Progress with Out of Sample Validation

Roger M. Cooke, Abby Colson
RFF, U of Strathclyde
Aug 28, 2015

Classical Model (1991)

- In addition to vbls of Interest, Experts give quantiles on seed vbls **FROM THEIR FIELD** [~10] whose values are known post hoc
- Experts scored wrt calibration (statistical accuracy) and informativeness
- Performance based weights (PW) (asymptotic strictly proper scoring rule) compared with Equal Weights (EW)
- Data sets from ~ 100 studies available

Contracting Party	Performed by	Study name	Subject	Reference nr
Robert Wood Johnson Foundation	Center for Disease Dynamics, Economics & Policy	Arkansas	Grant effectiveness, child health insurance enrollment	9.1
		Florida		
		Illinois		
		Nebraska		
		Washington		
		CoveringKids		
		Tobacco	Grant effectiveness tobacco control	9.2
		Obesity	Grant effectiveness, childhood obesity	
Disease Control Priority Project, 3rd Edition		Fistula	Effectiveness of obstetric fistula repair	3.29
		San_Diego	Effectiveness of surgical procedures	
Center for Disease Control and Prevention		CDC_ROI	Return on investment for CDC warnings	9.3
U Wisconsin, CREATE	V. Bier	Create	Terrorism	3.25, 3,26, 3,27
NOAA, EPA	U Notre Dame	Erie_Carp	Establishment of Asian Carp in Lake Erie	3.9, 3.22,
		GL_NIS	Costs of invasive species in Great Lakes	
EPA, U Maryland	B. Koch. U Maryland	UMD_NREMOVAL	Nitrogen removal in Chesapeake Bay	3.28
SANGUIN, UMC Utrecht	M. Janssen, UMC	Hemopilia	Hemophelia	4.7, 4.8, 4.9, 4.10
National Institute for Public Health and the	TU Delft	ATCEP	Air traffic Controllers Human Error	
		FCEP	Flight Crew Human Error	7.5
Brand Preventie		Daniela	Fire prevention and control	4.11
Liander		Liander	Underground cast iron gas-lines	
	W. Aspinall	Arsenic	Air quality levels for arsenic	--
		Biol_Agent	Human dose-response curves for bioterror agents	--
		CWD	Infection transmission risks: Chronic Wasting Disease from deer to humans	3.19; 3.20; 3.21
		eBPP	XMRV blood/tissue infection transmission risks	3.18; 3.24
		IceSheets	Contribution to sea level rise from Ice Sheets melting due to global warming	7.1
		PHAC_T4	Additional CWD factors	3.21
		Sheep	Risk management policy for sheep scab control	--
		SPEED	Volcano hazards (Vesuvius & Campi Flegrei, Italy)	2.6; 2.7
		TdC	Volcano hazards (Tristan da Cunha)	
		TOPAZ	Tectonic hazards for radwaste siting in Japan	6.16
New Zealand Earthquake Forecast Testing Centre	M. Gerstenberger	Gerstenberger	Seismic hazard modeling Christchurch	7.2,7.3, 7.4
Embry-Riddle	Goodheart	Goodheart	Airport safety	

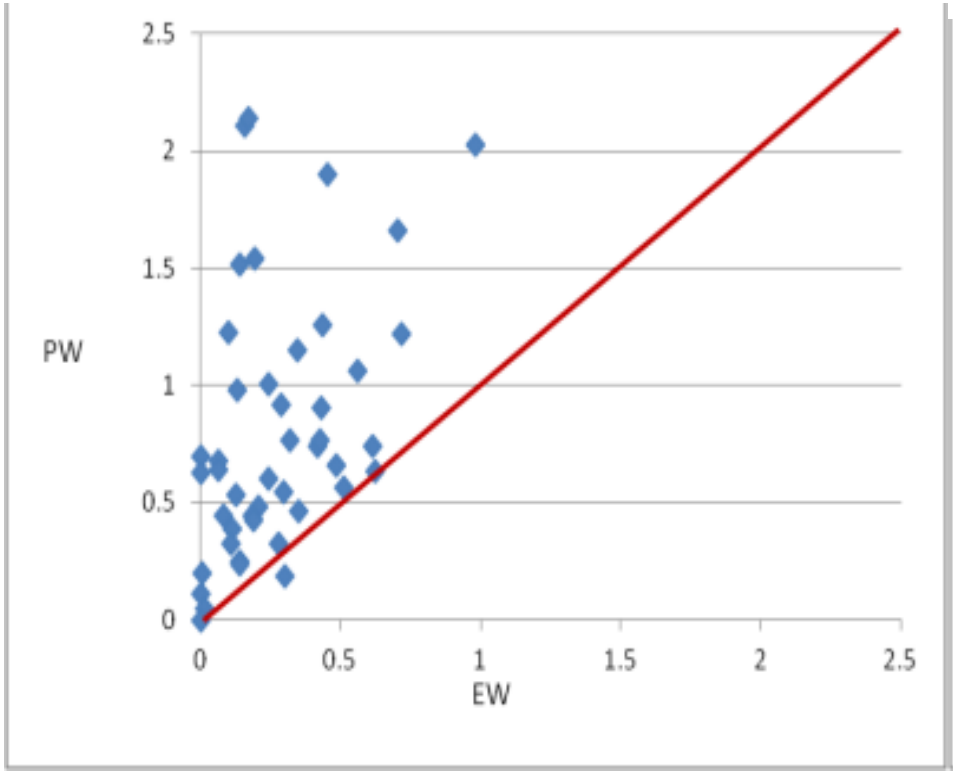
33 Applications post 2008

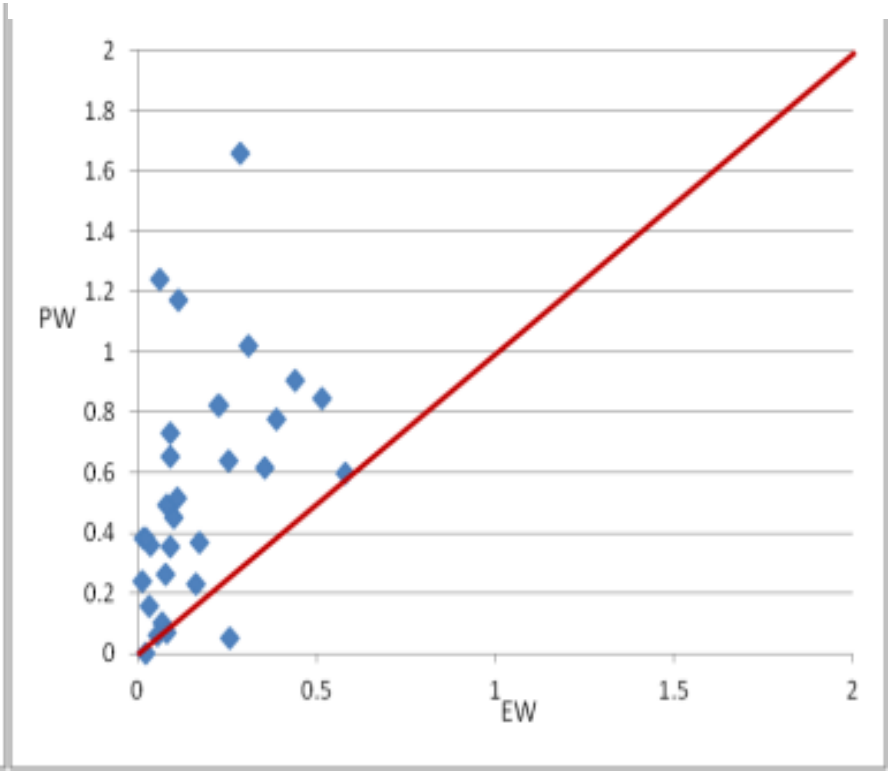
Combined score (calibration * information) Performance & Equal Weights

pre 2008

post 2008

In-sample validation





Averaging Quantiles? *(Lichtendahl et al 2013)*

F and *G* : CDFs from experts 1 and 2; *f*, *g* densities. *HW* , *hw* = CDF and density of the result of averaging the quantiles of *F, G*. Then

$$HW^{-1}(r) = \tfrac{1}{2}\left(F^{-1}(r) + G^{-1}(r) \right).$$

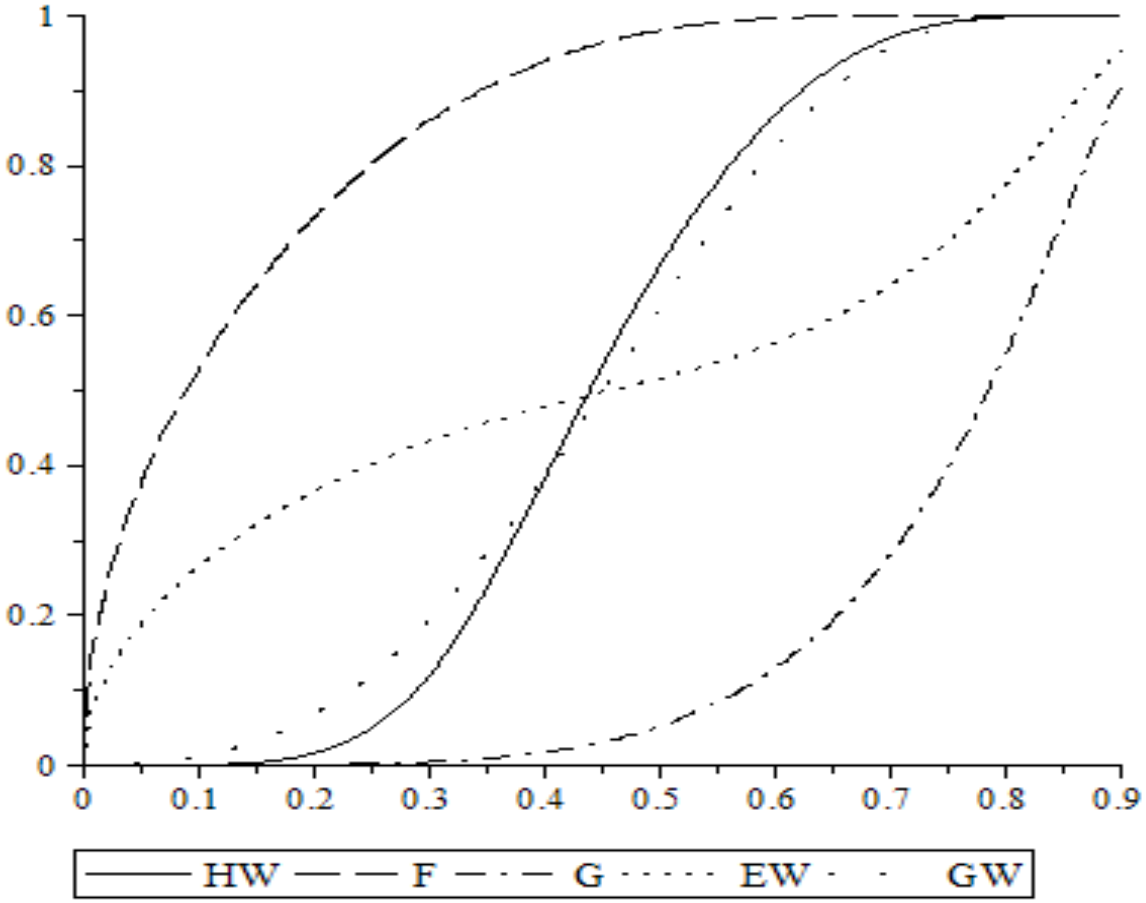
$$1/hw(HW^{-1}(r) = \tfrac{1}{2}\left(1/f\left(F^{-1}(r)\right) + 1/g(G^{-1}(r))\right),$$

$$hw(HW^{-1}(r)) = \frac{2}{\left(1/f\left(F^{-1}(r)\right) + 1/g(G^{-1}(r))\right)}.$$

