Why are Experts Correlated? Decomposing Correlations between Judges

Stephen B. Broomell & David V. Budescu

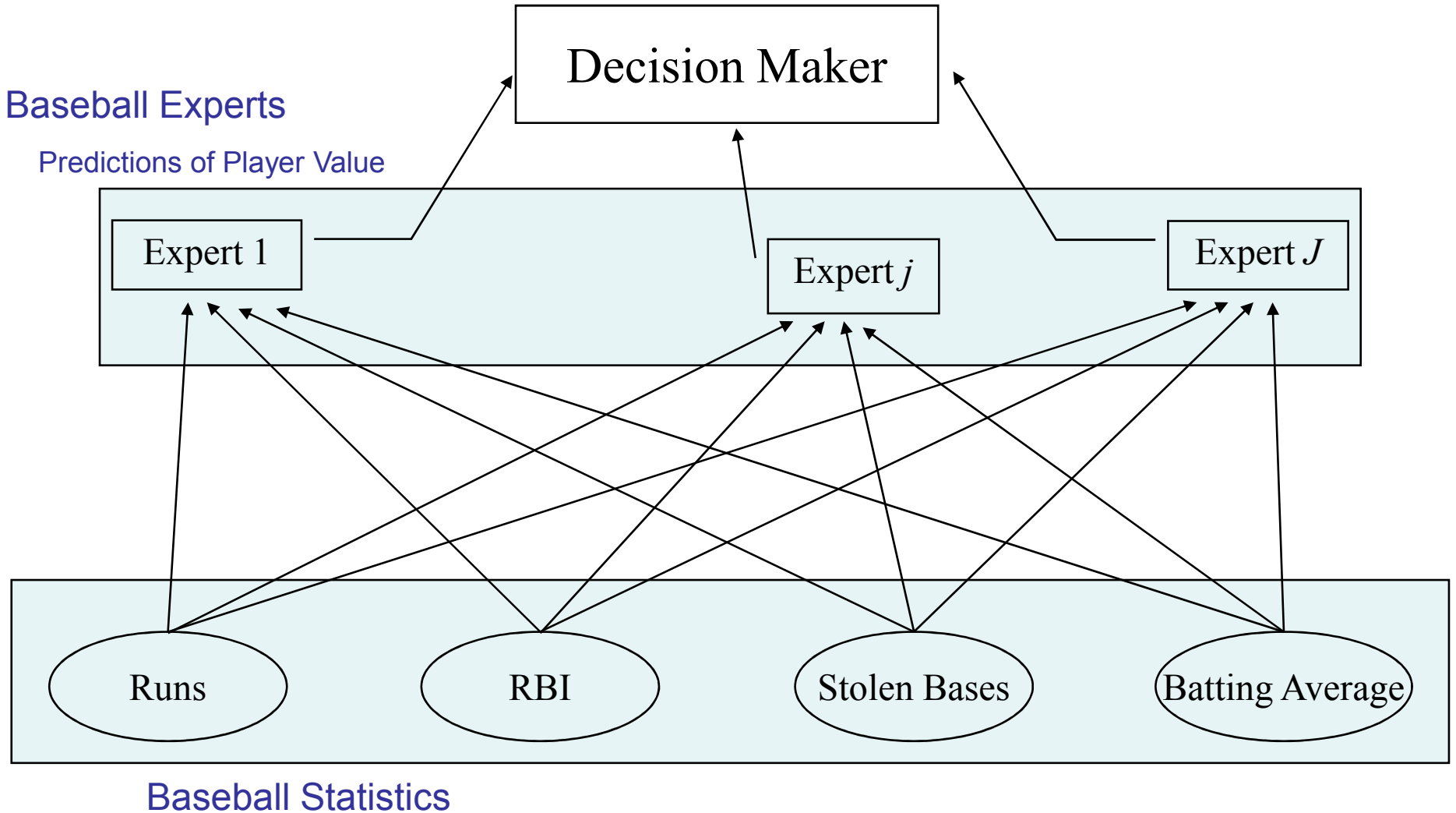
University of Illinois at Urbana-Champaign
Knowledge for All Lecture Series

November 2009

Outline

1. Introduction
   - Background
   - Underlying Factors
2. The Model
   - Symmetric Model
3. Predictions and Applications
   - Extreme Values
   - The Role of Validity
   - Judge vs. Cue
4. Confronting the Model with Data
   - Test of Symmetric Model
   - Impact of Player Names
5. Concluding Remarks

Baseball Example



Experts are Correlated

- Winkler (1971): Professional sportswriters predicting point spreads ($.73 \leq r \leq .97$)
- Ashton (1986): Executives and sales managers at TIME magazine (mean $r = .60$)
- Clemen and Winkler (1986): Econometric models predicting (a) nominal GNP and (b) real GNP (median $r = .91$)
- Winkler and Poses (1993): Physicians in an ICU predicting probabilities of survival for patients ($.76 \leq r \leq .79$)

Value of Additional Information

- The gains from adding experts are inversely related to the inter-correlations among the experts
  - Hogarth (1978)
- Highly redundant experts yield little additional information
  - Clemen & Winkler (1985)

Models of Expert Aggregation

- Clemen and Winkler (1985): Given J dependent judges, compute the equivalent number of independent judges n.

$$n(\sigma^2, \Sigma) = J[1 + (J-1)\rho]^{-1}$$

Models of Expert Aggregation

- Clemen and Winkler (1985): Given J dependent judges, compute the equivalent number of independent judges n.

$$n(\sigma^2, \Sigma) = J[1 + (J-1)\rho]^{-1} \Rightarrow n < J$$

Models of Expert Aggregation

- Clemen and Winkler (1985): Given J dependent judges, compute the equivalent number of independent judges n.

$$n(\sigma^2, \Sigma) = J[1 + (J-1)\rho]^{-1} \Rightarrow n < J$$

It takes fewer independent judges to achieve a certain accuracy.

Models of Expert Aggregation

- Clemen and Winkler (1985): Given J dependent judges, compute the equivalent number of independent judges n.

$$n(\sigma^2, \Sigma) = J[1 + (J-1)\rho]^{-1} \Rightarrow n < J$$

It takes fewer independent judges to achieve a certain accuracy.

