The Effect of Pooling and Evaluation Depth on Metric Stability

William Webber Alistair Moffat Justin Zobel

Department of Computer Science and Software Engineering
The University of Melbourne

The 3rd International Workshop on Evaluating Information Access

Australian Government
Australian Research Council

THE UNIVERSITY OF MELBOURNE

Overview

What effect do:

- Evaluation depth
- Assessment depth
- Normalization
- Choice of metric

have upon discriminative power in assessment?

Motivation

Moffat and Zobel designed the RBP metric.

RBP has nice mathematical properties.

RBP also has an intuitive, plausible user model.

RBP poor's discrimination

But studies showed RBP had poorer discrimination that AP, nDCG.

RBP and nDCG are very similar, rank-weighted metrics.

How RBP and nDCG differ

Main differences are:

- RBP is not normalized
- RBP weights decline smoothly, nDCG is steep–flat
- RBP typically not very deep

Hypotheses

nDCG is more discriminative because:

- of normalization
- because it validly makes use of more (deeper) relevance information
- because it is misled by evaluation beyond pooling depth

Evaluation and pooling depth

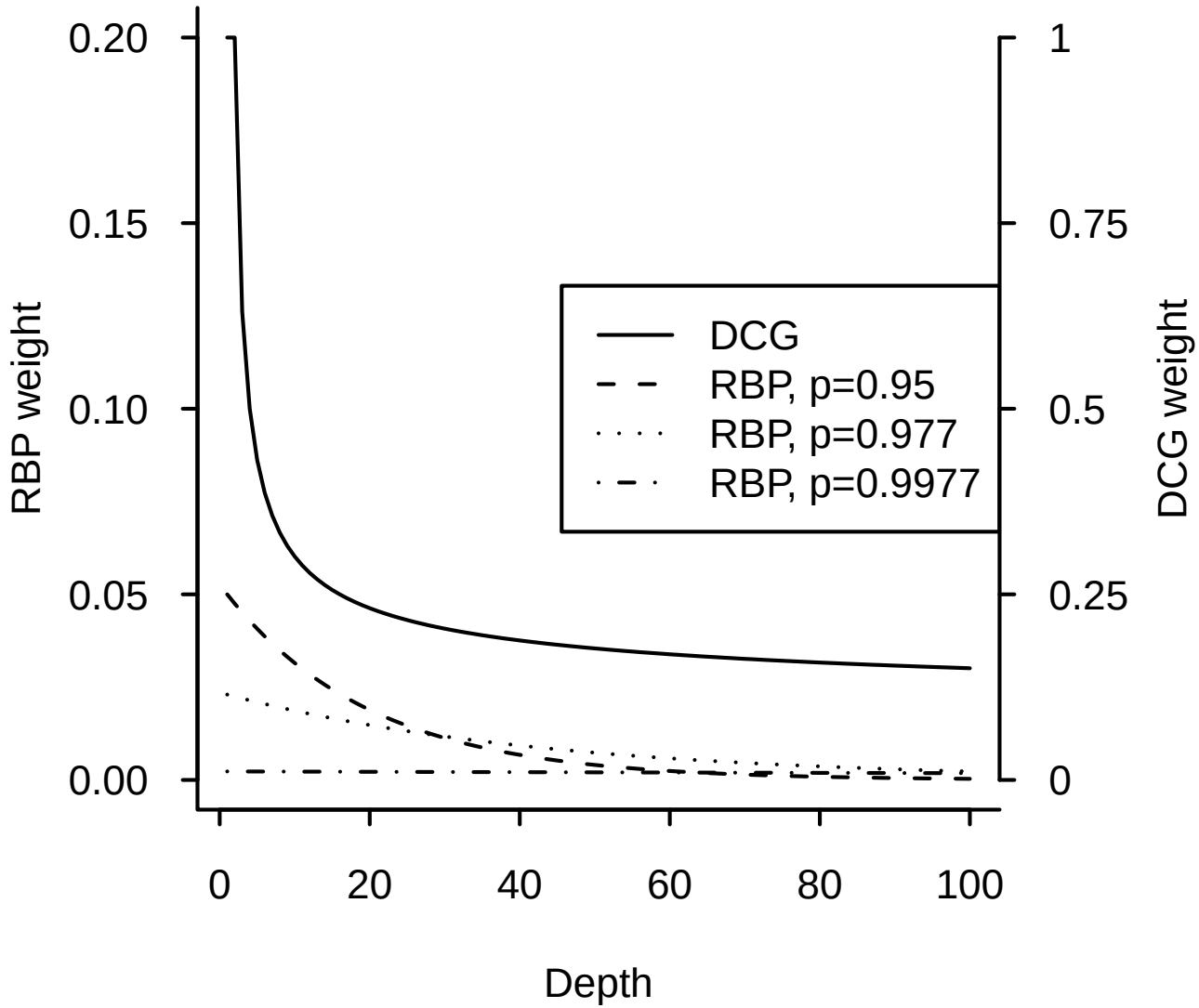
Allow that every system is pooled.

