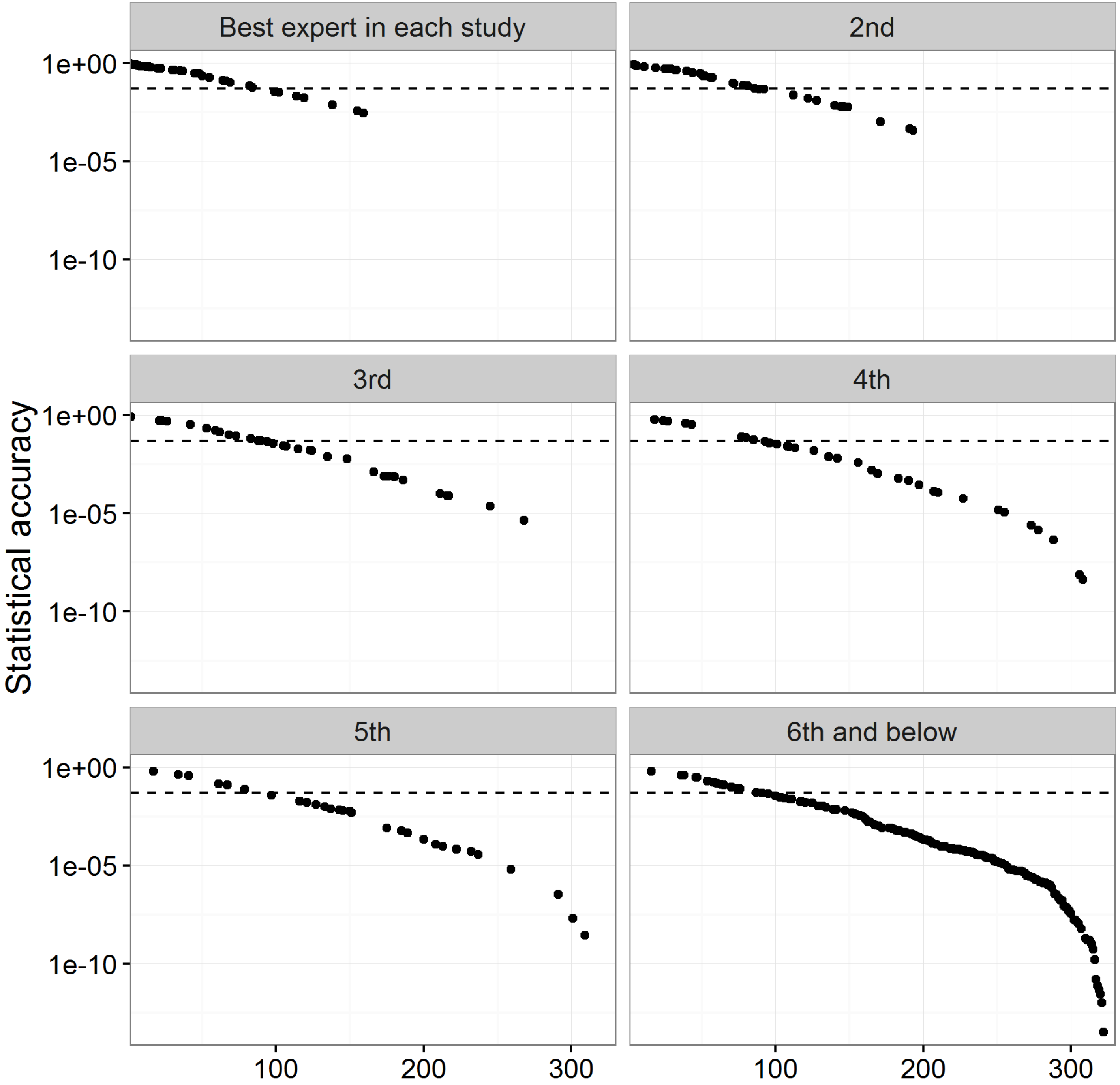
89 have SA > 0.05

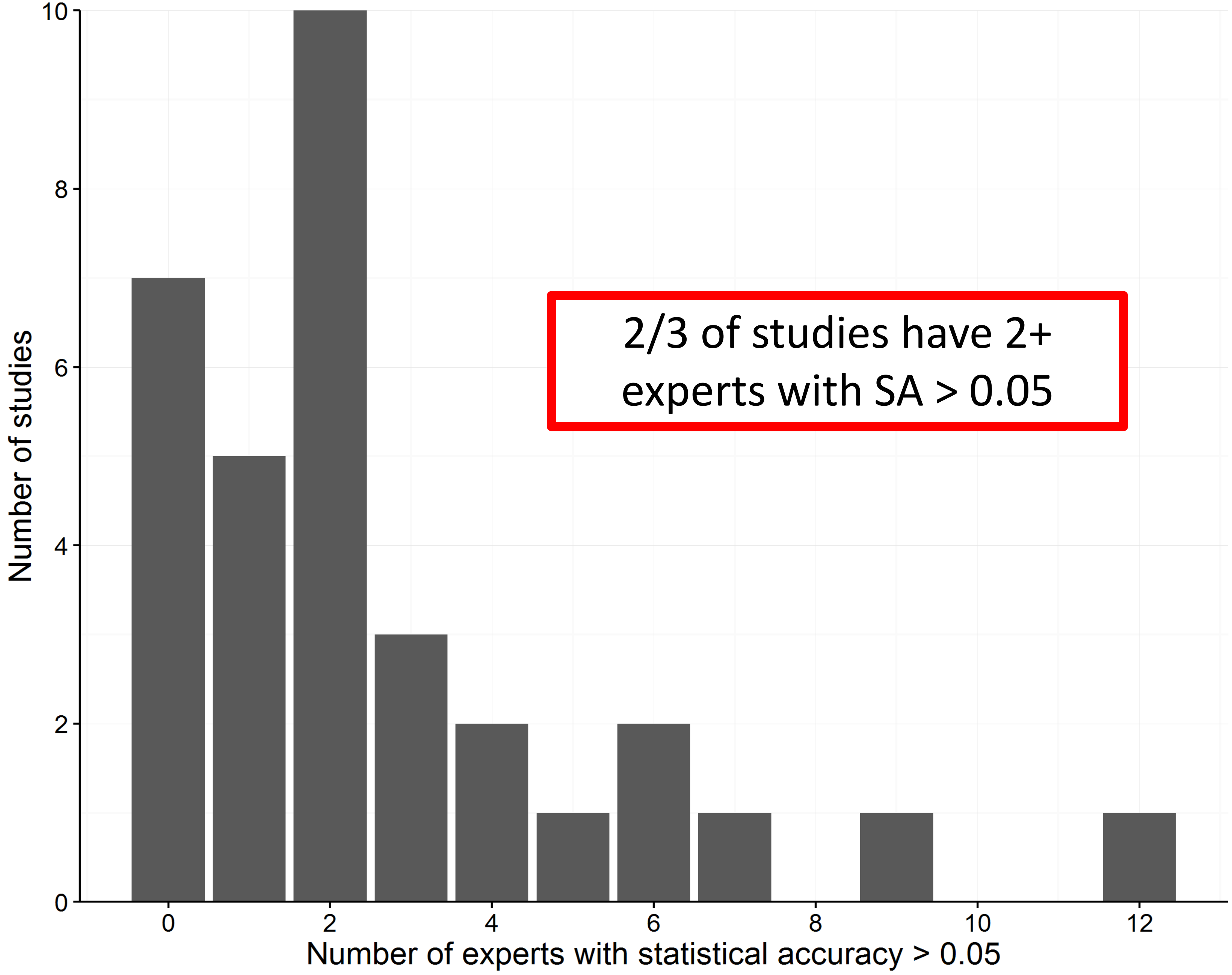
Over ½ have SA < 0.005

Approx ⅓ have SA < 0.0001

But it gets better!

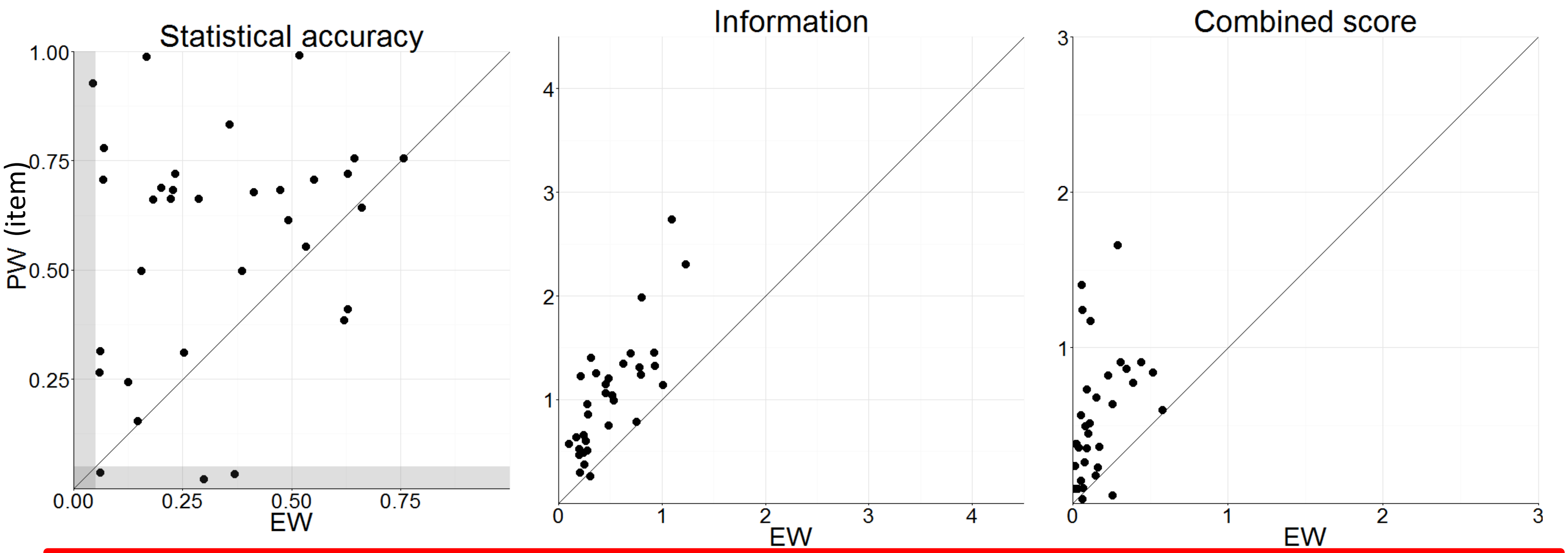