- Hogarth (1978): Shows group validity as a function of mean individual validity and inter-judge correlation.

$$\rho_{y\overline{x}} = J^{1/2}\overline{\rho}_{yx}[1 + (J-1)\overline{\rho}_{x_i x_j}]^{-1/2}$$

Models of Expert Aggregation

- Clemen and Winkler (1985): Given J dependent judges, compute the equivalent number of independent judges n.

$$n(\sigma^2, \Sigma) = J[1 + (J-1)\rho]^{-1} \Rightarrow n < J$$

It takes fewer independent judges to achieve a certain accuracy.

- Hogarth (1978): Shows group validity as a function of mean individual validity and inter-judge correlation.

$$\rho_{y\overline{x}} = J^{1/2}\overline{\rho}_{yx}[1 + (J-1)\overline{\rho}_{x_i x_j}]^{-1/2} \Rightarrow \lim_{J\to\infty} \rho_{y\overline{x}} = \frac{\overline{\rho}_{yx}}{[\overline{\rho}_{x_i x_j}]^{1/2}}$$

Models of Expert Aggregation

- Clemen and Winkler (1985): Given J dependent judges, compute the equivalent number of independent judges n.

$$n(\sigma^2, \Sigma) = J[1 + (J-1)\rho]^{-1} \Rightarrow n < J$$

It takes fewer independent judges to achieve a certain accuracy.

- Hogarth (1978): Shows group validity as a function of mean individual validity and inter-judge correlation.

$$\rho_{y\overline{x}} = J^{1/2}\overline{\rho}_{yx}[1 + (J-1)\overline{\rho}_{x_i x_j}]^{-1/2} \Rightarrow \lim_{J\to\infty} \rho_{y\overline{x}} = \frac{\overline{\rho}_{yx}}{[\overline{\rho}_{x_i x_j}]^{1/2}}$$

Dependencies in the group detract from the group level validity

Models of Expert Aggregation

- Clemen and Winkler (1985): Given J dependent judges, compute the equivalent number of independent judges n.

$$n(\sigma^2, \Sigma) = J[1 + (J-1)\rho]^{-1} \Rightarrow n < J$$

It takes fewer independent judges to achieve a certain accuracy.

- Hogarth (1978): Shows group validity as a function of mean individual validity and inter-judge correlation.

$$\rho_{y\overline{x}} = J^{1/2}\overline{\rho}_{yx}[1 + (J-1)\overline{\rho}_{x_i x_j}]^{-1/2} \Rightarrow \lim_{J\to\infty} \rho_{y\overline{x}} = \frac{\overline{\rho}_{yx}}{[\overline{\rho}_{x_i x_j}]^{1/2}}$$

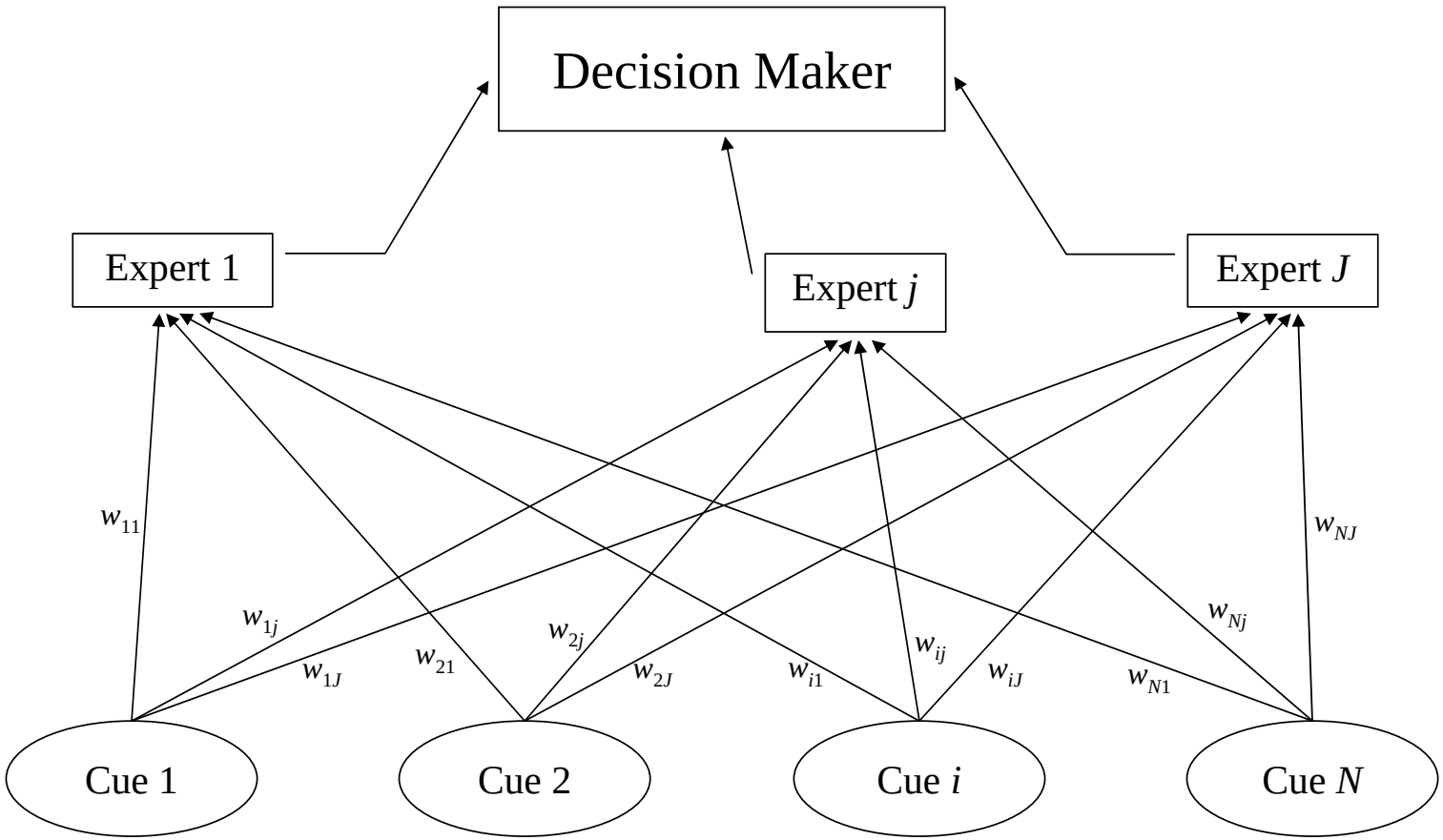
Dependencies in the group detract from the group level validity

- Wallsten et al. (1997) & Johnson, Budescu and Wallsten (2000): Demonstrate consistent results

Why is This Important?