Evaluation frequently performed beyond pooling depth.

TREC: pool depth 100, evaluate depth 1,000.

82% of DCG's rank weight to depth 1,000 falls beyond depth 100.

Rank weights for nDCG, RBP



Discriminative power

For a metric.

Measured on a particular set of runs.

Proportion of run pairs whose difference in effectiveness is statistically significant.

Popularized by Sakai, using bootstrap. We use *t* test.

Discriminative power

Metric	T5 AH	T8 AH	T01 Web	T04 Rob	T05 TB	mean
P@10	0.628	0.645	0.594	0.516	0.555	0.588
RBP, p=0.8	0.638	0.657	0.602	0.517	0.562	0.595
RBP, p=0.95	0.661	0.691	0.627	0.598	0.658	0.647
AP@1000	0.638	**0.725**	0.627	**0.680**	0.748	0.683
nDCG@1000	**0.693**	0.718	**0.673**	0.673	**0.762**	**0.704**
mean	0.651	0.687	0.624	0.597	0.657	0.643

Table: Discriminative power of standard metrics on different TREC collections. The most discriminative metric for each collection is highlighted.

Metric similarity

	R@ 10	AP@ 10	nDCG @10	RBP .8	P@ 1k	R@ 1k	AP@ 1k	nDCG @1k	RBP .9977
P@10	0.88	0.90	0.94	0.93	**0.74**	0.69	0.83	0.83	0.80
R@10		0.90	0.86	0.86	0.71	**0.68**	0.83	0.82	0.77
AP@10			0.90	0.90	0.73	0.70	**0.86**	0.85	0.79
nDCG@10				0.98	0.71	0.66	0.80	**0.81**	0.78
RBP.8					0.72	0.67	0.81	0.81	**0.79**
P@1k						0.88	0.81	0.85	0.90
R@1k							0.79	0.84	0.82
AP@1k								0.91	0.88
nDCG@1k									0.91

Table: Kendall's τ between system rankings on the TREC 8 AdHoc track participant systems, using different metrics.