= *Harmonic Mean*

The harmonic mean of 0.01 and 0.99 is 0.0198. Consider a flexible and tractable class of distributions on [0,1] (a>1, b>0):

$$F(x) = 1 - a^{-\frac{x^b}{1-x^b}}; \; F^{-1}(r) = \left(-\frac{\ln(1-r)}{\ln(a)} \cdot \left(1 - \frac{\ln(1-r)}{\ln(a)} \right)^{-1} \right)^{\frac{1}{b}}$$



	PW			EW			HW				
	cal	inf	comb	cal	inf	comb	cal	inf	comb	#seeds	#exprts
Arkansas	0.499	0.337	0.168	0.386	0.198	0.076	5.55E-02	0.640	3.55E-02	10	4
Arsenic D-R	0.036	2.739	0.098	0.061	1.095	0.067	7.99E-04	1.324	1.06E-03	10	9
ATCEP Err	0.683	0.227	0.155	0.124	0.247	0.031	5.99E-04	1.066	6.38E-04	10	5
Biol agents	0.678	0.610	0.414	0.413	0.244	0.101	3.60E-02	0.884	3.18E-02	12	12
CDC ROI	0.720	2.305	1.660	0.233	1.230	0.286	7.56E-01	1.565	1.18E+00	10	20
CoveringKi	0.720	0.431	0.310	0.628	0.274	0.172	9.03E-01	0.595	5.38E-01	10	5
CREATE	0.394	0.276	0.109	0.061	0.207	0.013	2.77E-04	0.52	1.44E-04	10	7
CWD	0.493	1.215	0.598	0.474	0.930	0.441	7.07E-01	1.494	1.06E+00	10	14
Daniela	0.554	0.634	0.351	0.533	0.168	0.089	1.82E-01	0.520	9.48E-02	7	4
dcpn_fistula	0.119	1.309	0.156	0.059	0.622	0.037	8.78E-08	1.125	9.88E-08	15	14
eBBP	0.833	1.406	1.172	0.358	0.316	0.113	8.04E-02	0.954	7.67E-02	8	14
EffusiveEru	0.664	1.123	0.745	0.286	0.796	0.228	2.65E-02	1.505	3.99E-02	15	10
Erie Carps	0.661	0.856	0.566	0.182	0.281	0.051	3.87E-01	0.754	2.92E-01	8	5
FCEP Error	0.664	0.574	0.381	0.222	0.099	0.022	1.75E-05	0.771	1.35E-05	10	8
Florida	0.756	1.133	0.857	0.756	0.455	0.344	6.98E-02	0.880	6.15E-02	10	7
GL-NIS	0.928	0.209	0.194	0.044	0.307	0.014	5.53E-02	0.842	4.66E-02	14	12
Gerstenberg	0.9302	1.095	1.018	0.6439	0.4815	0.31	8.10E-02	0.9659	7.82E-02	13	9
Goodheart	0.707	0.959	0.678	0.550	0.277	0.153	0.683	0.8884	6.07E-01	10	6
Hemophilia	0.312	0.494	0.154	0.254	0.202	0.051	3.12E-01	0.779	2.43E-01	8	18
IceSheet201	0.399	1.552	0.620	0.492	0.517	0.254	7.96E-02	1.201	9.56E-02	11	10
Illinois	0.337	0.647	0.218	0.620	0.264	0.163	2.37E-03	0.793	1.88E-03	10	5
Liander	0.228	0.524	0.120	0.228	0.484	0.111	2.81E-03	1.198	3.36E-03	10	11
Nebraska	0.033	1.447	0.048	0.368	0.695	0.256	2.40E-05	1.192	2.86E-05	10	4
Obesity	0.440	0.507	0.223	0.070	0.243	0.017	6.68E-04	0.745	4.98E-04	10	4
PHAC T4	0.178	0.351	0.062	0.298	0.211	0.063	1.64E-02	0.640	1.05E-02	13	10
San Diego	0.155	0.758	0.117	0.147	1.012	0.148	1.95E-03	1.545	3.02E-03	10	7
Sheep Scab	0.643	1.310	0.843	0.661	0.780	0.516	1.15E-02	1.411	1.63E-02	15	14
SPEED	0.676	0.777	0.525	0.517	0.751	0.389	2.97E-02	1.165	3.46E-02	16	14
TdC	0.989	1.256	1.242	0.166	0.364	0.060	1.24E-02	1.079	1.34E-02	17	18
Tobacco	0.688	1.062	0.730	0.200	0.451	0.090	2.11E-01	0.708	1.49E-01	10	7
Topaz	0.411	1.455	0.598	0.629	0.922	0.580	8.66E-05	1.528	1.32E-04	16	21
umd_nremo	0.706	1.988	1.404	0.068	0.804	0.054	2.40E-03	1.219	2.93E-03	11	10
Washington	0.200	0.724	0.145	0.155	0.529	0.082	4.2E-01	0.862	3.63E-01	10	5
nr < 0.05	2			1			18				
nr best			26			3			4		
Ave Inf		1.042			0.531			1.0761			

Summary In-sample performance

Geomean ratios (row/col) of combined scores for all post 2008 studies							
	NoOpt	**EW**	**PWg**	**Pwi**	**BE**	**2ndBE**	**HW**
NoOpt	1.000	1.700	0.523	0.453	0.911	0.894	19.183
EW	0.588	1.000	0.308	0.266	0.536	0.526	11.283
PWg	1.911	3.249	1.000	0.865	1.740	1.708	36.657
Pwi	2.210	3.757	1.156	1.000	2.012	1.975	42.391
BE	1.098	1.867	0.575	0.497	1.000	0.982	21.065
2ndBE	1.119	1.902	0.585	0.506	1.019	1.000	21.461
HW	0.052	0.089	0.027	0.024	0.047	0.047	1.000

Out of Sample Validation?

1. Conundrums in Literature ROAT & CODE
2. Cross-Validation
3. New data
4. Explaining OoS Validity

ROAT : Remove-One-at-a-Time

- Expert 1 $P_{heads} = 0.8$ Expert 2 $P_{heads} = 0.2$
- Weights w_1/w_2 = likelihood ratio Ex1 / Ex2

N Heads & N Tails,

LR = $0.8^N \times 0.2^N / 0.2^N \times 0.8^N = 1.$

Remove **one H**, LR = 0.2/0.8 = ¼ = w_1/w_2

$P^{DM}_{heads} = (1/5) \times 0.8 + (4/5) \times 0.2 = 0.32.$

$P^{DM}_{tails} = 0.32.$

P^{DM}_{heads} used to predict Heads

P^{DM}_{tails} used to predict Tails

N=10, LR PW/EW = $(0.32/0.5)^{10} = 0.012.$

Lin and Cheng (2008) examined 28 of the 45 studies and found PW significantly out performing EW, although PW's out-of-sample performance was degraded. **Lin and Cheng (2009)** used ROAT on 40 studies finding no significant difference between PW and EW. These publications do not report that their code has been vetted against EXCALIBUR, and there are very large differences between Lin and Cheng 2008 and Cooke and Goossens 2008

	Lin&Cheng study	Lin & Cheng Parameters		TUD Calbr Vbls / eff nr	Lin and Cheng 2008, Table 1 "within sample"		Cooke and Goossens 2008 Table 1	
	NAME	# expert	#calibration vbls		PWComb	EWComb	PWComb	EWComb
1	Acrylonitrile	7	10	Same	0.47	0.44	0.764	0.423
3	Dike ring	17	47	Same	0.42	0.03	0.2456	0.03768
4	Flanges	10	8	Same	0.6	0.2	0.905	0.4274
5	Crane risk	8	10	12/11	0.93	0.28	1.148	0.345
6	Groundwater	7	10	Same	0.95	0.05	2.106	0.158
7	Space debris	7	26	**26**/18	6.0E-06	0.13	0.25	0.14
8	Composite materials	6	12	Same	0.55	0.21	0.39	0.111
10	Dry deposition	8	14	Same	0.48	0.003	0.697	0.001
11	Atmospheric dispersion	8	23	Same	0.38	0.18	0.9785	0.129
12	Early health effects	7/9	15	Same	0.06	0.01	0.0496	0.01153
14	Soil transfer	4	31	Same	1.0E-06	1.0E-07	1.0E-04	9.7E-05
15	Wet deposition	7	19	Same	0.11	0.002	0.113	0.00073
20	Movable barriers	8	14	Same	0.06	0.13	0.535	0.125
21	Real estate	5	31	Same	0.7	0.001	0.6296	0.0009
22	River dredging	6	8	Same	0.54	0.18	0.447	0.185
23	Sulpher trioxide	4	7	Same	2.53	0.3	0.547	0.294

The out-of-sample code of **Flandoli et al (2011)** has errors in optimization and scaling. Two of the 4 cases analysed had 15 and 16 calibration variables, enabling comparison with results from the Eggstaff code. Flandoli et al draw 500 random samples from training sets of fixed size and compute the scores on the complementary tsest set. The EW scores agree reasonably, but the PW scores do not.

		PW			EW		
		Sa	Inf	Comb	Sa	Inf	Comb
Pbearl 8 training, 7 test	Eggstaff	0.149	0.617	0.072	0.271	0.167	0.046
	Flandoli Table 8	0.229	0.407	0.093	0.273	0.167	0.046
Vesuvius 8 training 8 test	Eggstaff	0.277	1.176	0.240	0.520	0.756	0.383
	Flandoli Table 4	0.449	0.896	0.377	0.519	0.720	0.365

ROAT Used by

- Cooke, R.M. (2008) Special issue on expert judgment, Editor's Introduction Reliability Engineering & System Safety, 93, Available online 12 March 2007, Volume 93, Issue 5, May 2008, Pages 655-656.
- Clemen, R.T (2008)" Comment on Cooke's classical method" Reliability Engineering & System Safety, Volume 93, Issue 5, May 2008, Pages 760-765
- Lin, Shi-Woei, and Bier, V.M. (2008) "A Study of Expert Overconfidence" Reliability Engineering & System Safety, 93, 775-777, Available online 12 March 2007. Volume 93, Issue 5.
- Lin, Shi-Woei, Cheng, Chih-Hsing (2008) "Can Cooke's Model Sift Out Better Experts and Produce Well-Calibrated Aggregated Probabilities?" Department of Business Administration, Yuan Ze University, Chung-Li, Taiwan Proceedings of the 2008 IEEE IEEM
- Lin, Shi-Woei, Cheng, Chih-Hsing (2009) "The reliability of aggregated probability judgments obtained through Cooke's classical model", Journal of Modelling in Management, Vol. 4 Iss: 2, pp.149 – 161,
- Shi-Woei Lin, Ssu-Wei Huang, (2012) "Effects of overconfidence and dependence on aggregated probability judgments", Journal of Modelling in Management, Vol. 7 Iss: 1, pp.6 - 22
- Cooke, R.M. (2012) "Pitfalls of ROAT Cross Validation Comment on Effects of Overconfidence and Dependence on Aggregated Probability Judgments" ,Journal of Modelling in Management, vol.7, nr. 1, pp 20-22, ISSN 1746-5664.