- Wisdom of the crowd.
  - Independence is required for a crowd to be wise!
- How independent are information sources?
- Experts are exposed to extensive training
  - Will highly overlapping training reduce the value of additional expert opinions?

Representation of an "Expert Opinion"



Parameter Description

Assumptions about Nature	
ρ_c	The average correlation between naturally occurring information sources used by experts (medical tests, stock performance, etc.).
N	The amount of total information (cues) available in a given decision task.
Assumptions about Experts	
σ_w^2	The variance of the weights: A measure of the experts differentiation between cues.
ρ_w	The correlation between the weights assigned by the experts' to the various cues.
δ	Noise-to-signal ratio A measure of the unreliability of the expert.

A Model of Inter-Judge Correlation

Underlying Assumptions

- Let an "opinion", x_j, be defined as a weighted sum of information cues.

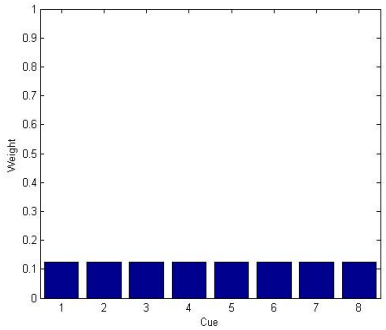
$$x_j = \sum_{i=1}^{N} w_{ij} c_i + e_{ij} \quad where \quad e_{ij} \sim (0, \sigma^2)$$

- Weights are non-negative and normalized to sum to 1.

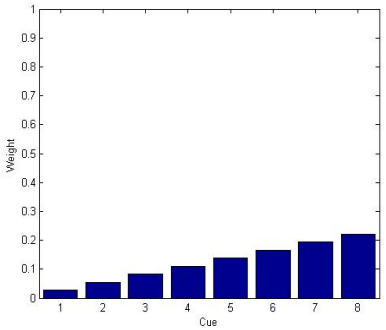
$$\sum_{i=1}^{N} w_{ij} = 1 \quad for \quad j = 1, \ldots, J$$

Weighting Examples ($N = 8$ cues)

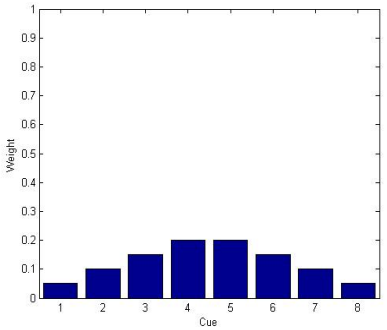
Uniform Weighting



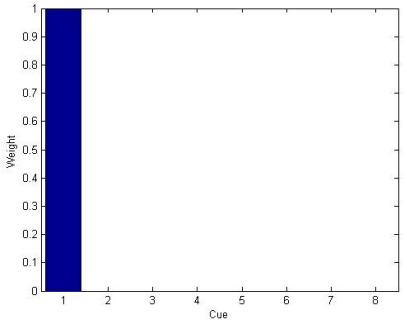
Proportional to Ranks



Triangular



Single



Each weighting pattern has a variance
Each pair of weighting patterns has a covariance

Derivation

Consider a correlation

$$\rho_{xy} = \frac{\sigma_{x,y}}{\sqrt{\sigma_x^2}\sqrt{\sigma_y^2}}$$

- We can define the variance and covariance between two judges' opinions (x_j and $x_{j'}$)

$$\sigma_{x_j}^2 = \underline{w}_j^T \mathbf{\Sigma} \underline{w}_j + \sigma_{e_j}^2$$

$$\sigma_{x_j x_{j'}} = \underline{w}_j^T \mathbf{\Sigma} \underline{w}_{j'}$$

- Thus, we can compute the correlation between two judges, $\rho_{x_j x_{j'}}$

Inter-Judge Correlation

The Model of Inter-Judge Correlation

Both judges are restricted to have the same cue differentiation ($\sigma_w^2 = \sigma_{w'}^2$) and accuracy ($\delta = \delta'$).

$$\rho_{x_j x_{j'}} = f(\rho_c, \rho_w, \sigma_w^2, \delta, N)$$

$$\rho_{x_j x_{j'}} = \frac{\rho_c + (1-\rho_c)[(N-1)\rho_w \sigma_w^2 + \frac{1}{N}]}{[\rho_c + (1-\rho_c)((N-1)\sigma_w^2 + \frac{1}{N})](1+\delta)}$$