nDCG at depths

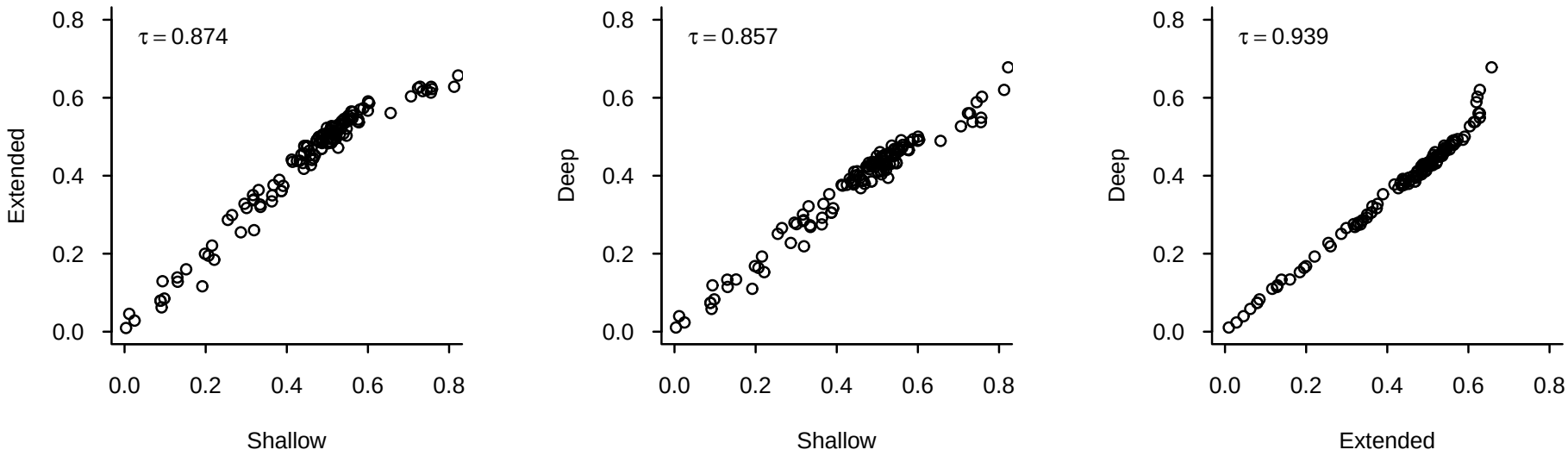
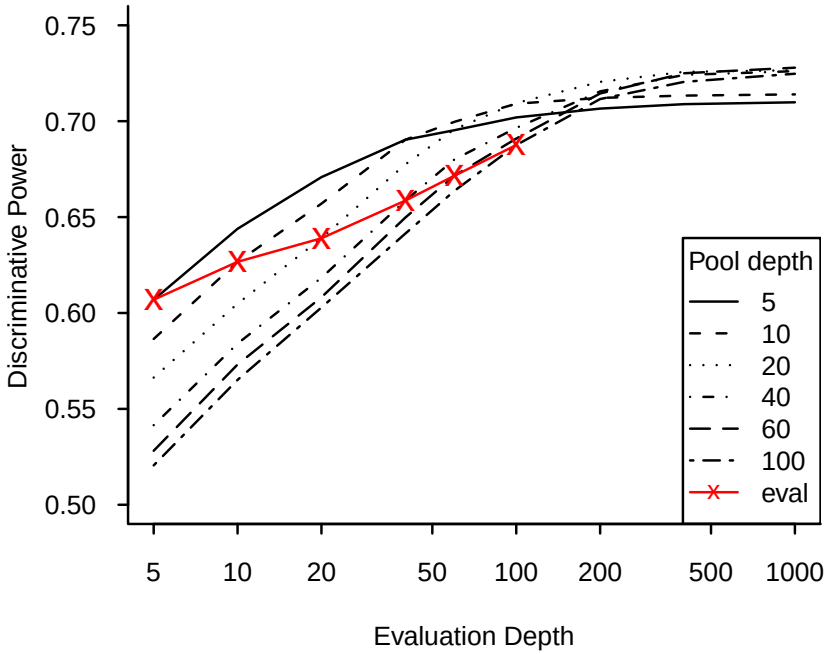
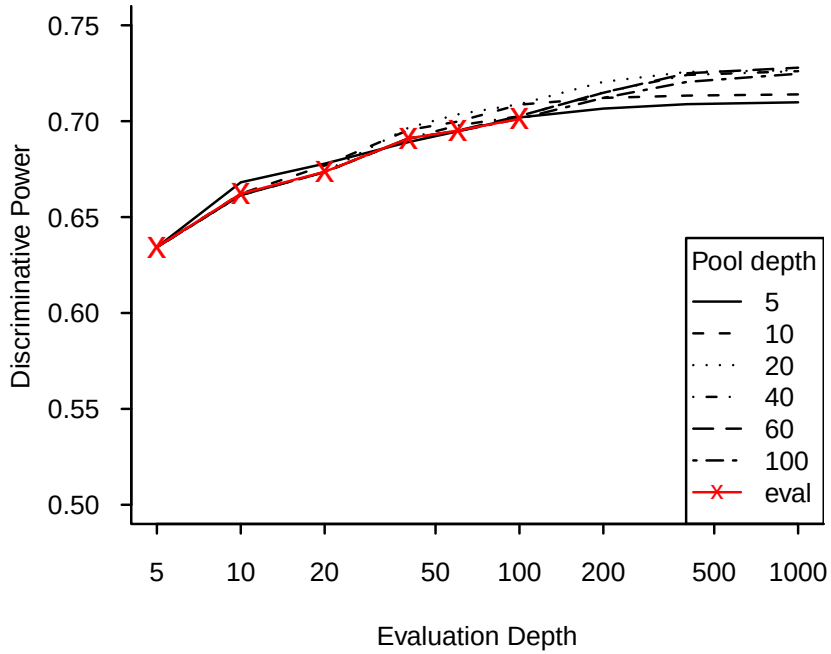