CROSS Validation

- Cooke, R.M., (2008) Response to Comments, Special issue on expert judgment Reliability Engineering & System Safety, 93, 775-777, Available online 12 March 2007. Volume 93, Issue 5, May 2008.
- Flandoli, F. Giorgi W.P. Aspinall, W. and Neri A (2010). " Comparing the performance of different expert elicitation models using a cross-validation technique" appearing in Reliability engineering and System Safety.
- Eggstaff,J.W., Mazzuchi,T.A. Sarkani, S. (2014) The Effect of the Number of Seed Variables on the Performance of Cooke's Classical Model, Reliability Engineering and System Safety 121 (2014) 72–82.
- Burgman, M. et al (20??) Intelligence Game, IARPA shoot-out

Out-of-sample Cross Validtion

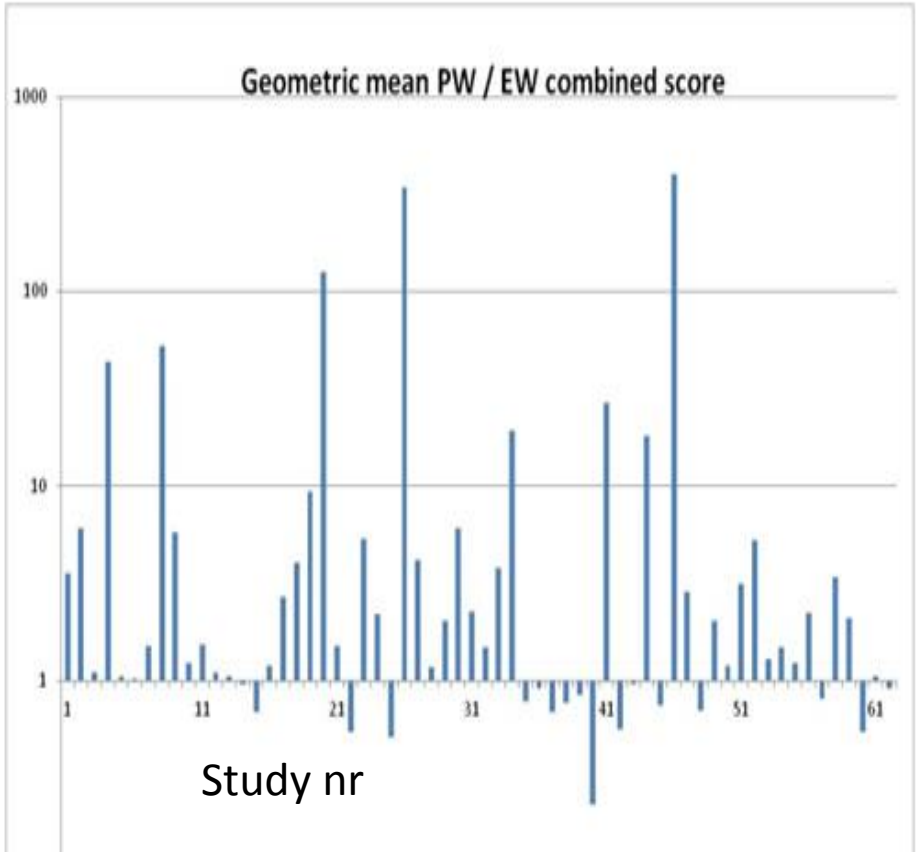
- N seed vbls
- K < N training set; N-K test set
- **WHICH K?**

- K small, low power to resolve experts
- K large, low power to resolve DM
- K = N-1, ROAT bias
- K = N/2…all k-tuples Law of Large Numbers??

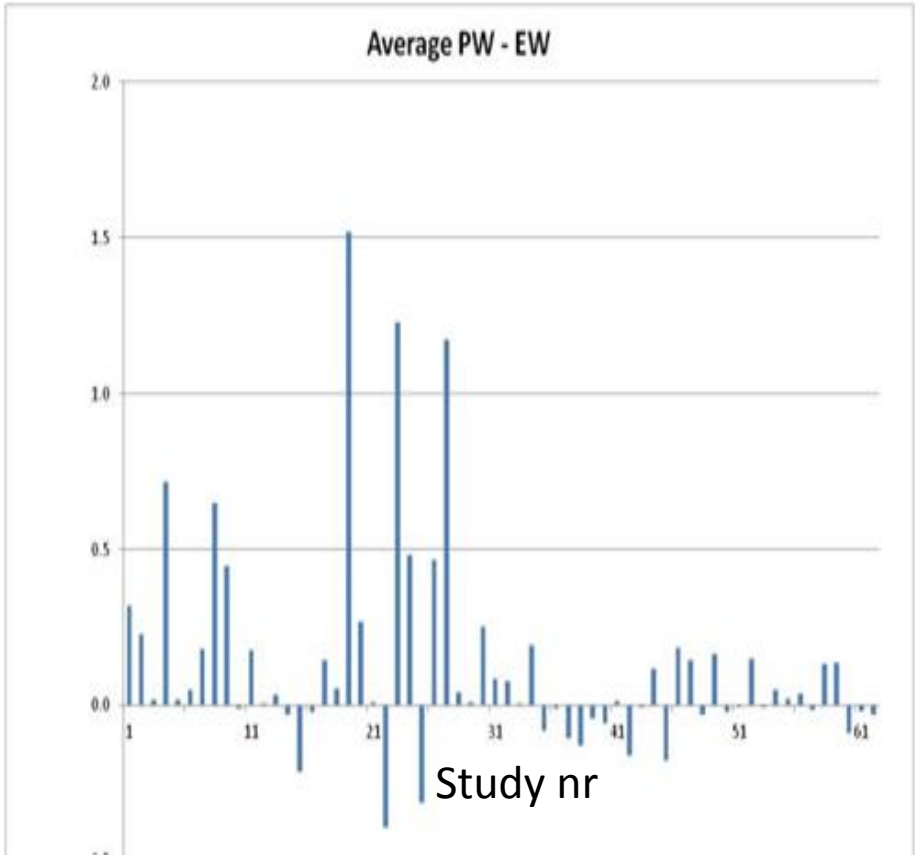
Eggstaff et al

- For K = 1….#seeds = N;
  - Initialize on EACH training sets size K
  - Score PW and EW on each test set
  - For given K average PW and EW scores
- Aggregate over all K by
  - Arithmean of PW-EW [affected by statistical power loss as K ↗]
  - Geomean of PW/EW [better, dimensionless]

%(PW > EW) = 73% (Eggstaff et al)