Statistical accuracy of the best experts looks less dismal.





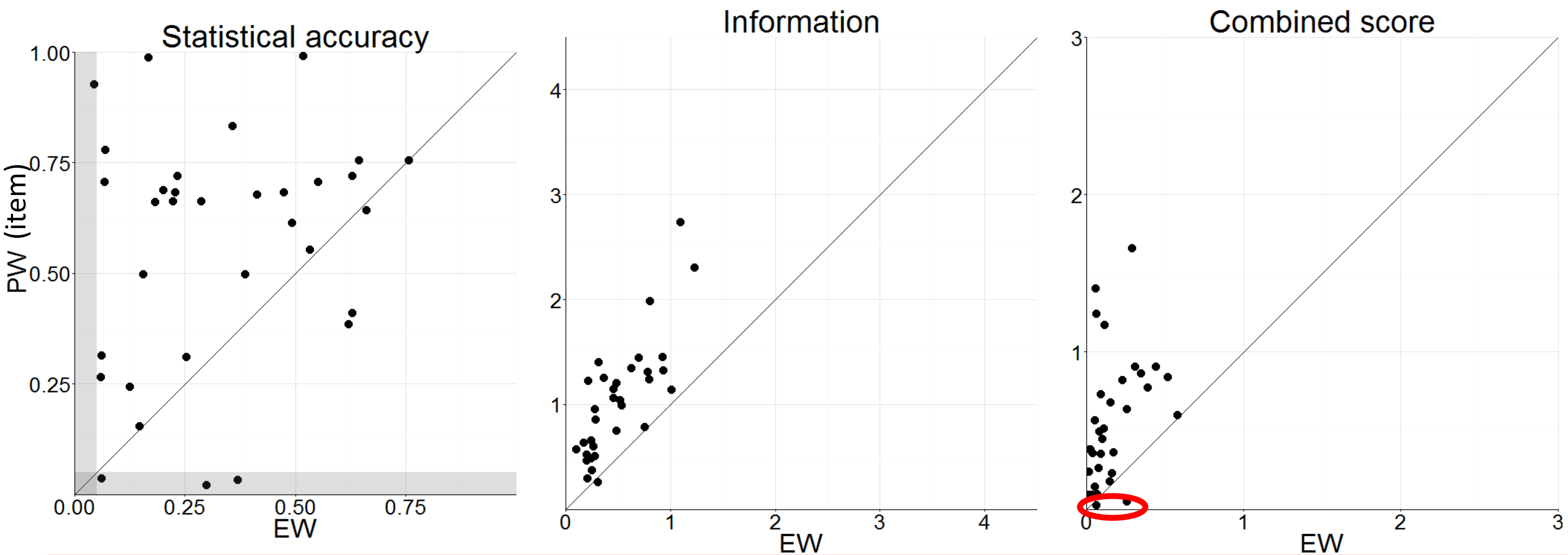
2/3 of studies have 2+
experts with SA > 0.05
Number of studies
Number of experts with statistical accuracy > 0.05
0
2
4
6
8
10
12