Figure: Relationship of system mean nDCG scores at different pooling and evaluation depths, for the TREC 8 AdHoc runset.

Cutoff depths for normalization



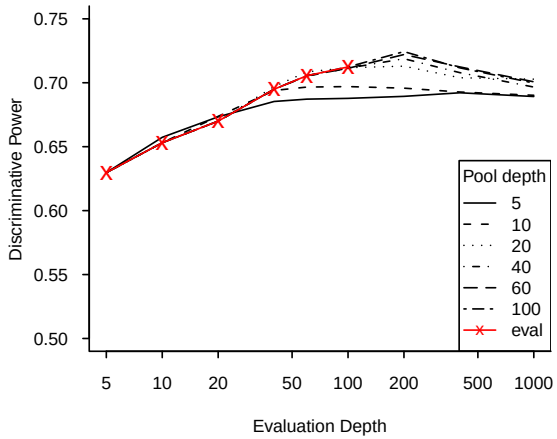
$\mathrm{AP} \equiv \mathrm{eAP}$



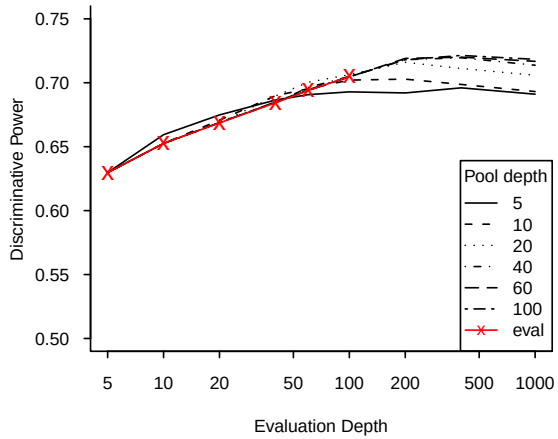
aAP

Figure: AP normalized by R versus AP normalized by $\max(k, R)$

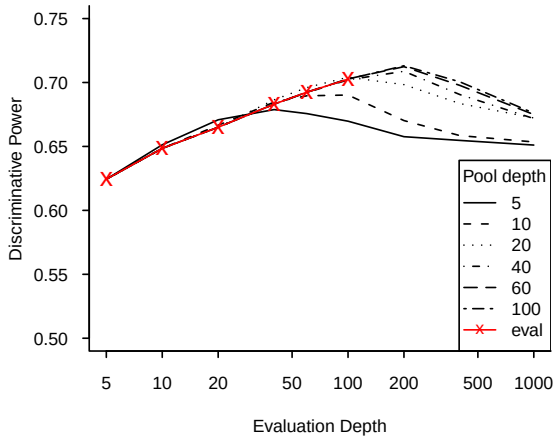
Cutoff, pooling, discrimination



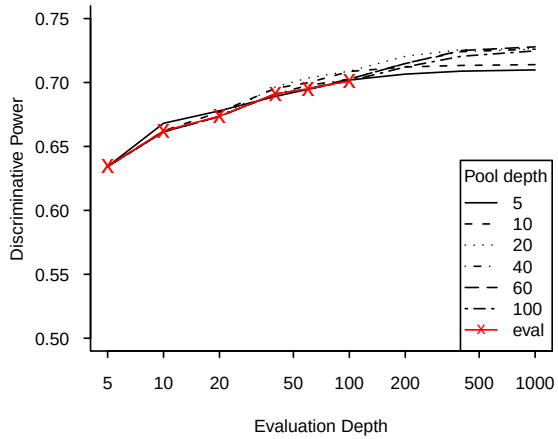
DCG



nDCG



RBP, resid $= 0.1$



aAP

Figure: Pooling and evaluation depth, and discriminative power

Correlation of significance

Pool	Eval	Metric	Pool@10 Eval@10		Pool@10 Eval@100			Pool@100 Eval@100		
			nDCG	RBP	aAP	nDCG	RBP	aAP	nDCG	RBP
10	10	aAP	0.89	0.88	***0.73***	*0.72*	*0.67*	***0.74***	*0.74*	*0.73*