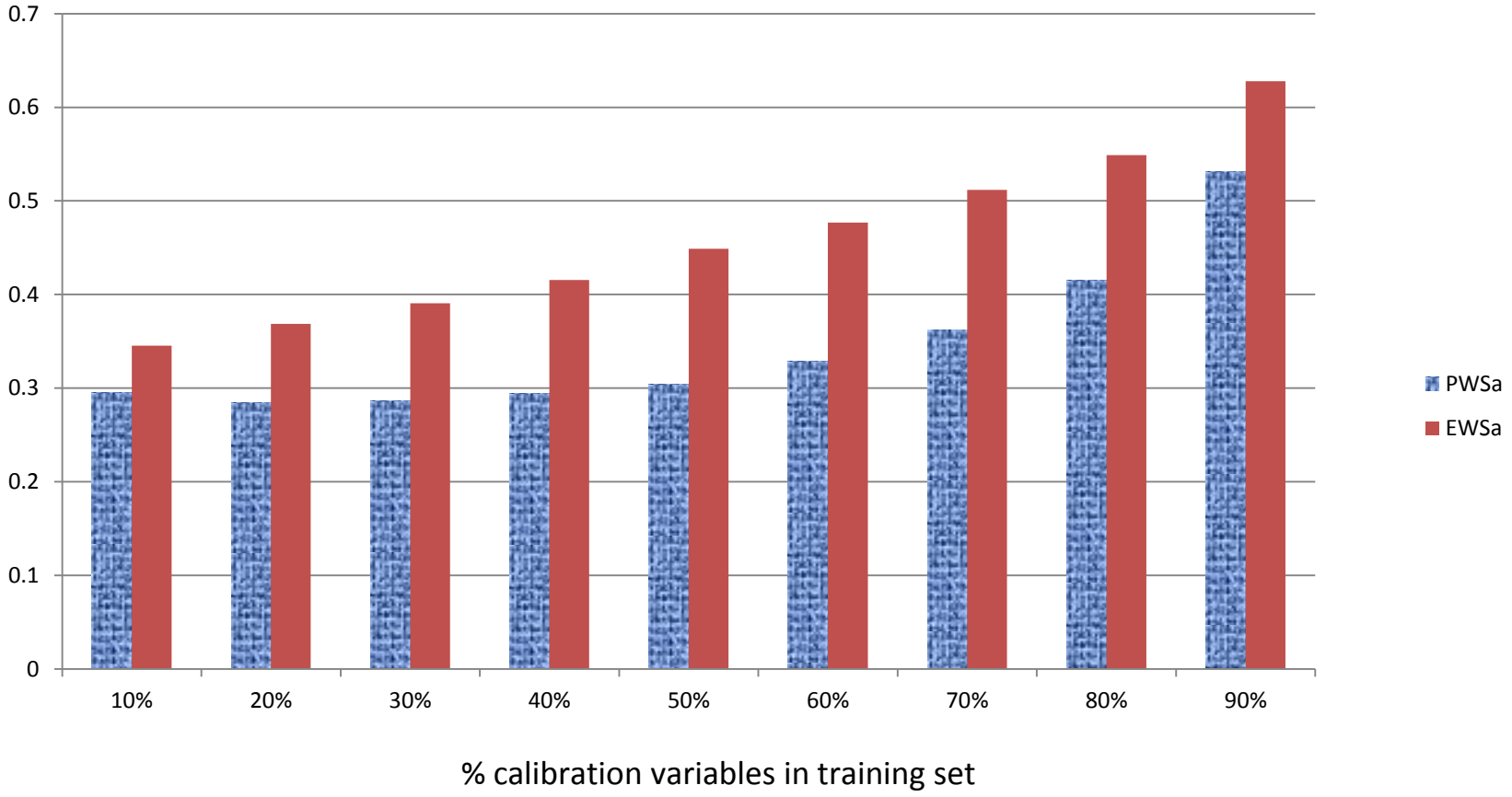


Smallest to largest # seeds

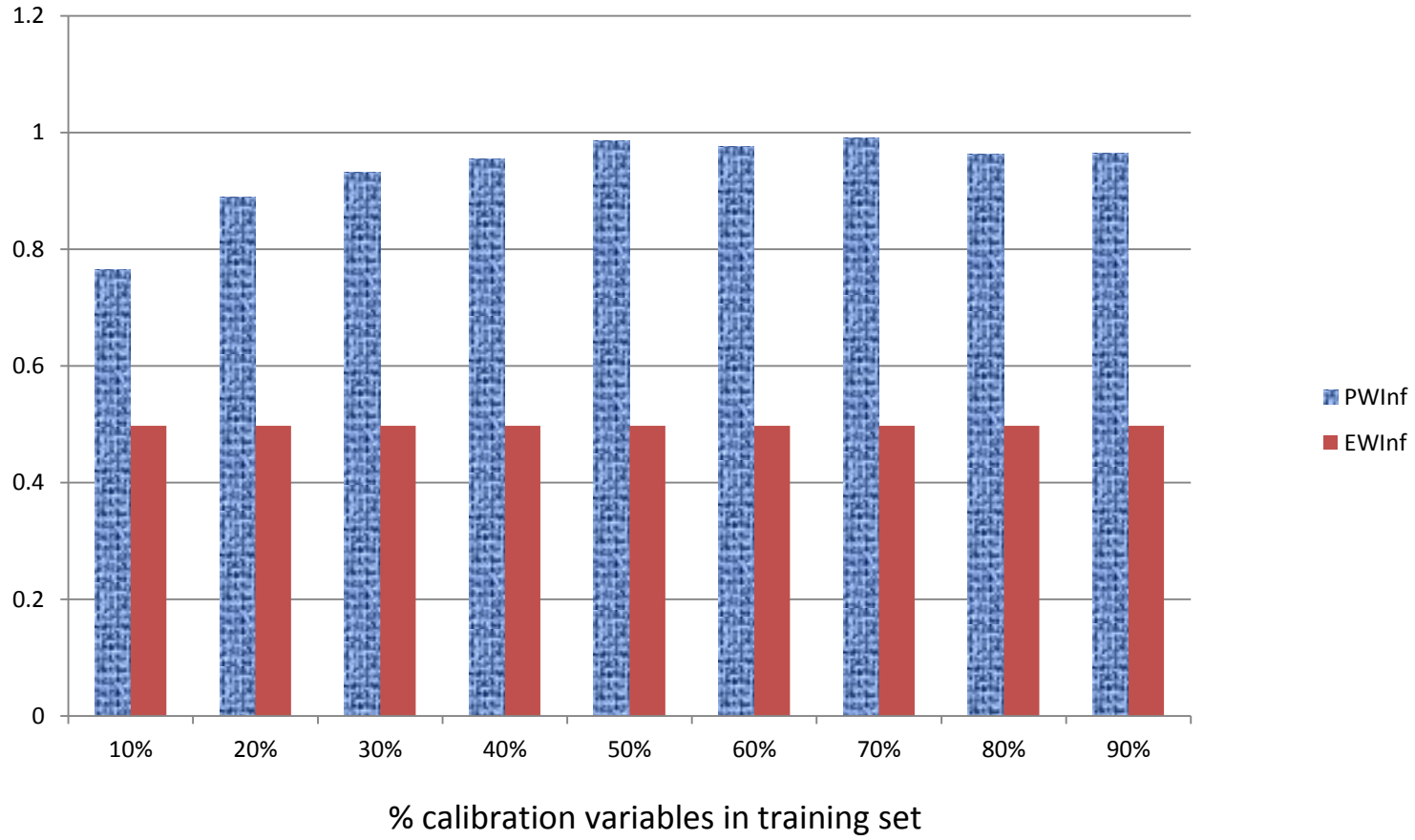


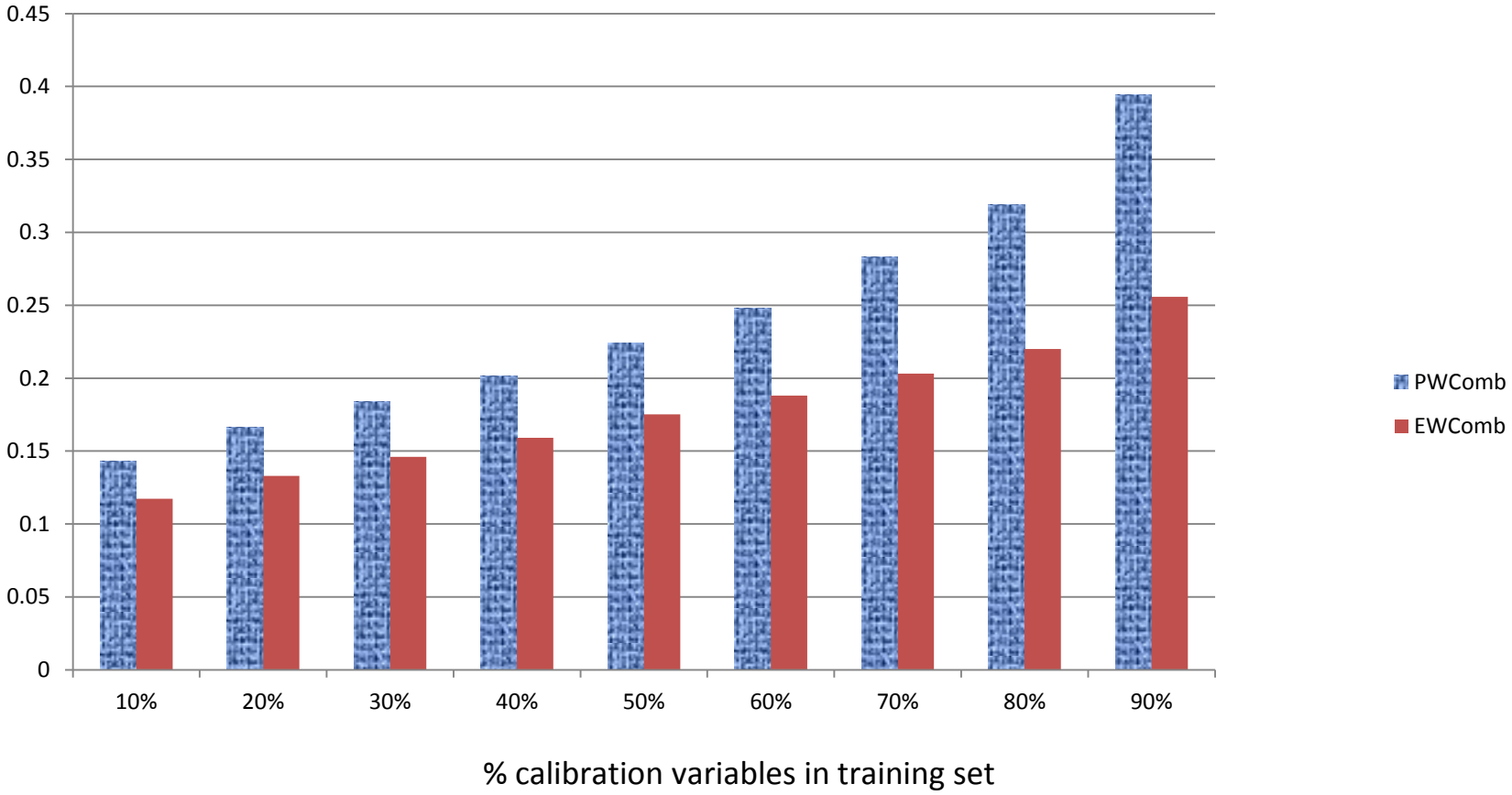
Smallest to largest # seeds

Average over all studies per % training set size of the average *PWSa* and average *EWSa*

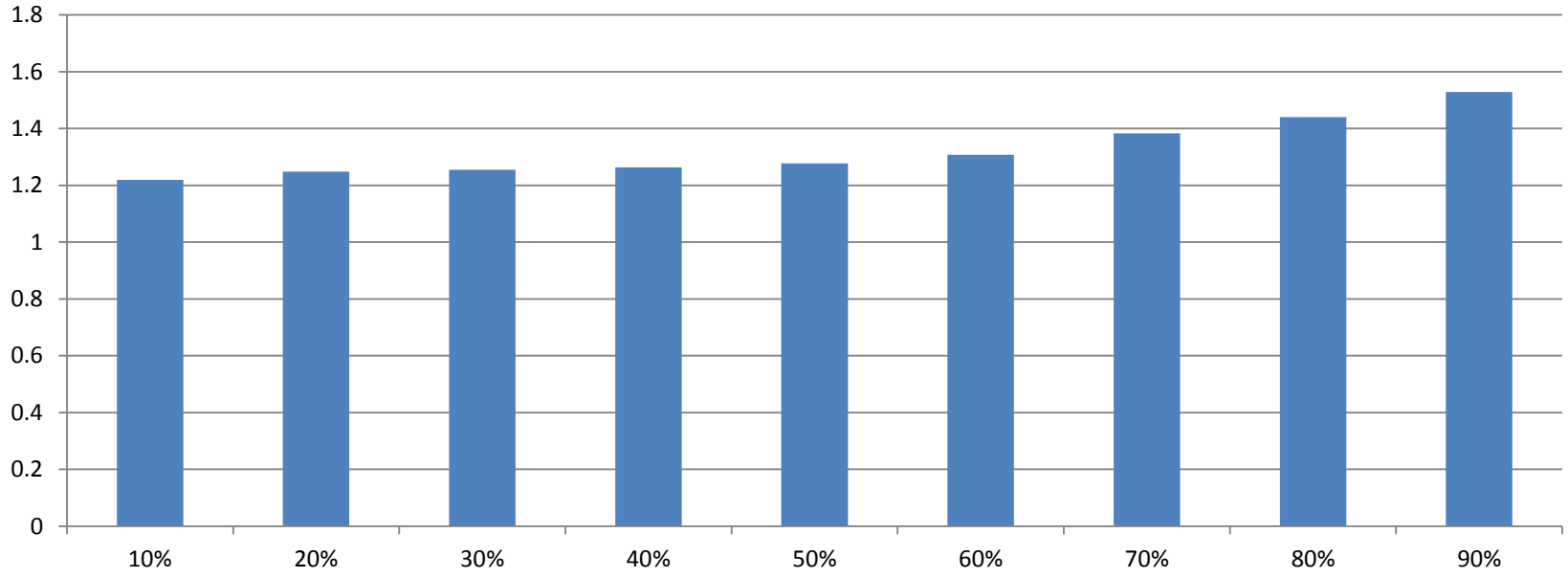


Average over all studies per % training set size of the average *PWInf* and average *EWInf*