The benefit of performance weighting



Looking at combined scores:
PWi > PWg 14 studies
PWi = PWg in 13 studies
PWi = best expert 12 studies

The benefit of performance weighting



Looking at combined scores:

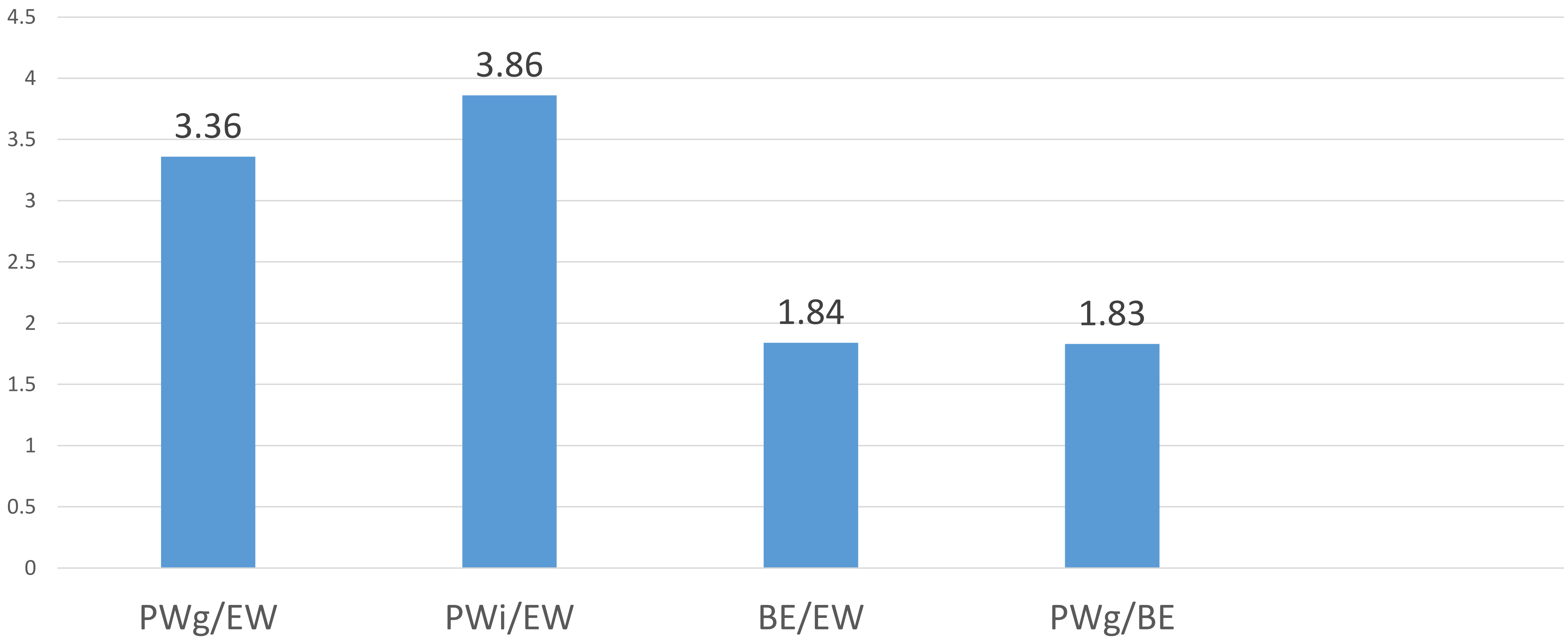
PWi > PWg 14 studies