		nDCG		0.96	*0.73*	***0.75***	*0.70*	*0.73*	***0.76***	*0.74*
		RBP			*0.72*	*0.74*	***0.69***	*0.72*	*0.75*	***0.73***
10	100	aAP				0.88	0.84	**0.86**	0.86	0.81
		nDCG					0.88	0.79	**0.88**	0.83
		RBP						0.75	0.81	**0.85**
100	100	aAP							0.87	0.82
		nDCG								0.88

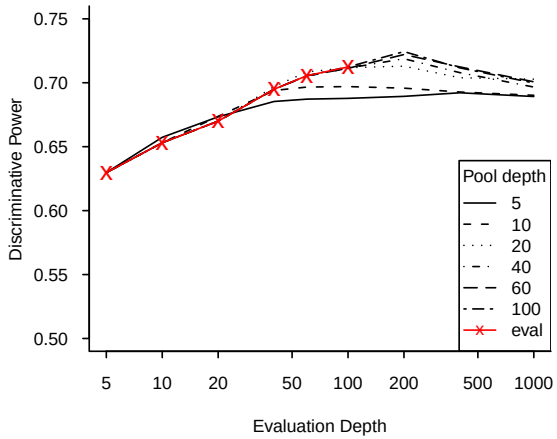
Table: Kendall's τ between p values assigned to TREC 8 AdHoc system pairs by paired, two-tailed t tests.

Recapitulating hypotheses

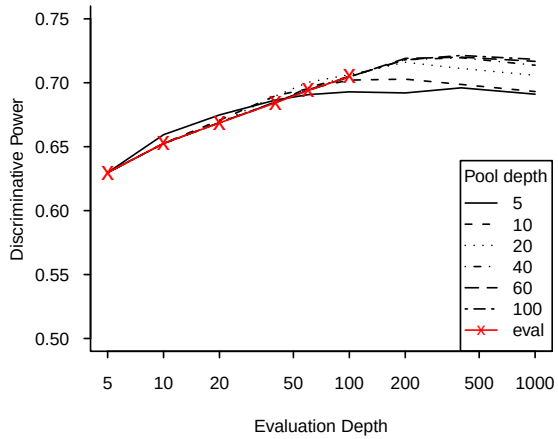
The original hypotheses:

- Normalization helps discriminative power
- Evaluating beyond pooling depth misleadingly helps discriminative power
- Greater evaluation depth helps discriminative power

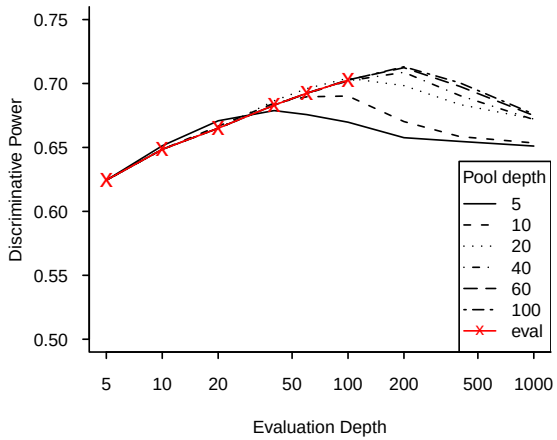
Normalization doesn't help discriminative power