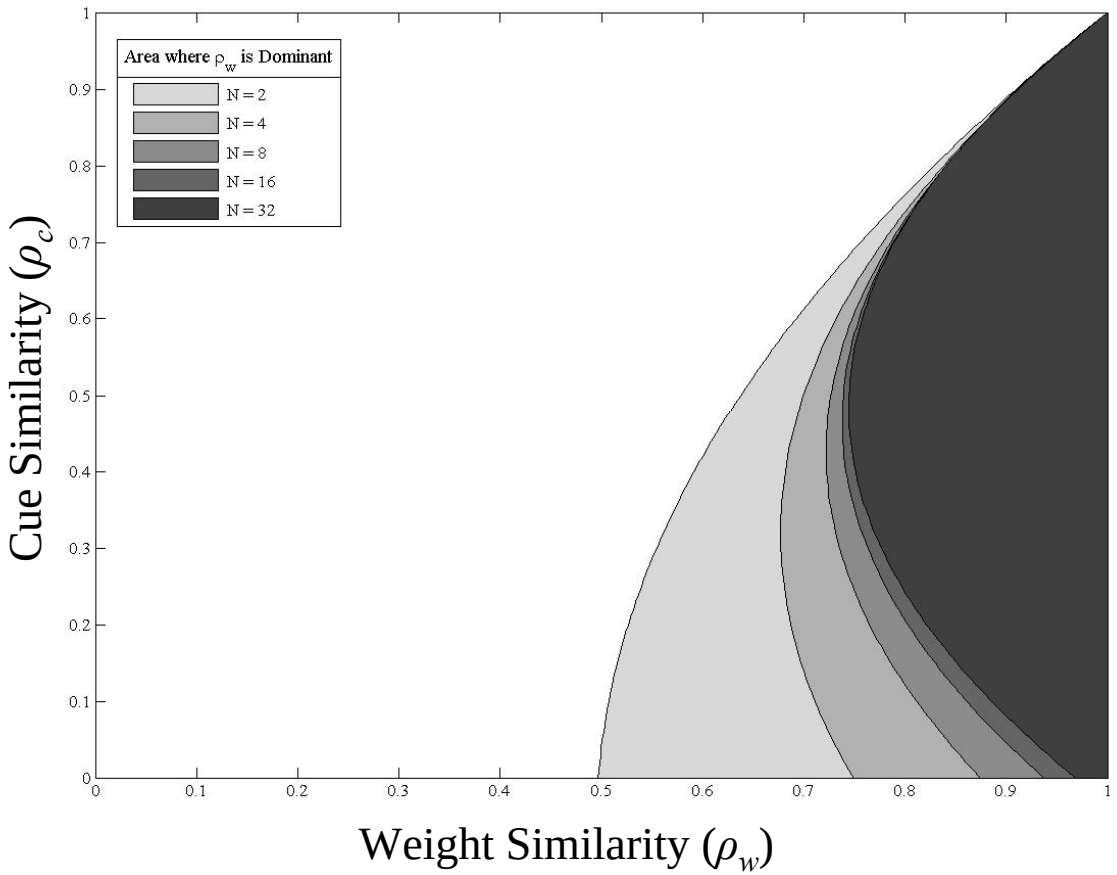
Underlying Factors

Marginal Effects

- Inter-judge correlation increases as a function of
  - number of cues,
  - inter-cue correlations, and
  - inter-weight correlation
- Inter-judge correlation decreases as a function of
  - variance of weighting scheme, and
  - unreliability of judges

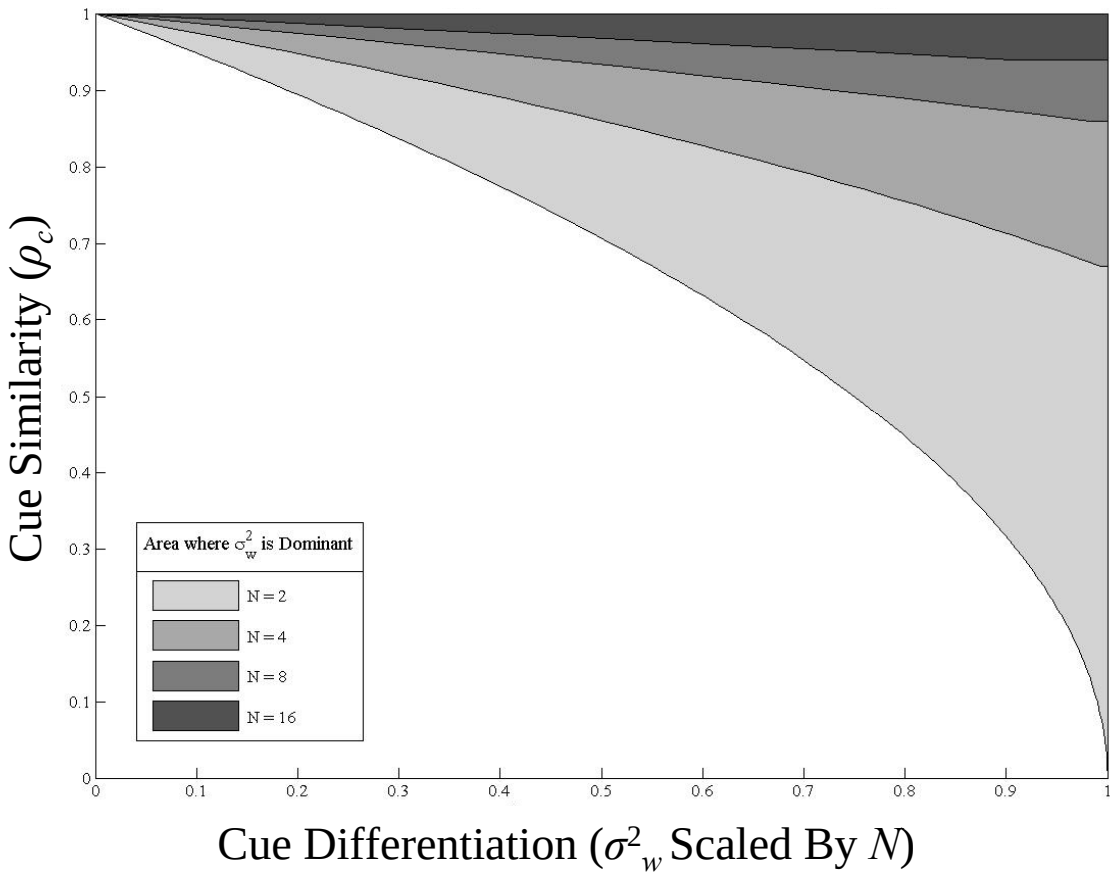
Relative Importance

Iso-Influence Curves for ρ_c vs. ρ_w



Relative Importance

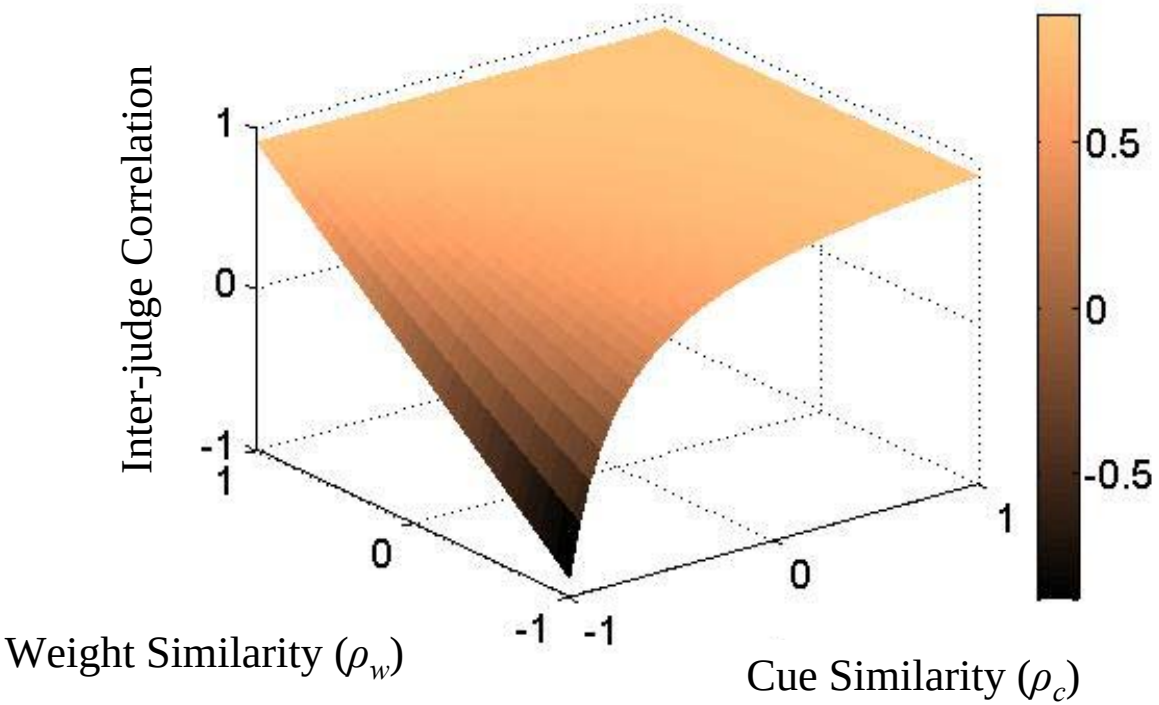
Iso-Influence Curves ρ_c vs. σ_w^2 (Scaled by N)