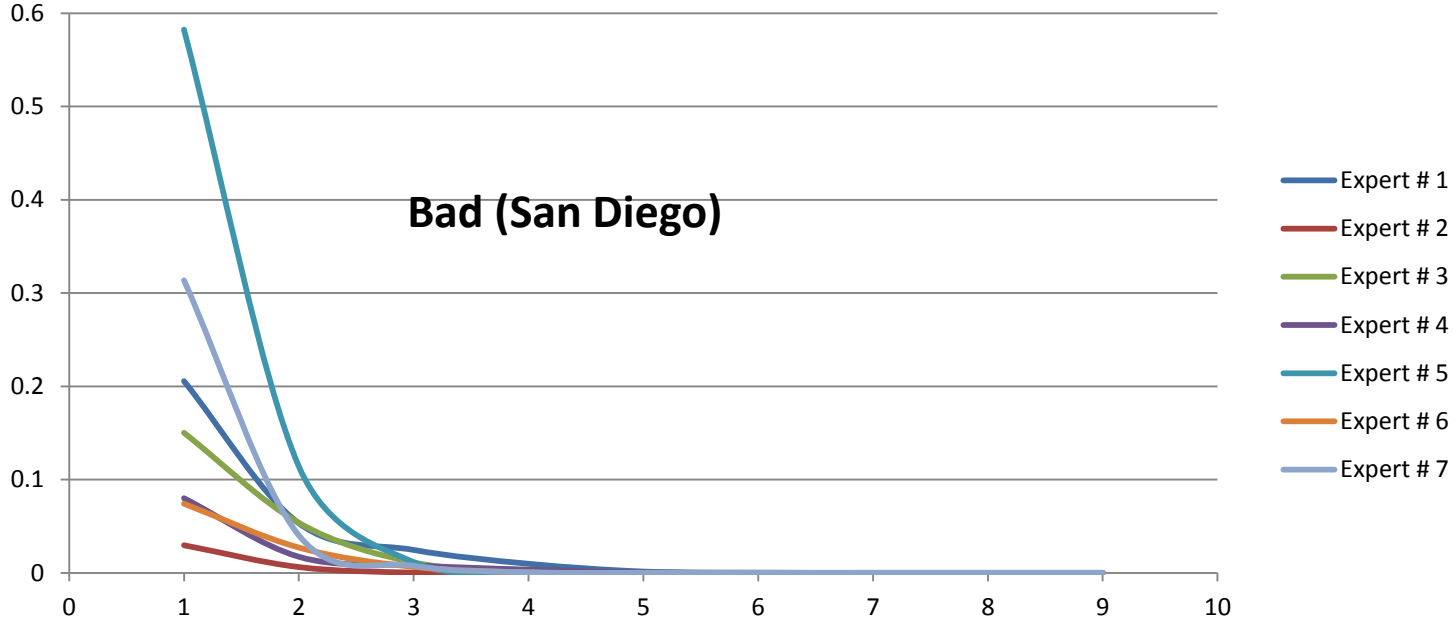


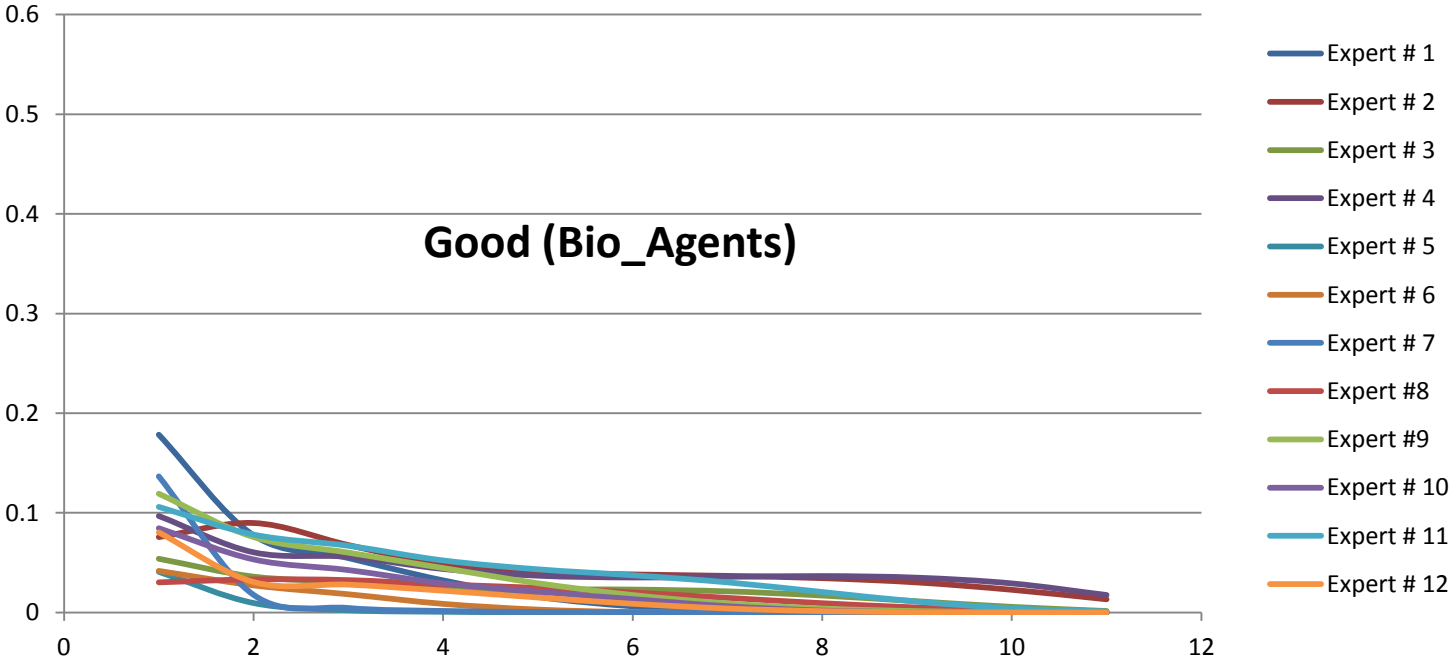

0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
10%
20%
30%
40%
50%
60%
70%
80%
90%
% calibration variables in training set
PWComb
EWComb

Geomean(AvPWComb/AvEWComb) over all studies

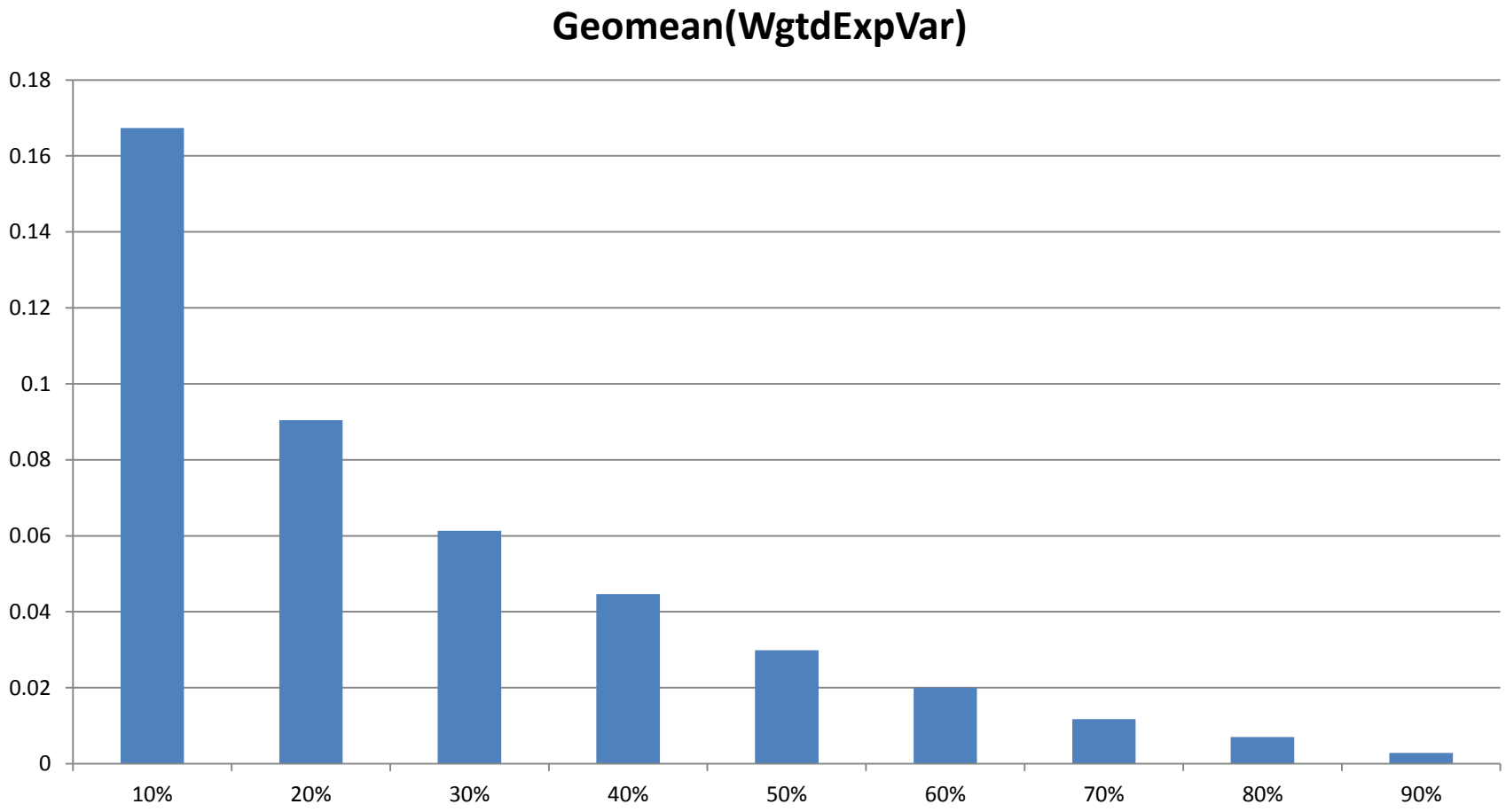


Variance of experts' combined score



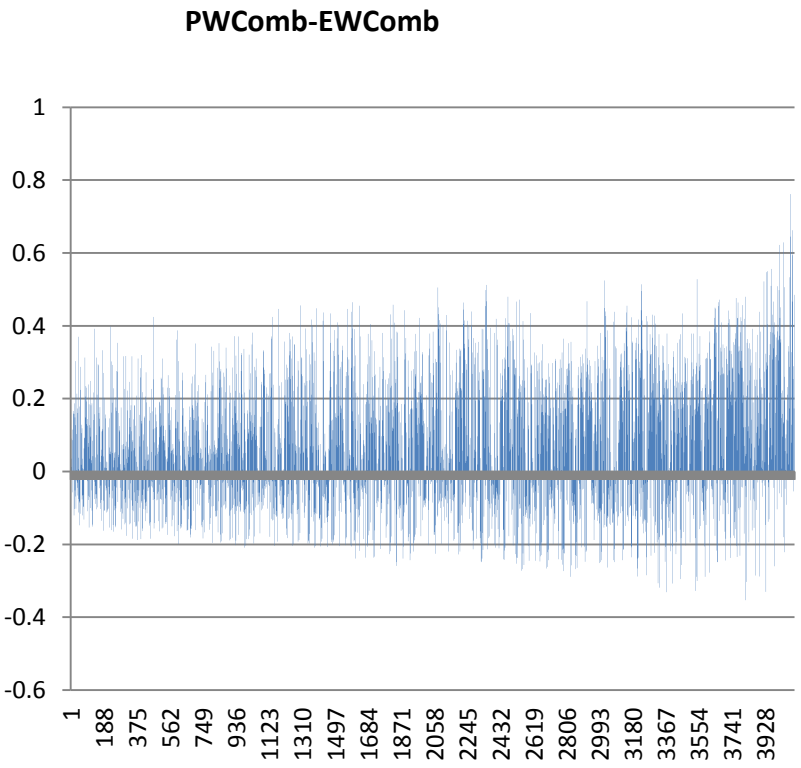


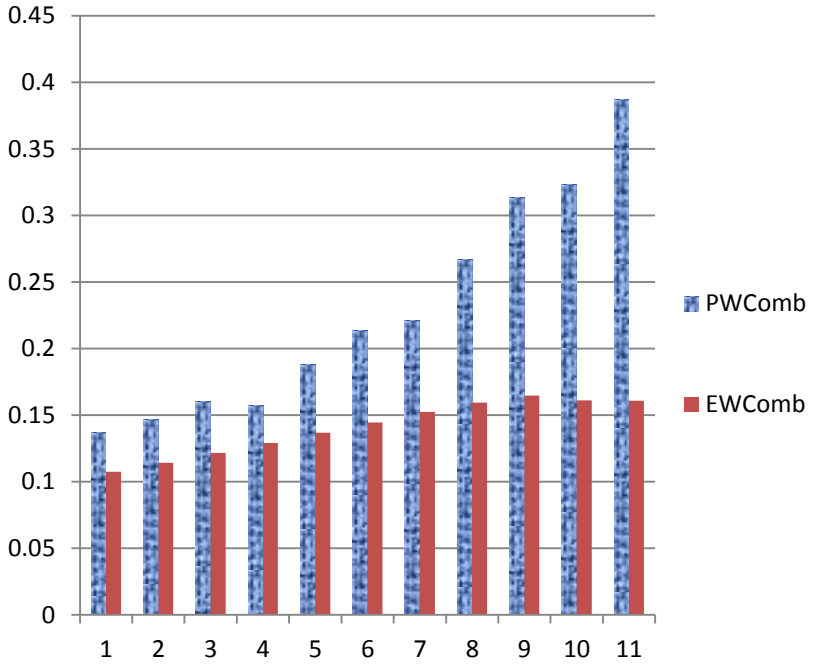
Overall variance in experts' combined score



Biol_Agents

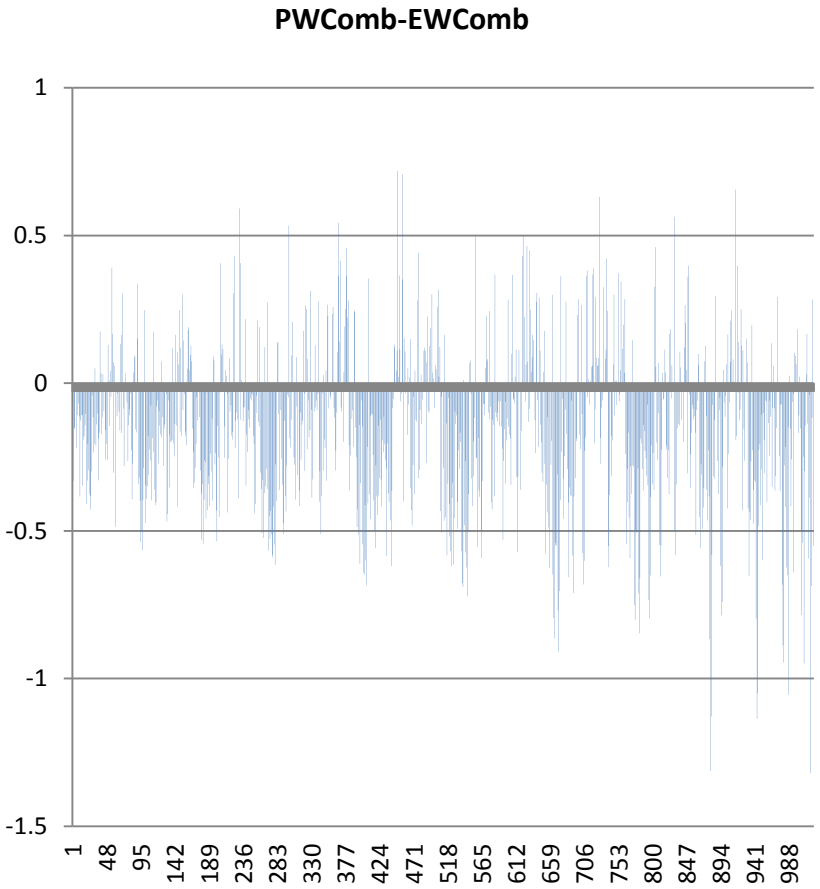
Left, differences of combined scores for *PW* and *EW* for all training sets, from size 1 to size 10. Right, combined scores of *PW* and *EW* averaged per training set size,