PWi = PWg in 13 studies

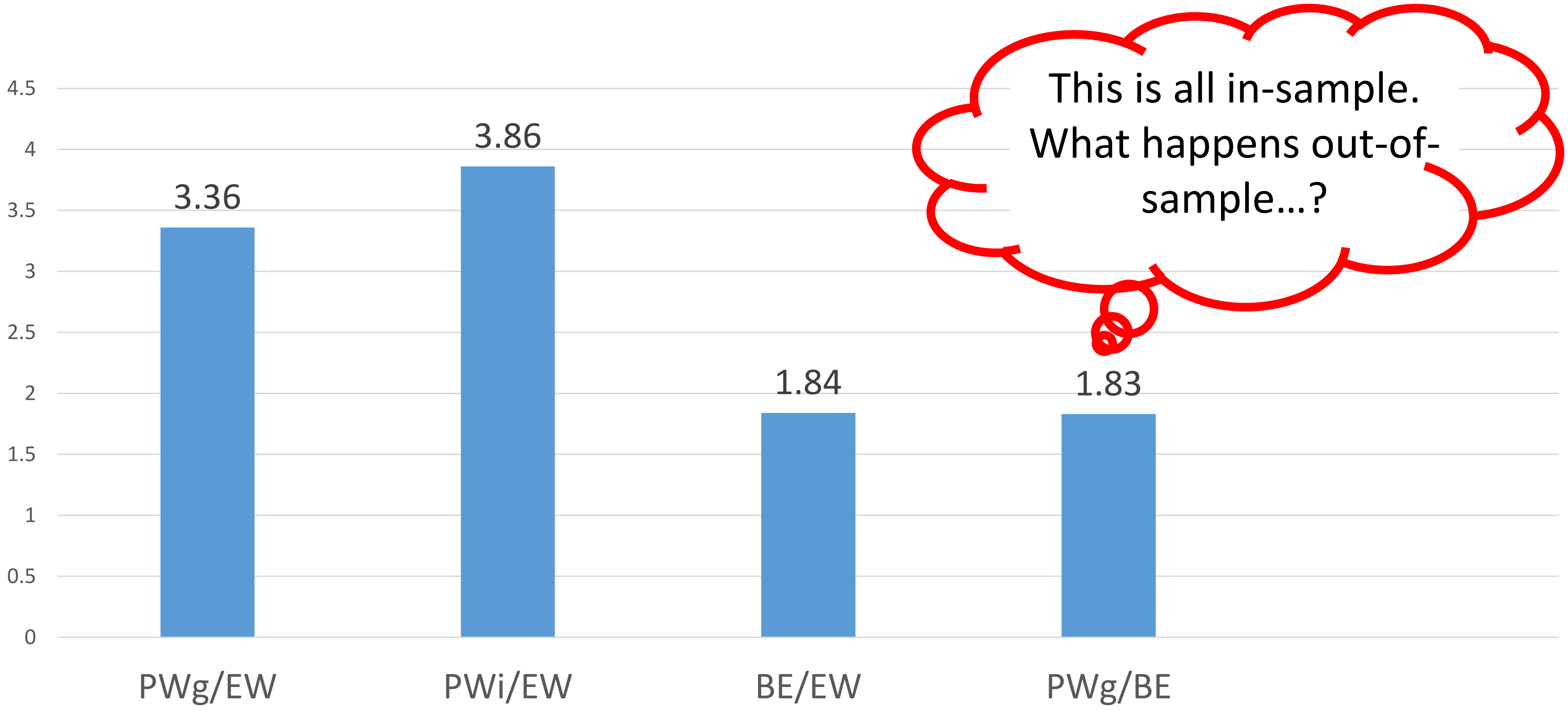
PWi = best expert 12 studies

EW is the best in 1 study

Geometric mean of ratios over all studies



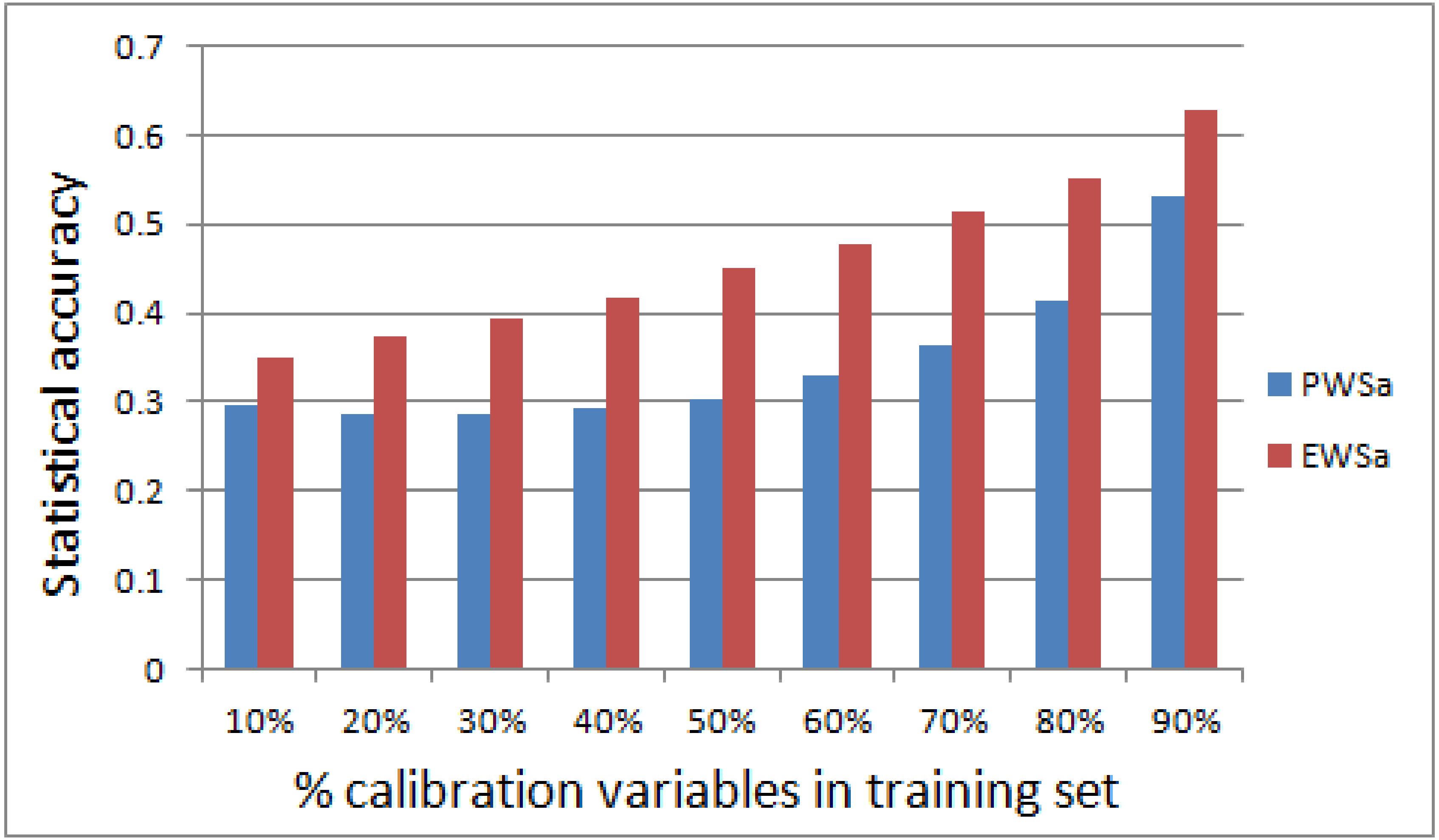
Geometric mean of ratios over all studies