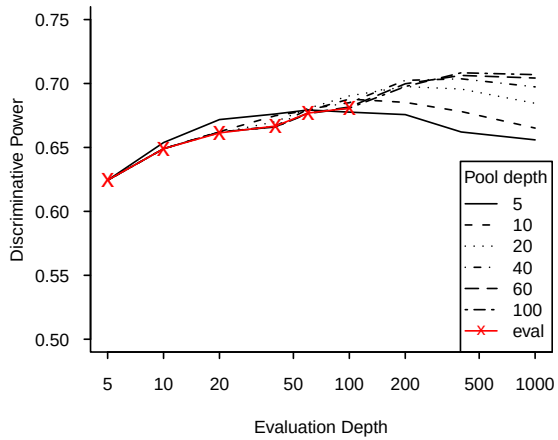
DCG



nDCG



RBP, resid $= 0.1$



nRBP, resid $= 0.1$

Figure: Pooling and evaluation depth, and discriminative power

Evaluation beyond pooling depth is not misleading

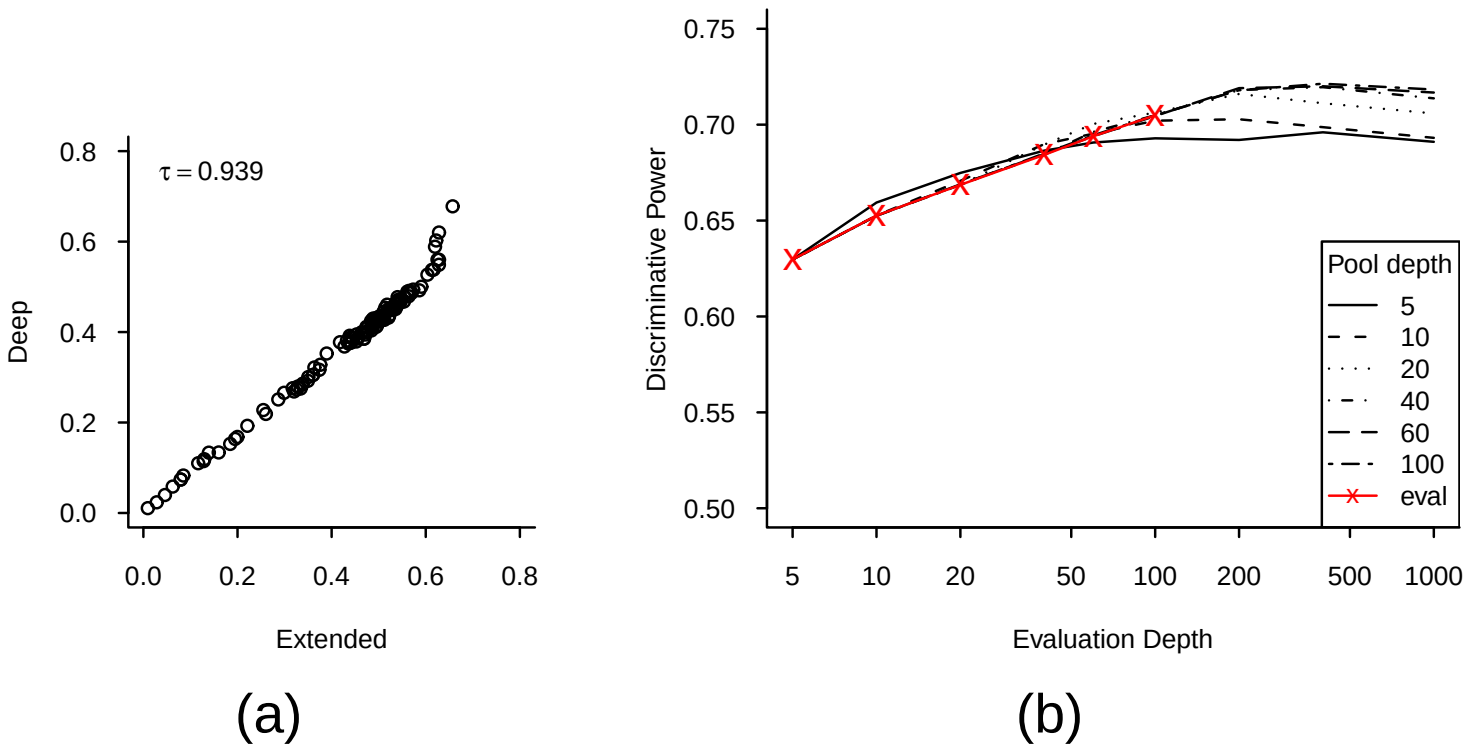
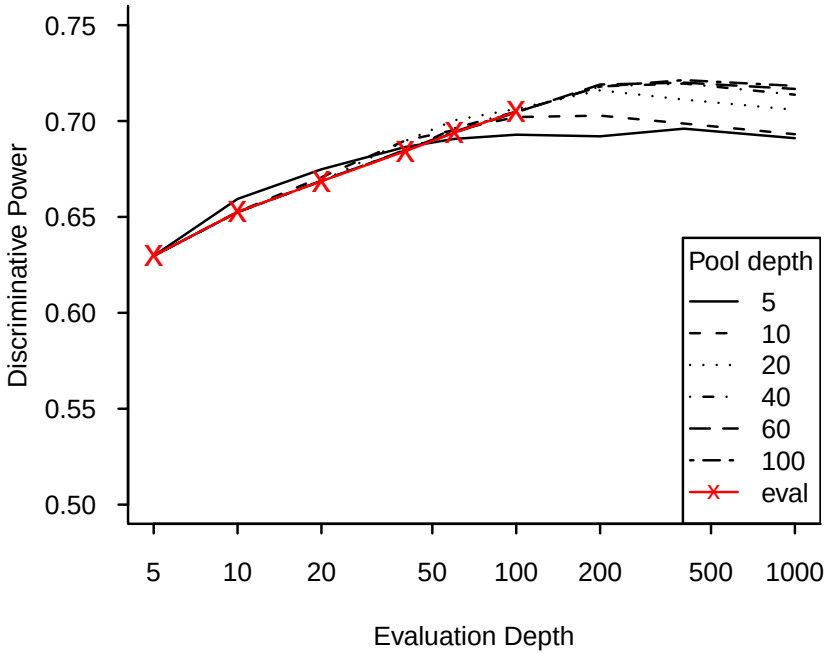
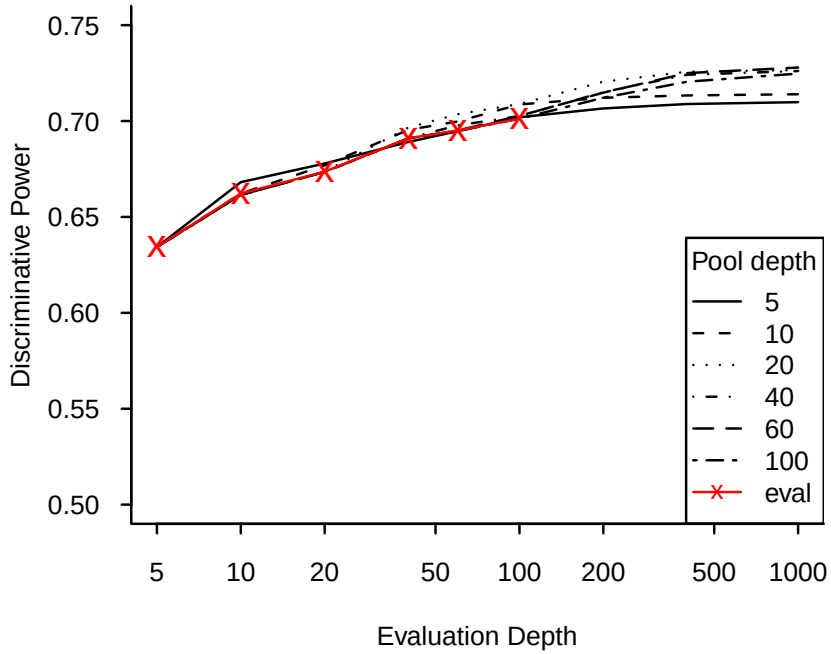