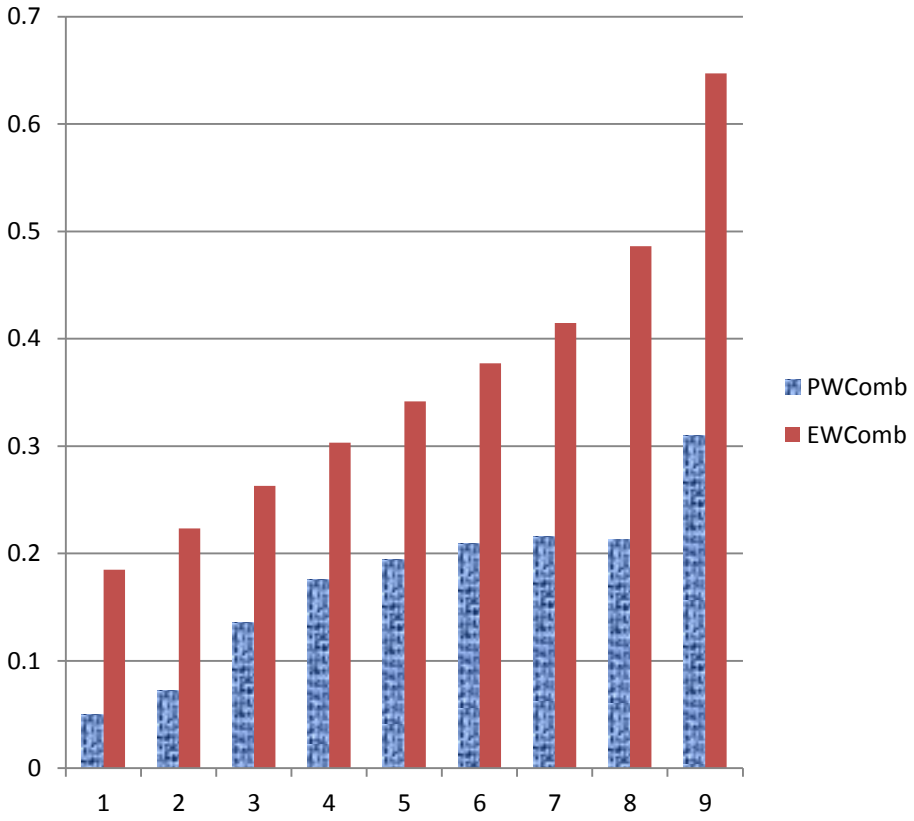




San Diego

Left, differences of combined scores for *PW* and *EW* for all comparisons, from size 1 to size 11; Right, combined scores of *PW* and *EW* averaged per training set size,





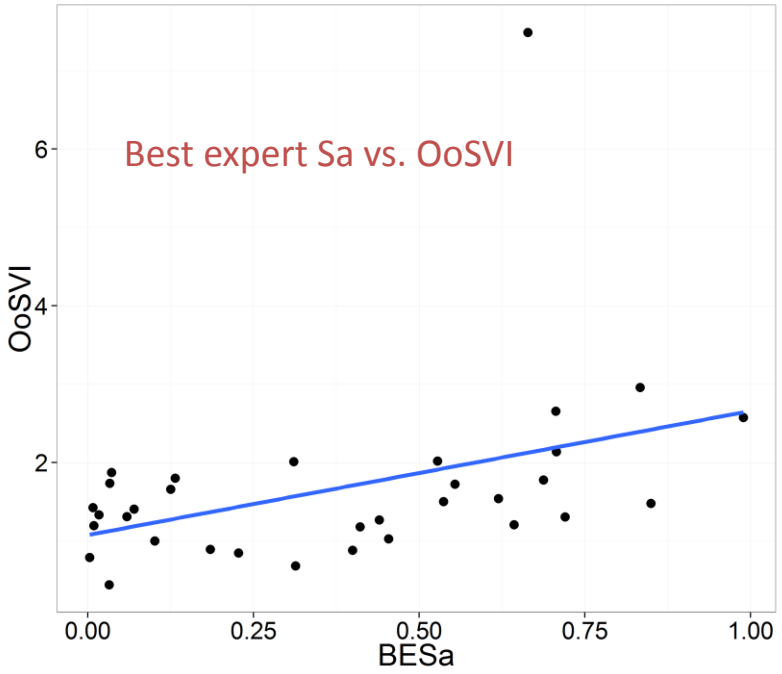
Average *PWComb* /Average *EWComb* for % training sets

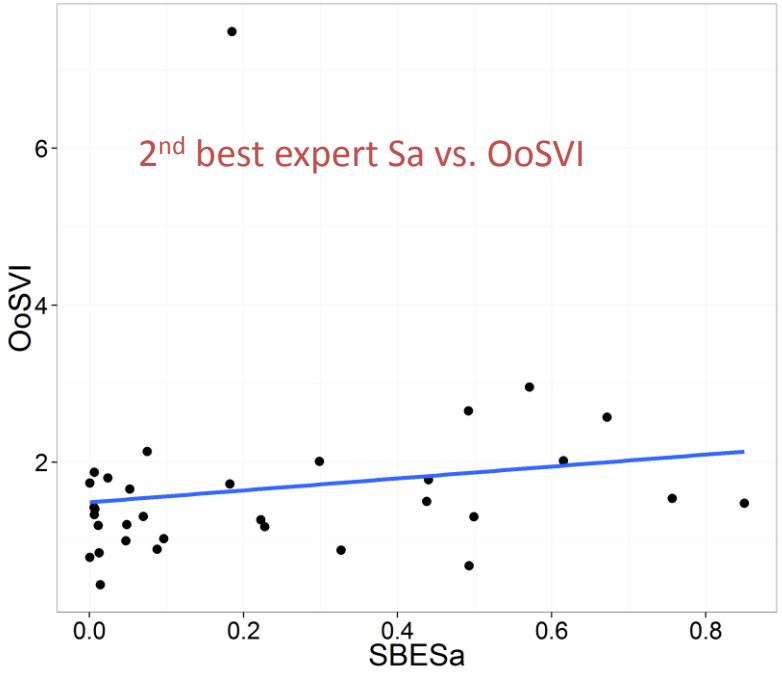
	Training set size as percent of calibration variables									Row
	10%	20%	30%	40%	50%	60%	70%	80%	90%	Geomean
Arkansas	1.132	1.256	1.523	1.382	1.423	1.322	1.306	1.404	1.718	**1.376**
Arsenic	1.039	1.035	1.073	1.117	1.293	1.431	1.722	1.874	1.961	**1.352**
ATCEP	1.955	1.677	1.380	1.455	1.166	1.149	1.156	0.997	0.799	**1.262**
Biol_Agent	1.278	1.280	1.217	1.373	1.477	1.453	1.676	2.008	2.405	**1.536**
CDC_ROI	1.006	1.199	1.112	1.229	1.004	1.109	1.107	1.305	1.399	**1.157**
CoveringKids	1.032	1.510	1.407	1.478	1.487	1.463	1.517	1.538	1.427	**1.420**
Create	0.890	0.789	0.763	0.817	1.046	1.277	1.278	1.331	1.142	**1.014**
CWD	1.328	0.980	1.031	0.907	0.812	0.729	0.708	0.680	0.756	**0.862**
Daniela	1.051	1.051	1.086	1.099	1.137	1.137	1.815	1.721	1.721	**1.279**
Fistula	0.262	0.964	0.918	1.039	1.147	1.354	1.362	1.426	1.910	**1.037**
eBPP	1.859	1.844	2.027	1.778	2.402	2.576	2.727	2.958	4.033	**2.384**
Eff_Erup	0.965	0.903	0.903	0.796	0.651	0.664	0.892	0.892	0.919	**0.835**
Erie_Carp	2.920	2.612	2.684	2.567	1.856	1.787	2.047	2.017	2.909	**2.339**
FCEP	3.843	7.704	7.704	7.908	8.897	8.826	7.485	7.485	5.713	**7.091**
Florida	0.920	0.445	0.657	0.695	0.750	0.886	0.979	1.364	1.412	**0.851**
Gerstenberger	1.056	1.183	1.152	1.683	1.651	1.670	1.562	1.501	1.604	**1.431**
GL_NIS	2.177	1.847	1.672	1.477	1.186	1.134	1.066	1.024	0.809	**1.316**
Goodheart	1.180	1.291	1.595	1.441	1.366	1.480	1.611	2.136	2.607	**1.586**
Hemopilia	1.638	2.019	2.019	1.862	2.938	1.534	1.476	1.476	2.808	**1.913**
IceSheets	1.266	0.861	0.867	0.814	0.850	0.779	0.807	0.880	0.903	**0.883**
Illinois	0.671	0.697	0.821	0.798	0.867	1.126	1.407	1.800	2.484	**1.073**
Liander	0.881	0.746	0.488	0.614	0.669	0.780	0.870	0.788	0.575	**0.700**
Nebraska	0.559	0.340	0.389	0.517	0.692	0.978	1.393	1.733	1.892	**0.789**
Obesity	3.569	2.383	2.430	2.105	1.842	1.586	1.361	1.267	1.815	**1.945**
PHAC_T4	1.057	0.833	0.709	0.650	0.853	0.974	1.106	1.195	1.180	**0.931**
San_Diego	0.273	0.327	0.516	0.578	0.569	0.555	0.519	0.439	0.478	**0.460**
Sheep	0.772	0.866	0.828	0.902	1.018	1.033	1.119	1.204	1.432	**1.001**
SPEED	0.575	0.661	0.595	0.614	0.632	0.750	0.783	0.844	0.835	**0.692**
TDC	3.413	3.850	4.207	3.416	2.794	2.754	2.667	2.573	2.557	**3.088**
Tobacco	2.179	2.167	2.019	1.965	1.861	1.928	1.830	1.778	1.472	**1.900**
Topaz	0.863	0.860	0.860	0.941	0.966	1.050	1.119	1.178	1.182	**0.994**
UMD_NREMOVAL	1.882	5.221	4.555	4.508	3.634	3.340	3.192	2.654	2.236	**3.300**
Washington	3.614	2.148	1.726	1.370	1.119	1.119	1.142	1.308	1.334	**1.529**
Column Geomean	**1.219**	**1.249**	**1.254**	**1.263**	**1.277**	**1.308**	**1.383**	**1.440**	**1.528**	**1.321**
number > 1	**22**	**19**	**20**	**20**	**22**	**24**	**26**	**26**	**25**	**23**
P(this many success or more in 33 trials) on null hypothesis	0.040	0.243	0.148	0.148	0.040	0.007	0.0007	0.001	0.002	0.018