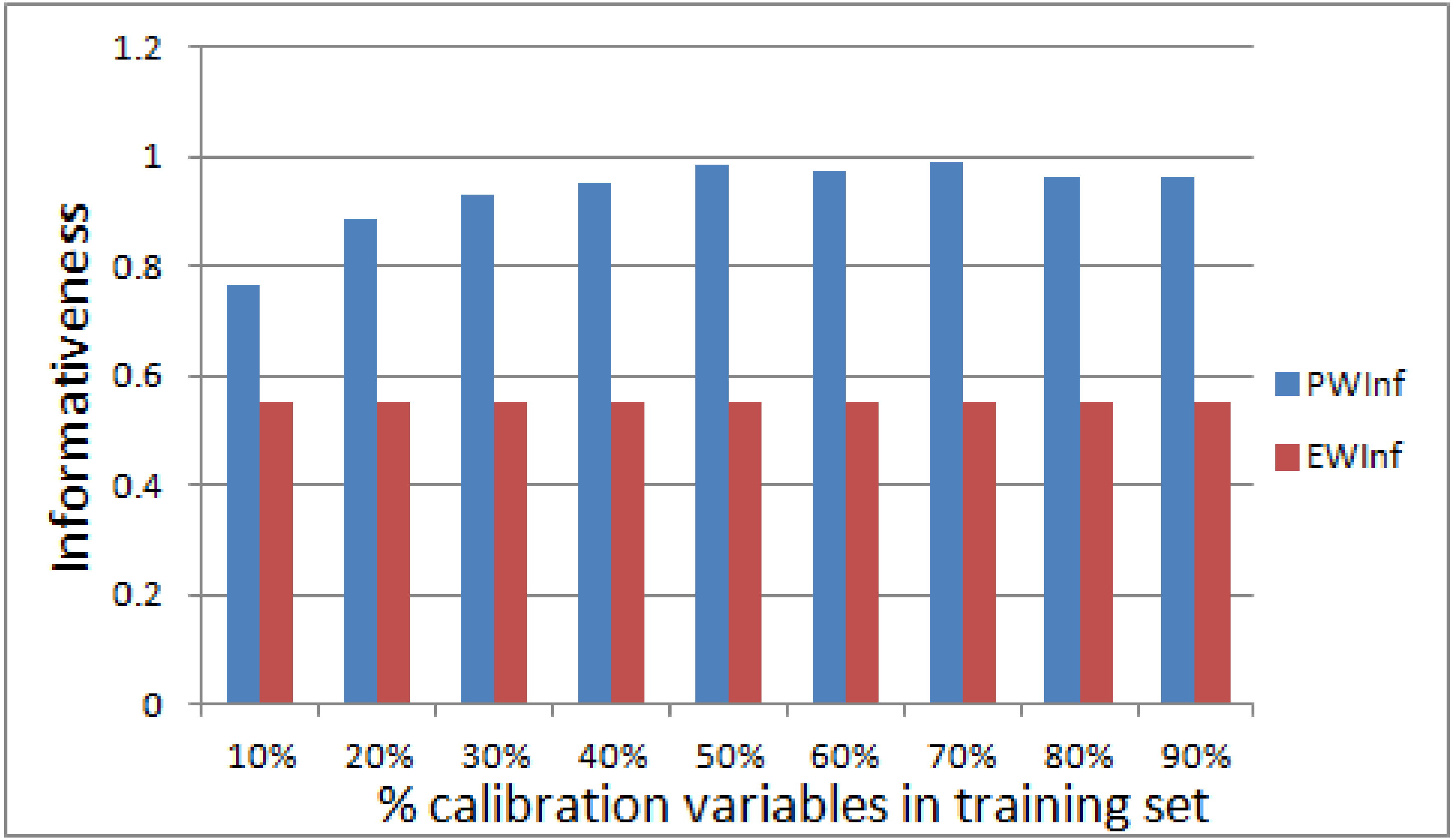
What about out-of-sample performance?

- True out-of-sample validation is rarely possible.
- Alternative methods
  - ROAT (Clemen 2008, Cooke 2008, Lin and Cheng 2008, Lin and Cheng 2009 Cooke 2011)
  - 50/50 splits (Cooke 2008)
  - Sampling 70/30 splits; test set at least 8 (Flandoli 2011)
  - Looking at all possible training/test splits (Eggstaff 2014)

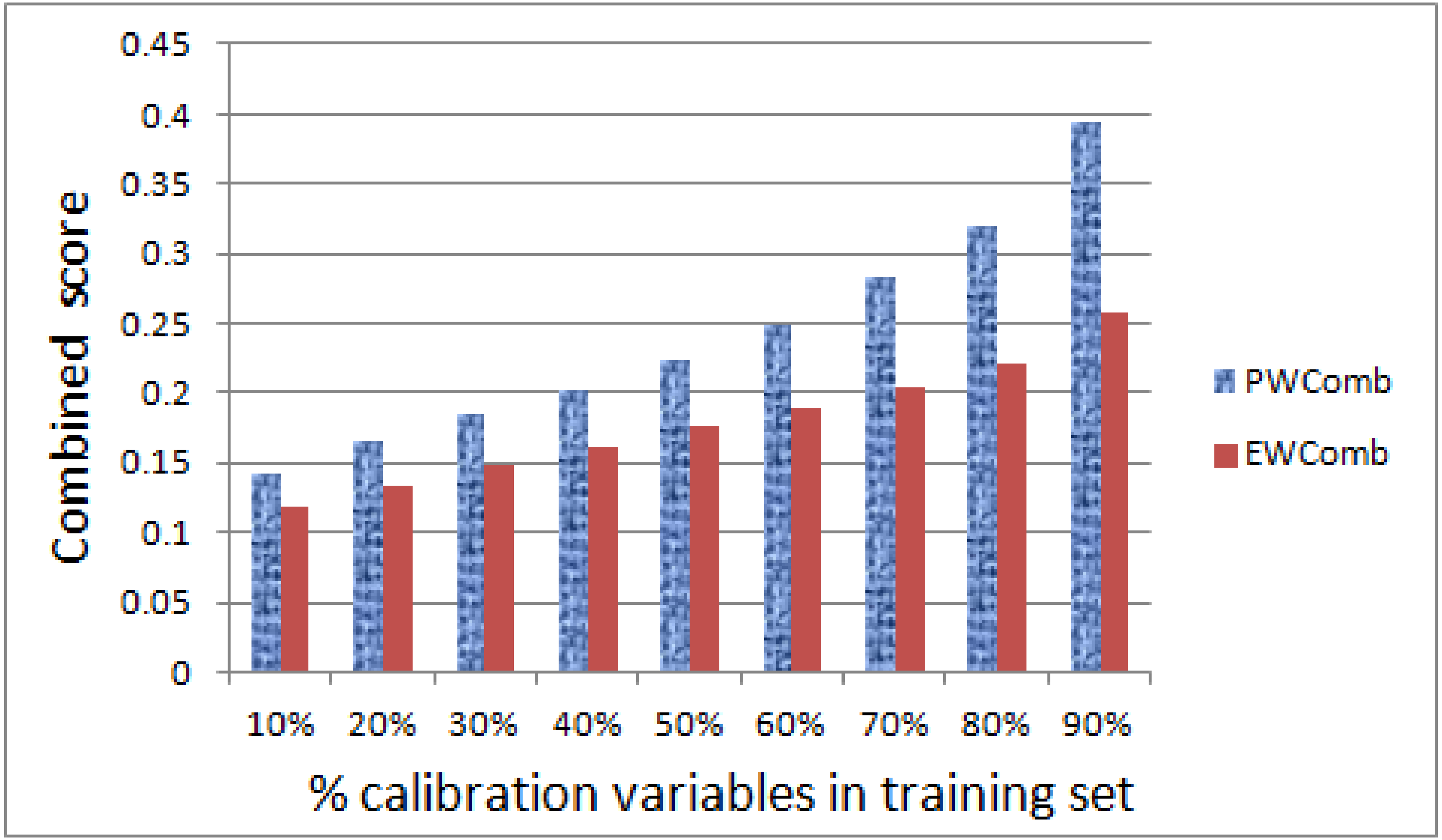
PWSa and EWSa by % training set, averaged over all studies