Underlying Parameters

Inter-weight Correlation vs. Inter-Cue Correlation ($N = 2$)

$\delta = 1/9$ $\sigma^2_w = 0.06$ (Ranked Weights)

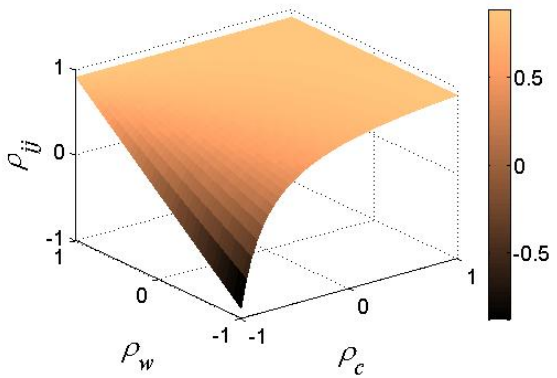


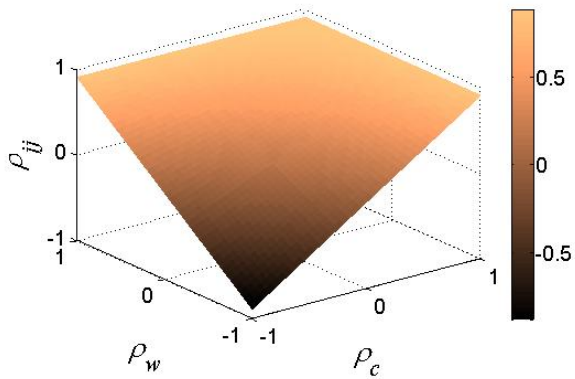
Underlying Parameters

Inter-weight Correlation vs. Inter-Cue Correlation ($N = 2$)

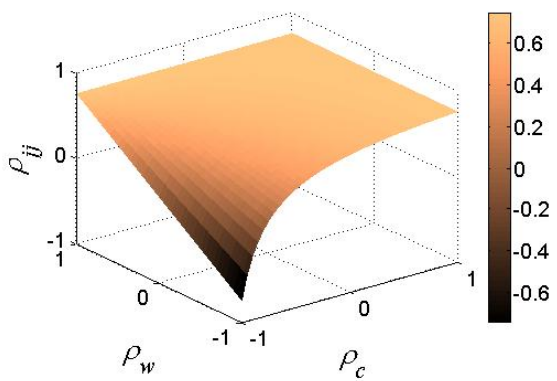
$\sigma^2_w = 0.06$ (Ranked Weights) $\sigma^2_w = 0.5$ (Single Weight)