Figure: (a) short and full pooling similar scores; (b) short and full pooling similar discrimination

Deep evaluation picks up useful information



nDCG



aAP

Figure: Evaluation depth the most important, consistent determinant of discriminative power.

New hypothesis: DCG weights good

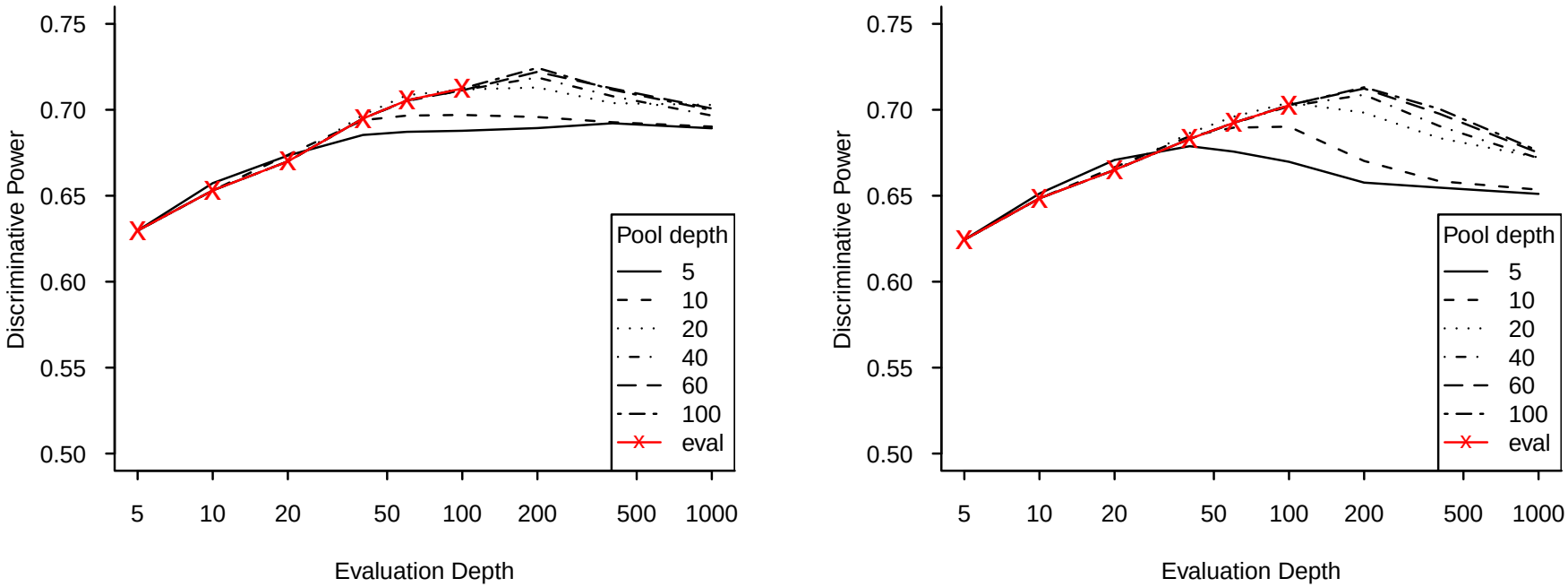


Figure: Effect of increasing evaluation depth on discriminative power

New hypothesis: DCG weights good

Deepening evaluation by raising RBP *p* harms discriminative power with short pooling.

Has little effect on DCG.

Steep–flat weighting of DCG may actually be (by chance?) well suited.

Questions