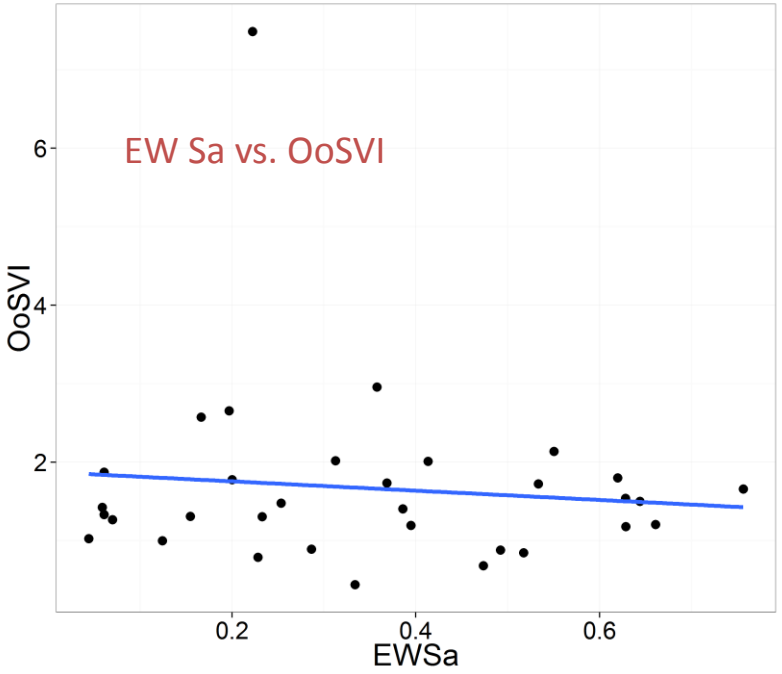
Daniela	1.051	1.051	1.086	1.099	1.137	1.137	1.815	1.721	1.721	**1.279**
Fistula	0.262	0.964	0.918	1.039	1.147	1.354	1.362	1.426	1.910	**1.037**
eBPP	1.859	1.844	2.027	1.778	2.402	2.576	2.727	2.958	4.033	**2.384**
Eff_Erup	0.965	0.903	0.903	0.796	0.651	0.664	0.892	0.892	0.919	**0.835**
Erie_Carp	2.920	2.612	2.684	2.567	1.856	1.787	2.047	2.017	2.909	**2.339**
FCEP	3.843	7.704	7.704	7.908	8.897	8.826	7.485	7.485	5.713	**7.091**
Florida	0.918	0.446	0.738	0.769	0.822	1.194	1.294	1.658	1.841	**0.989**
Gerstenberger	1.056	1.183	1.152	1.683	1.651	1.670	1.562	1.501	1.604	**1.431**
GL_NIS	2.177	1.847	1.672	1.477	1.186	1.134	1.066	1.024	0.809	**1.316**
Goodheart	1.180	1.291	1.595	1.441	1.366	1.480	1.611	2.136	2.607	**1.586**
Hemopilia	1.638	2.019	2.019	1.862	2.938	1.534	1.476	1.476	2.808	**1.913**
IceSheets	1.266	0.861	0.867	0.814	0.850	0.779	0.807	0.880	0.903	**0.883**
Illinois	0.671	0.697	0.821	0.798	0.867	1.126	1.407	1.800	2.484	**1.073**
Liander	0.881	0.746	0.488	0.614	0.669	0.780	0.870	0.788	0.575	**0.700**
Nebraska	0.559	0.340	0.389	0.517	0.692	0.978	1.393	1.733	1.892	**0.789**
Obesity	3.569	2.383	2.430	2.105	1.842	1.586	1.361	1.267	1.815	**1.945**
PHAC_T4	1.057	0.833	0.709	0.650	0.853	0.974	1.106	1.195	1.180	**0.931**
San_Diego	0.273	0.327	0.516	0.578	0.569	0.555	0.519	0.439	0.478	**0.460**
Sheep	0.772	0.866	0.828	0.902	1.018	1.033	1.119	1.204	1.432	**1.001**
SPEED									335	**0.692**
TDC									557	**3.088**
Tobacco									472	**1.900**
Topaz									182	**0.994**
UMD_NREMOVAL	1.882	5.221	4.555	4.508	3.634	3.340	3.192	2.654	2.236	**3.300**
Washington	3.614	2.148	1.726	1.370	1.119	1.119	1.142	1.308	1.334	**1.529**
Column Geomean	**1.219**	**1.249**	**1.259**	**1.267**	**1.281**	**1.320**	**1.395**	**1.449**	**1.541**	**1.327**
number > 1	**22**	**19**	**20**	**20**	**22**	**25**	**27**	**26**	**25**	**23**
P(this many success or more in 33 trials) on null hypothesis	0.040	0.243	0.148	0.148	0.040	0.002	0.0002	0.001	0.002	0.018

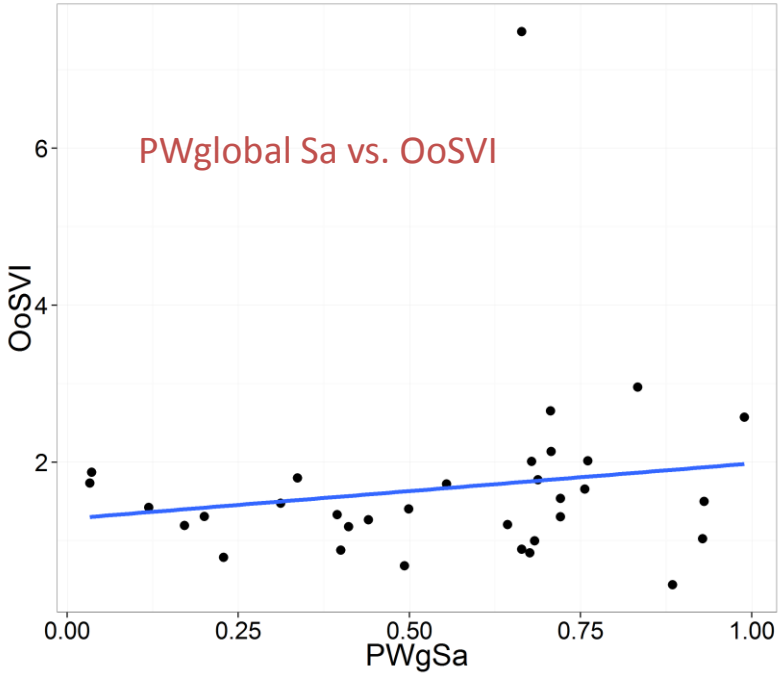
For each percentage split of 50% or more, the null hypothesis would be rejected.

What Explains OoS Validity?