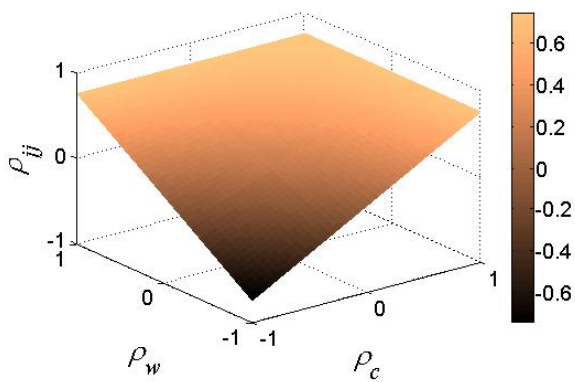
$\delta = 1/9$





$\delta = 1/3$

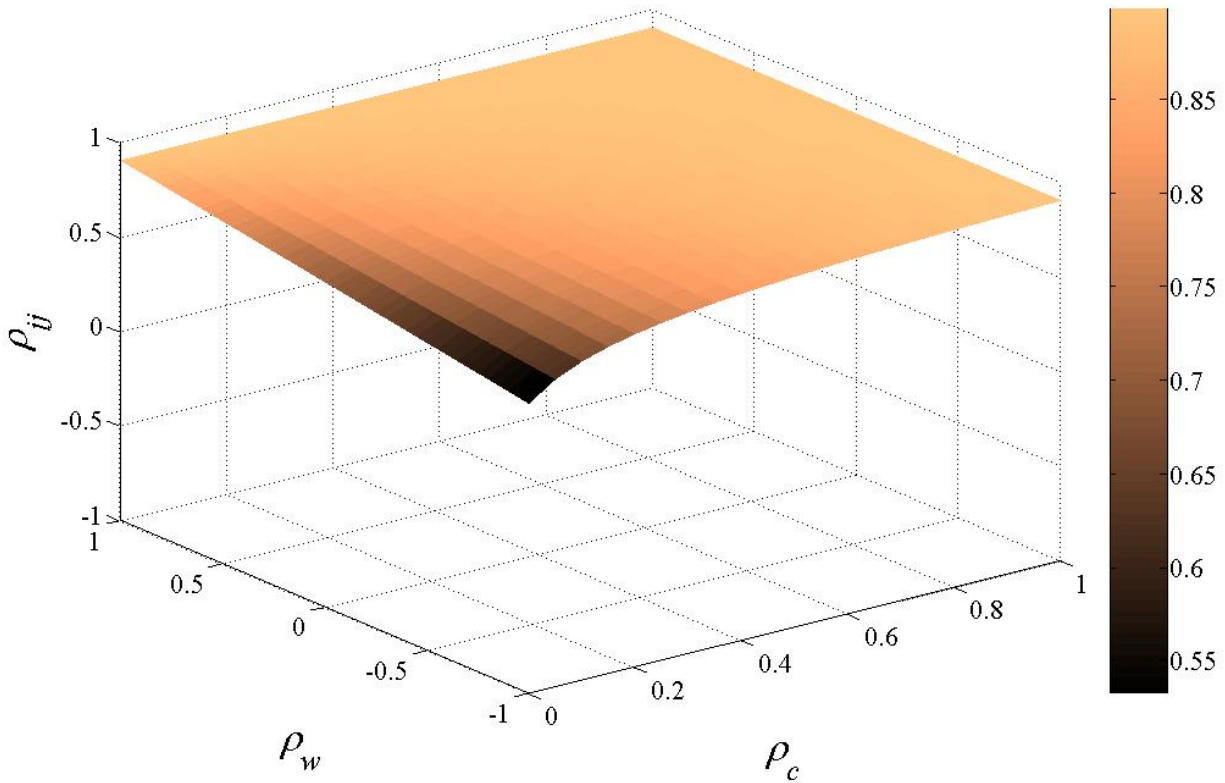




Underlying Parameters

Inter-weight Correlation vs. Inter-Cue Correlation ($N = 8$)

$\delta = 1/9$ $\sigma^2_w = 0.005$ (Ranked)



Predictions and Applications

When are Judges Uncorrelated?

Uncorrelated Judges

$$\rho_c = 0 \quad \& \quad \sigma_{w_j w_{j'}} = \frac{-1}{N(N-1)}$$

- Requires two non-overlapping sets of uncorrelated cues.

When are Judges Perfectly Correlated?

Perfectly Correlated Judges

- A necessary condition for a perfect correlation is the absence of noise $(\delta = 0)$
- Case 1: $\rho_w = 1$
- Case 2: $\rho_c = 1$

Partialling out Validity

Assumptions:

- Let ρ_{yc} be the average cue validity.
- It follows that judge validity is defined as:

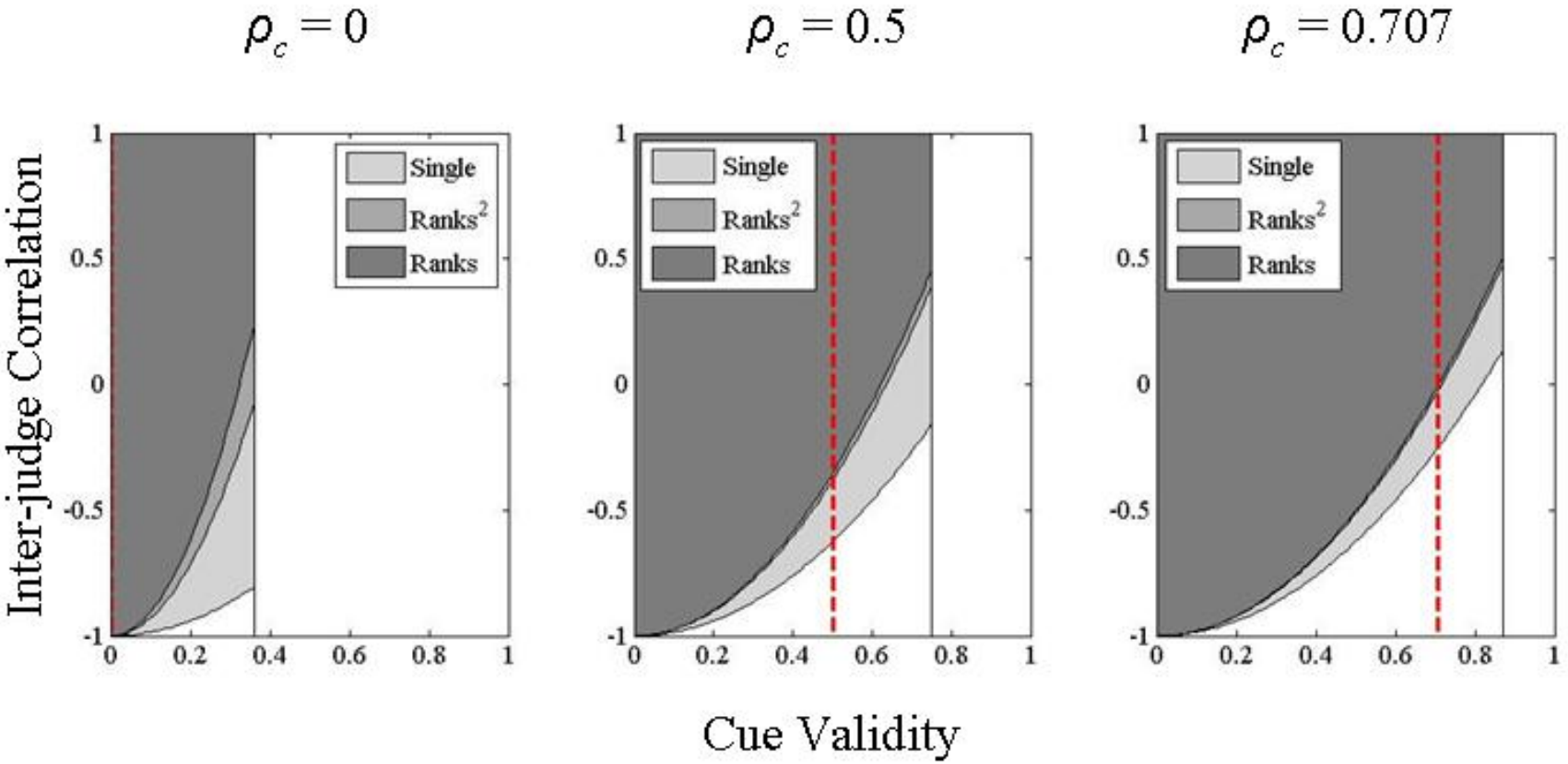
$$\rho_{x_j y} = \frac{\rho_{yc}}{\sqrt{\sigma^2_{x_j}}}$$

- The restriction of range on inter-judge correlation is:

$$\frac{2\rho^2_{yc} - \sigma^2_{x_j}}{\sigma^2_{x_j}} \leq \rho_{x_j x_{j'}} \leq 1$$