PWInf and EWInf by % training set, averaged over all studies



PWComb and EWComb by % training set, averaged over all studies



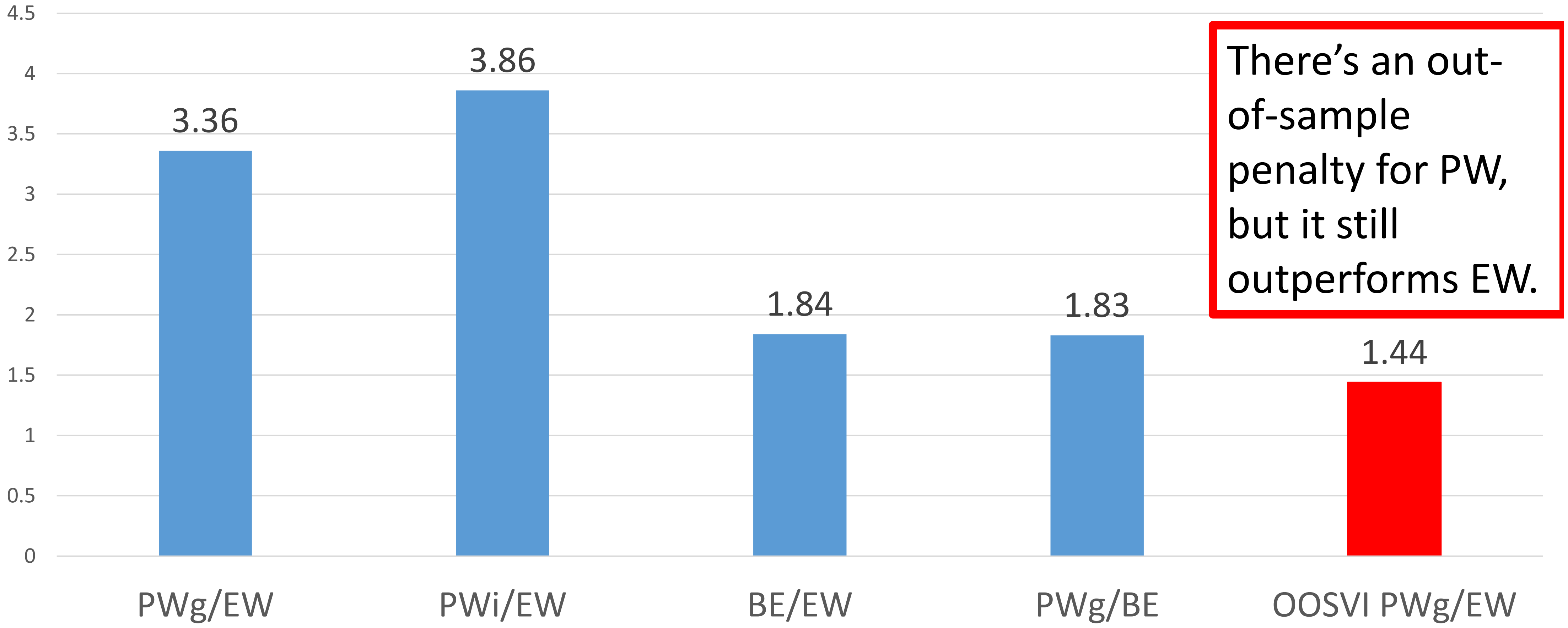
Choosing a summary measure is a tricky balance.

Image by Raimond Spekking / CC BY-SA 4.0 (via Wikimedia Commons)

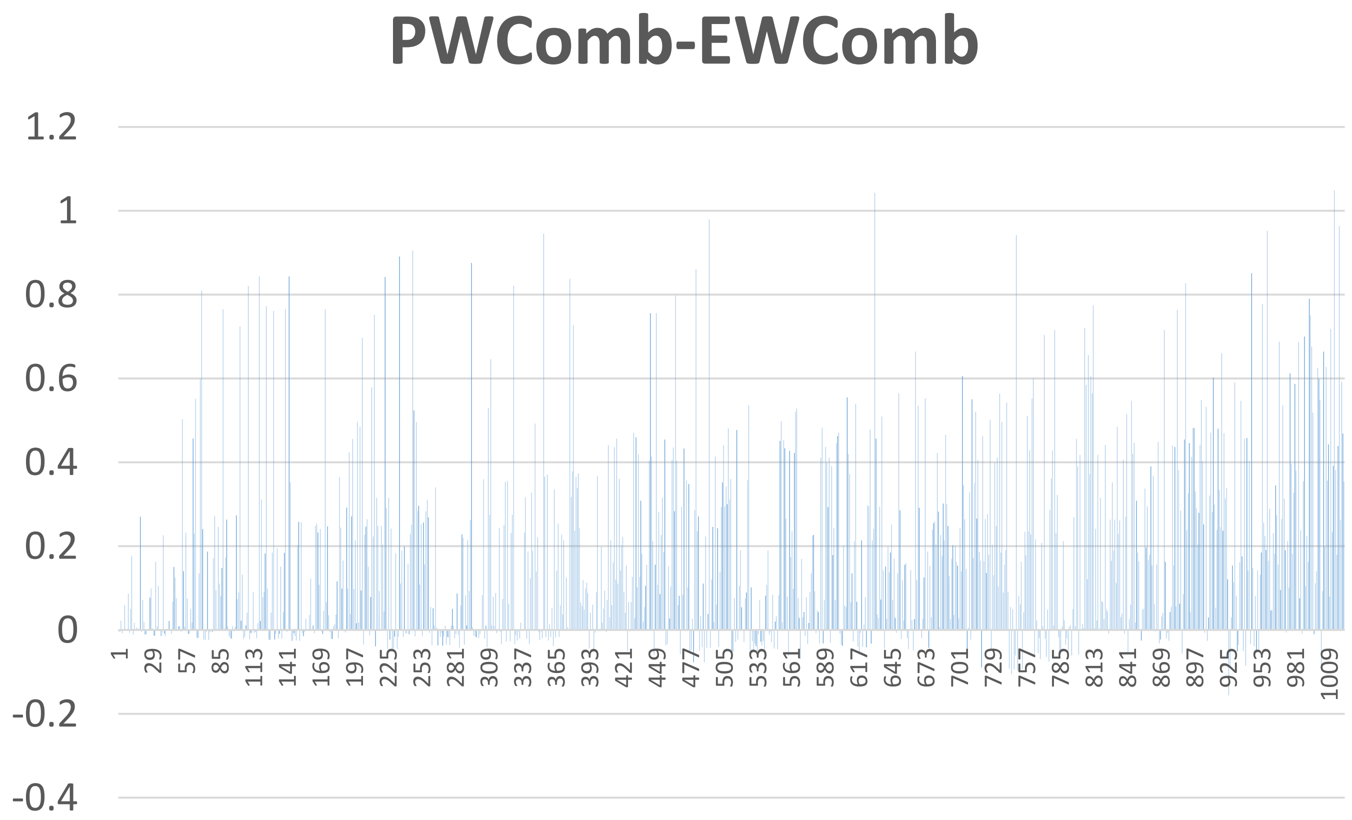
Out of Sample Validity Index: use training sets that are 80% of the entire set of calibration variables.