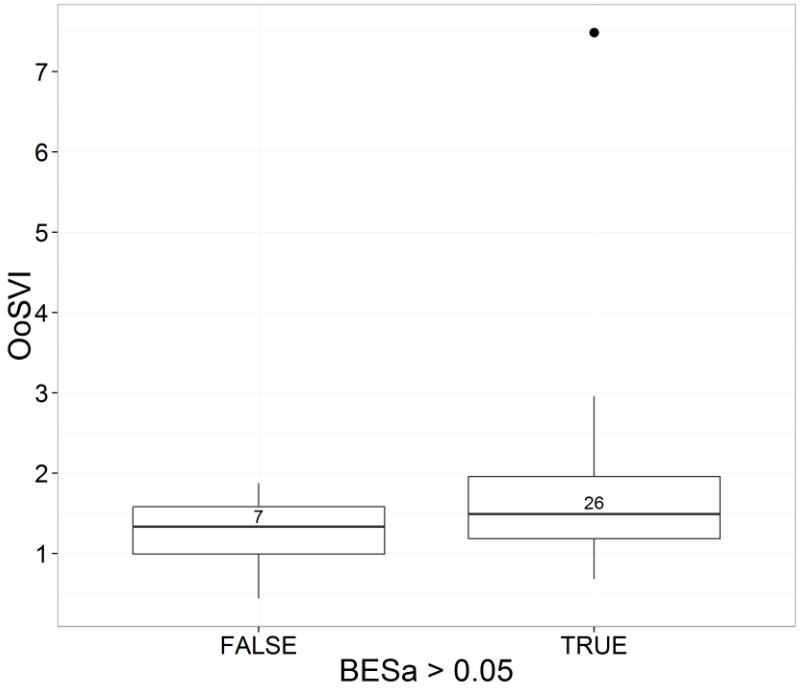

Best expert Sa vs. OoSVI
OoSVI
BESa
0.00
0.25
0.50
0.75
1.00
2
4
6

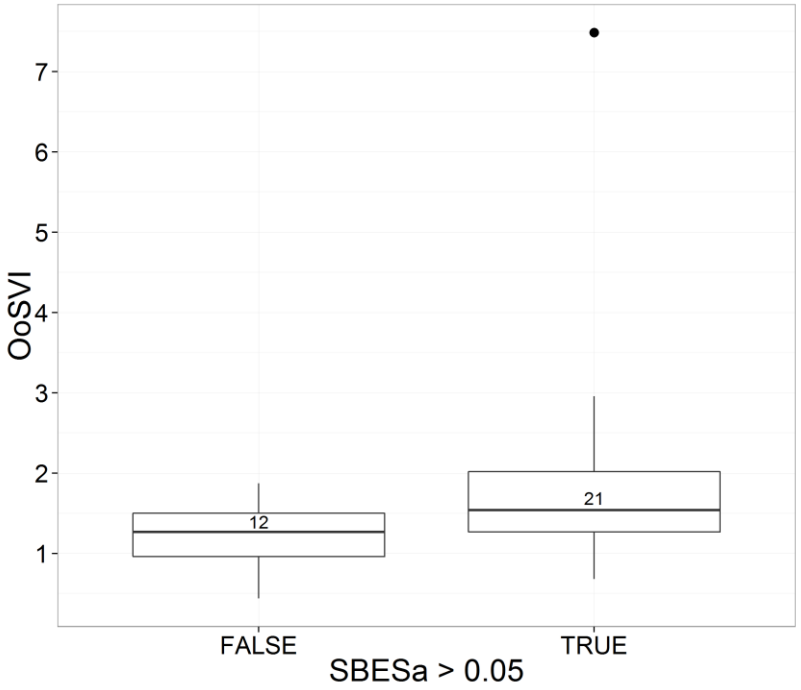

2nd best expert Sa vs. OoSVI
OoSVI
SBESa
0.0
0.2
0.4
0.6
0.8
2
4
6


EW Sa vs. OoSVI
OoSVI
EWSa
0.2
0.4
0.6
2
4
6


PWglobal Sa vs. OoSVI
OoSVI
PWgSa
0.00
0.25
0.50
0.75
1.00
2
4
6

OoSVI improves when BESa and SBESa > 0.05





Conclusions

1. Use OoSVI to study out of sample validity

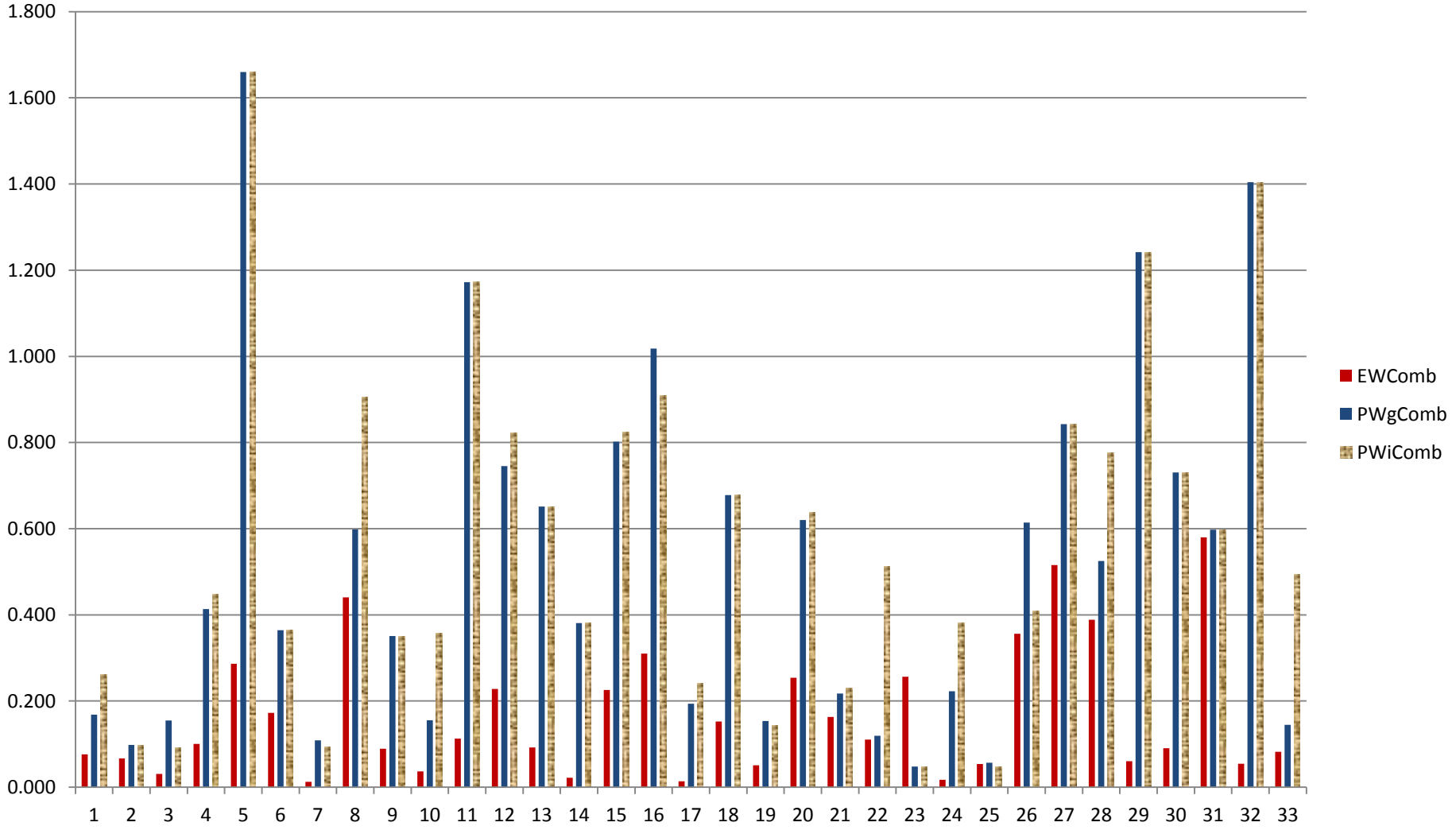
2. Hypothesis:

H_0 : *PWgs not better than EW: P(OoSVI >1) = 0.5, studies independent*

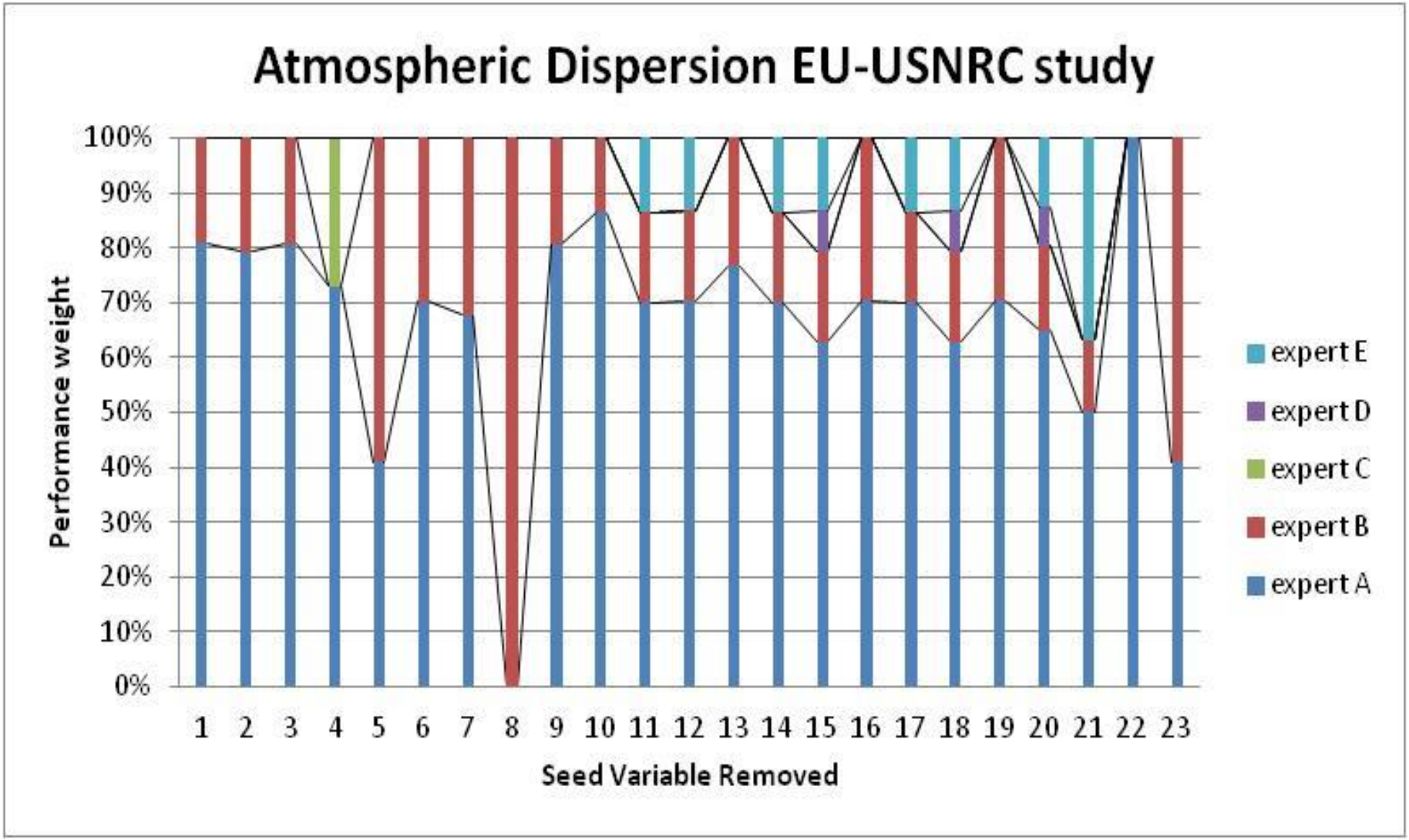
$P(H_0 \mid 33 \text{ studies}) = 0.001.$

$P(H_0 \mid \text{all data}) = 2.5E\text{-}5$

33 post 2006 studies were contracted and overseen by inter alia the Robert Wood Johnson Foundation, US EPA, US NOAA, US DHS, Public Health Agency of Canada, PrioNet (Canada), Sanguin, British Government, European Community, NUMO (Japan), and Bristol University (UK).

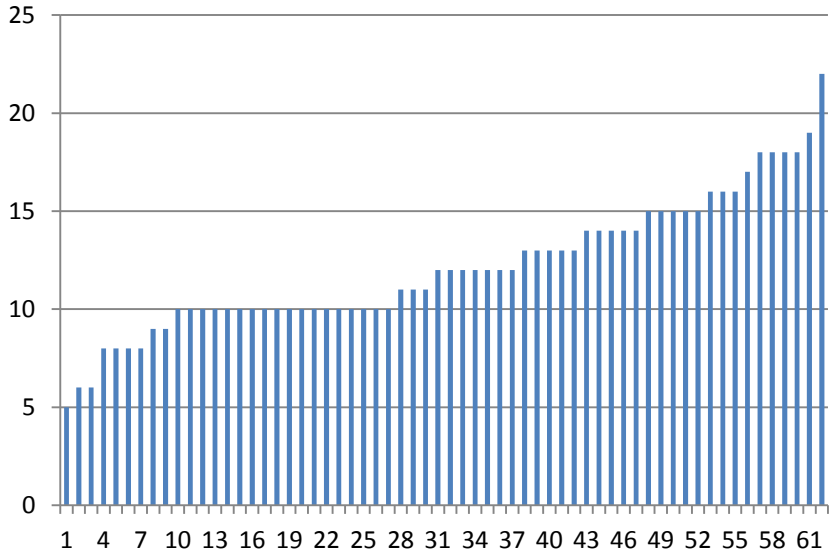


Variation of expert weights under one-at-a-time seed variable exclusion.

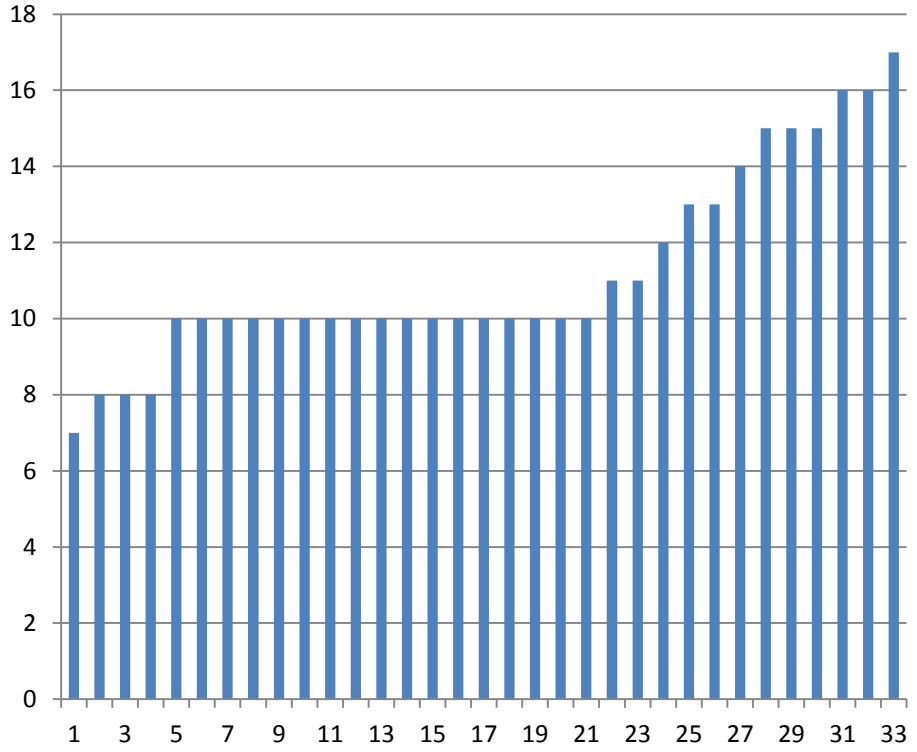


Distribution of number of calibration variables (vertical axis) in Eggstaff et al 2014 and the present study. The horizontal axis is study number.

Eggsstaff et al (2013)



Present study



Isolate the growth of *PWComb* that is not due to decreasing statistical power

PWComb(t,s) = PW combined score on training set *t* of study *s.*

$Av_{\#t=k}$ *PWComb(t,s)* = average of *PWComb(t,s)* over all training sets of size *k* of study *s.*

Similar for *EWComb*.

Fix *s* and fixing training size *t;* ***EWSa(t,s)*** **and** ***EWInf(t,s)*** **are nearly independent**: Mean and standard deviation over all studies and all training percentage sizes of

$Av_{\#t=k}EWComb(t,s) - Av_{\#t=k}EWSa(t,s) \times Av_{\#t=k}EWInf(t,s)$ = (-4.3E-4, 6.5E-4). Therefore, for all *s*

$$Av_{\#t=k}\ \frac{PWSA(t,s)}{Av_{\#t=k}\,EWSa(t,s)} \times \frac{PWInf(t,s)}{Av_{\#t=k}\,EWInf(t,s)} = \frac{Av_{\#t=k}\,PWComb(t,s)}{Av_{\#t=k}\,EWSa(t,s) \times Av_{\#t=k}\,EWInf(t,s)}$$

Because of independence, RHD differs very little from

$$\frac{Av_{\#t=k}\,PWComb(t,s)}{Av_{\#t=k}\,EWComb(t,s)}\ .$$