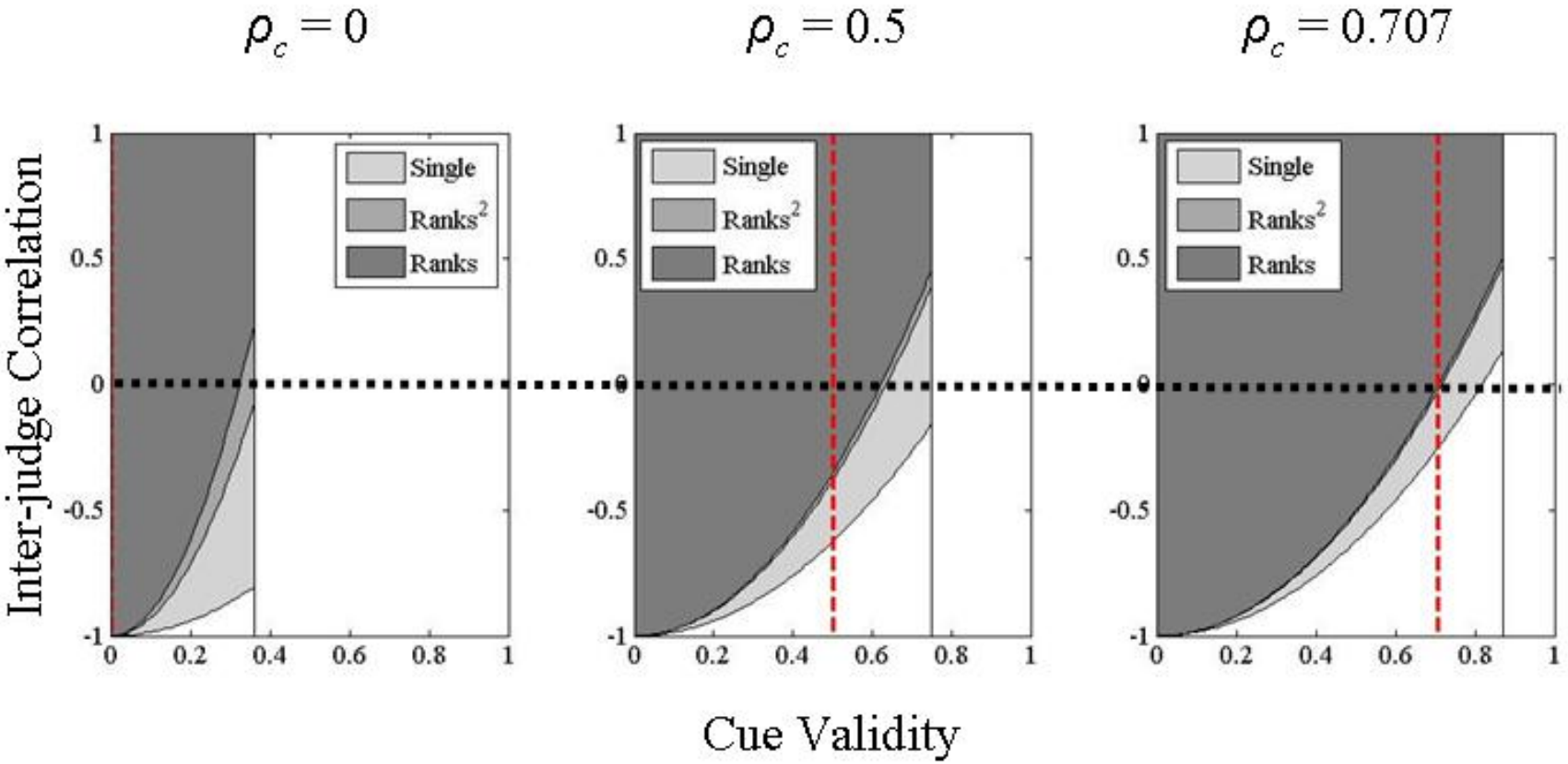
Range Restriction

Range of Inter-judge Correlation vs. Cue Validity



Range Restriction

Range of Inter-judge Correlation vs. Cue Validity



Additional Judge vs. Additional Cue

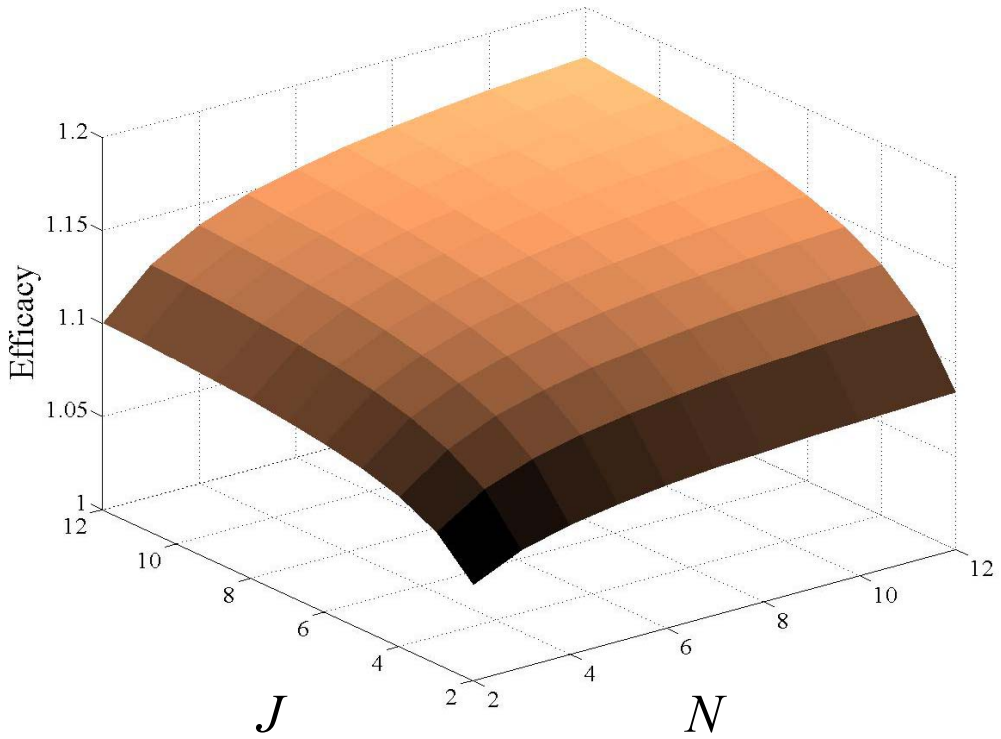
- A policy maker has access to J judges and N cues.
- Needs to decide whether to invest in another judge or another cue.