- The expert weights have low volatility.
- The expert weights more closely resemble the weights used in the actual study based on all calibration variables.
- For studies assessing 5-, 50- and 95-percentiles on 10 calibration variables, the possible statistical accuracy scores range over a factor 31, which is ample for distinguishing EW and PW.

Geometric mean of ratios over all studies

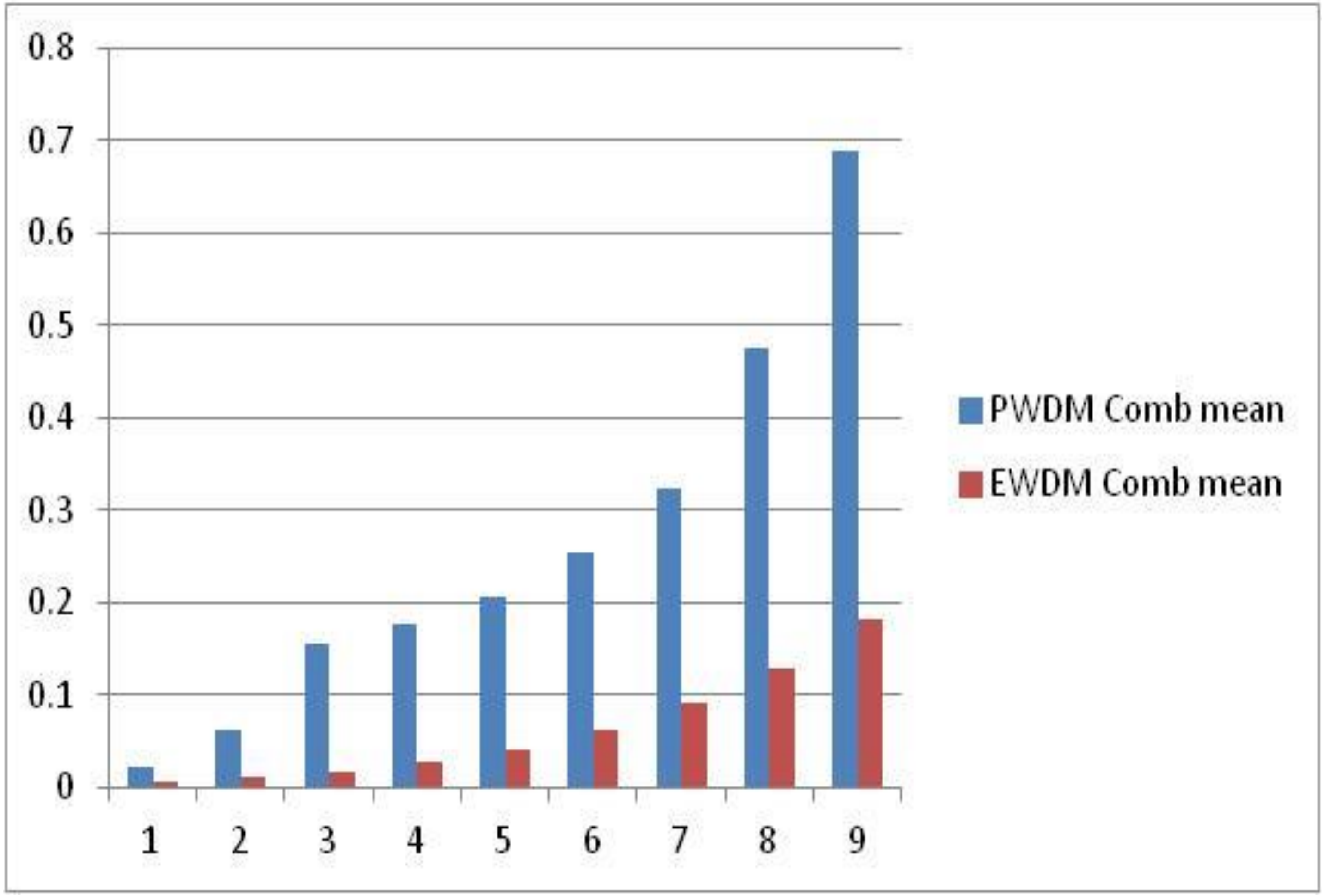


Example study: UK AMR

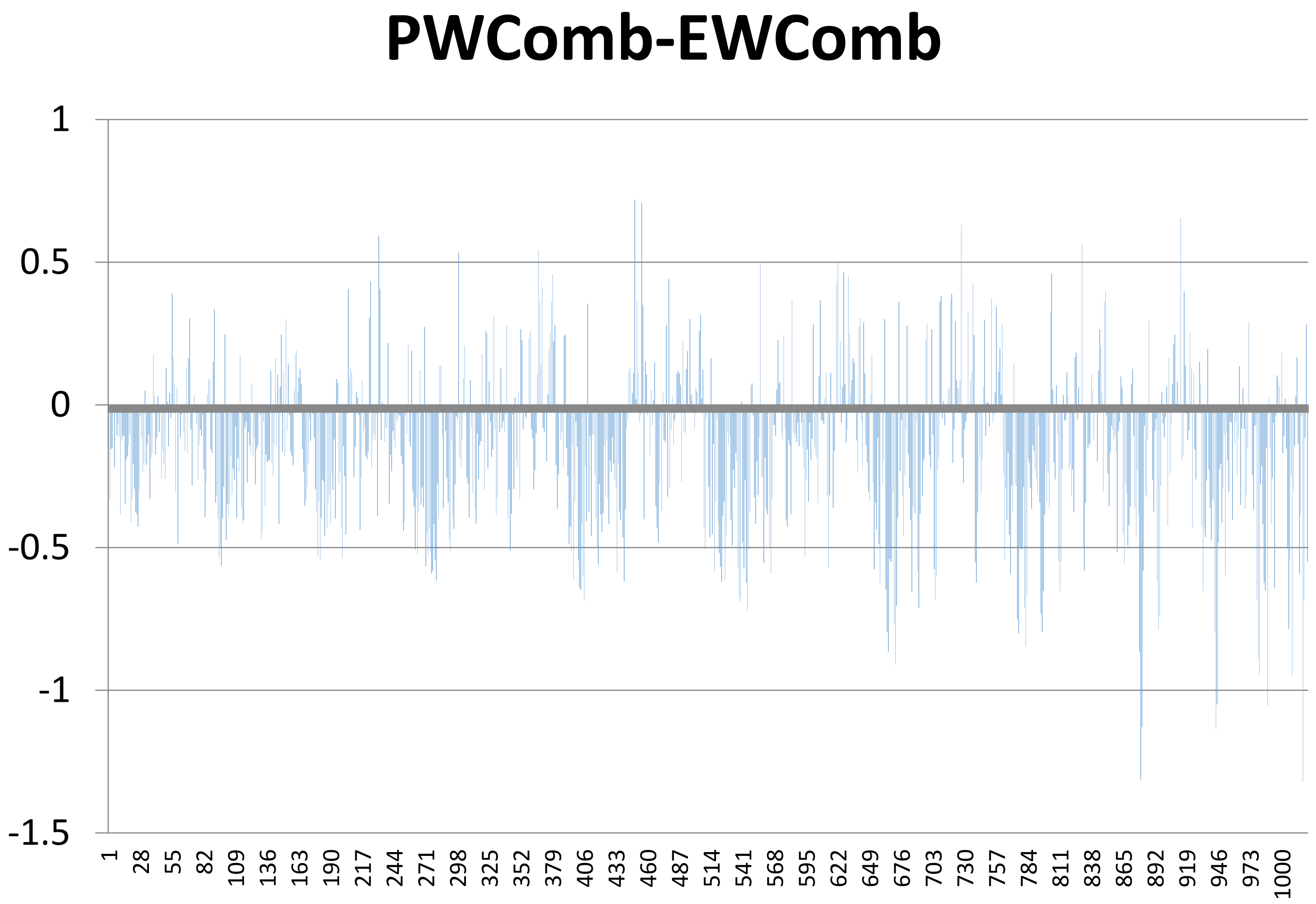


Example study: UK AMR

OOSVI =3.286

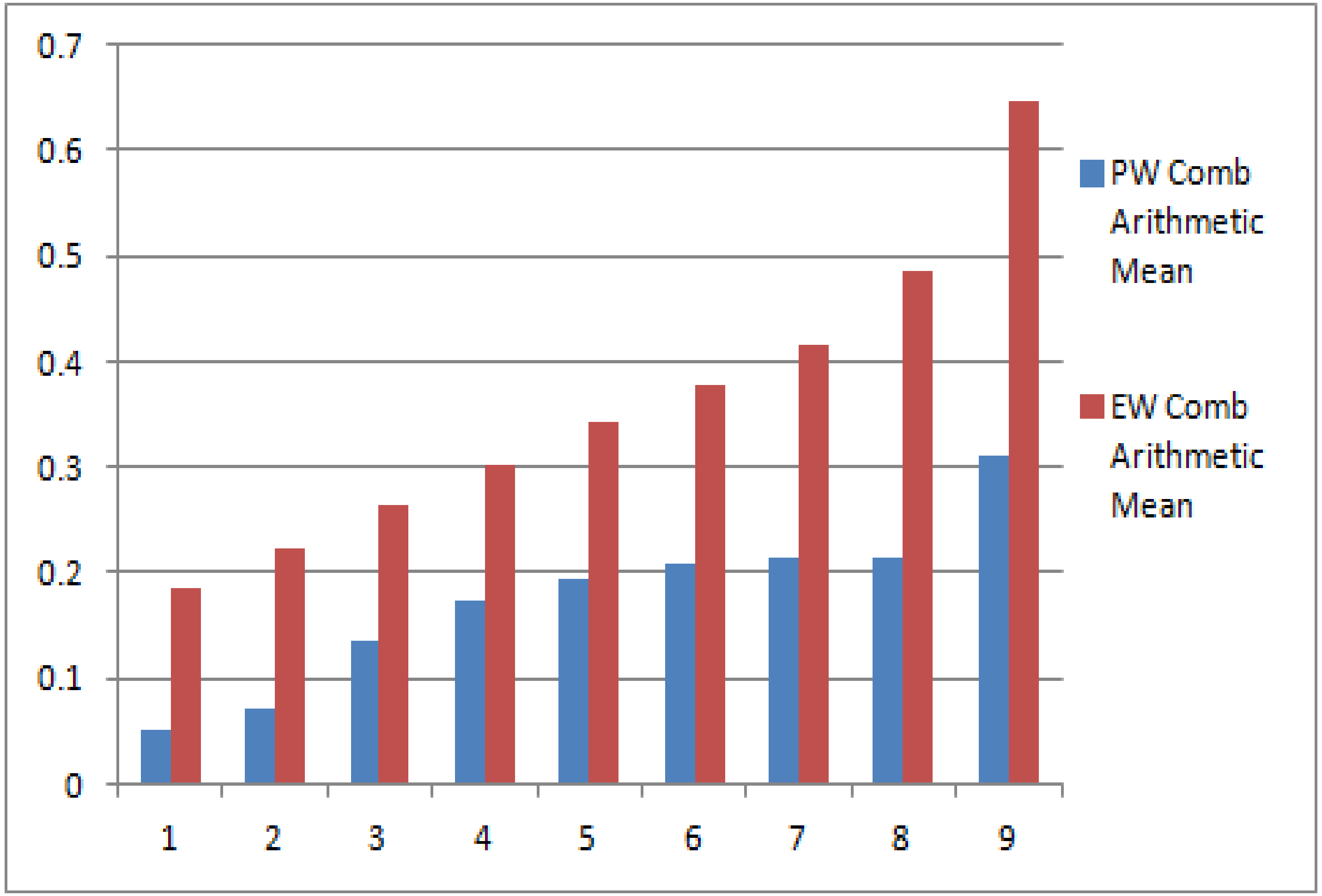


Example study: San Diego



Example study: San Diego

OOSVI = 0.439



How can we improve OOSVI?

- Number of experts?
- Number of calibration variables?
- 3 vs 5 quantiles?
- Plenary vs. 1-on-1?

How can we improve OOSVI?

- Number of experts? **No**
- Number of calibration variables? **No**
- 3 vs 5 quantiles? **No**
- Plenary vs. 1-on-1? **No**

	BE SA < 0.05	BE SA > 0.05
OOSVI	1.14	1.54

	SBE SA < 0.05	SBE SA > 0.05
OOSVI	1.17	1.64

Good OOSVI depends on good experts

What comes next?

- We need OOSV with item weights.
- Surely there’s something to say about study covariates and in/out-of-sample performance…
- The “updated” dataset is already woefully out of date.

Thanks!