$$\text{Efficacy} = \frac{\rho_{y\overline{x}}}{\overline{\rho}_{yx}} = J^{1/2}[1 + (J-1)\overline{\rho}_{x_i x_j}]^{-1/2}$$

- Let $\overline{\rho}_{x_i x_j} = f(\rho_c, \rho_w, \sigma_w^2, \delta, N)$

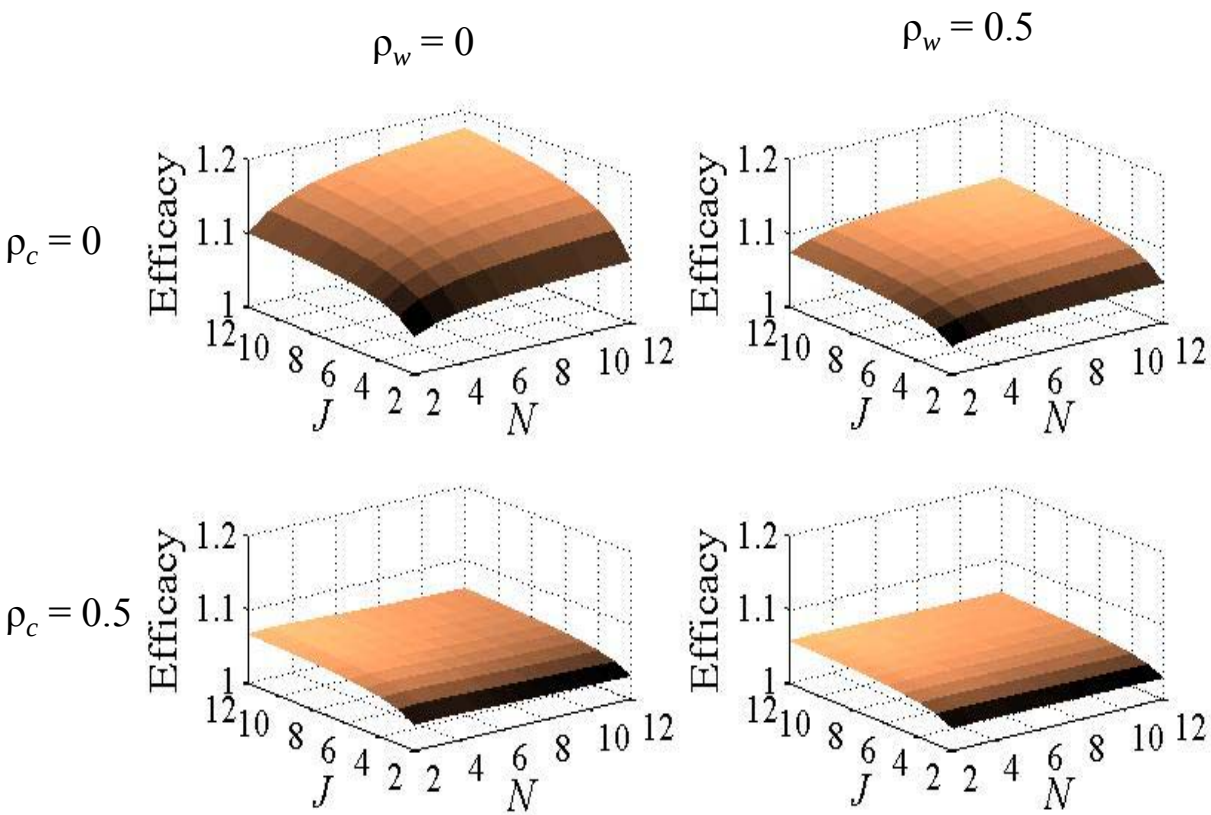
Additional Judge vs. Additional Cue

$\rho_c = 0$ $\rho_w = 0$



$$\text{Efficacy} = \frac{\rho_{y\overline{x}}}{\overline{\rho}_{yx}} = J^{1/2}[1 + (J-1)\overline{\rho}_{x_i x_j}]^{-1/2}$$

Additional Judge vs. Additional Cue



$$\text{Efficacy} = \frac{\rho_{y\overline{x}}}{\overline{\rho}_{yx}} = J^{1/2}[1 + (J-1)\overline{\rho}_{x_i x_j}]^{-1/2}$$

Additional Judge vs. Additional Cue

			$\rho_w = 0$					$\rho_w = 0.5$				
			N					N				
			2	3	4	5		2	3	4	5
ρ_c=0	J	2	-0.07	-0.09	-0.08	-0.07	J	2	-0.14	-0.12	-0.10	-0.09
		3	**0.05**	-0.01	-0.02	-0.03		3	**0.01**	-0.04	-0.04	-0.04
		4	**0.09**	**0.01**	0.00	-0.01		4	**0.06**	0.00	-0.01	-0.02
		5	**0.11**	**0.03**	0.00	0.00		5	**0.09**	**0.02**	0.00	-0.01
			N					N				
			2	3	4	5		2	3	4	5
ρ_c=0.5	J	2	0.21	0.20	0.21	0.21	J	2	0.06	0.09	0.10	0.11
		3	0.39	0.35	0.35	0.35		3	0.30	0.28	0.28	0.27
		4	0.45	0.41	0.40	0.40		4	0.40	0.36	0.36	0.35
		5	0.48	0.44	0.43	0.43		5	0.45	0.41	0.40	0.39

$$\text{Efficacy}\{(J+1),\ N,\ \overline{\rho}_c\} = \text{Efficacy}\{J,\ (N+1),\ \overline{\rho}_c*\}$$

Confronting the Model with Data

Baseball Experiment: Design

- Participants: 16 self-proclaimed baseball experts
- Expertise: Established in a previous study (Miller, 2008)
- Task: Predict the value of MLB players based on selected statistical information (cues)
  - Groups 1-3: Statistics for Batters (e.g., Runs, Batting Average, ...)
  - Group 4: Statistics for Pitchers (e.g., Strikes, WHIP, ...)
- Design: Inter-cue correlation was manipulated
  - Group 1: $N = 5; \rho_c = 0.61$
  - Group 2: $N = 4; \rho_c = 0.49$
  - Group 3: $N = 4; \rho_c = 0.74$
  - Group 4: $N = 5; \rho_c = 0.27$
- Predictions were generated for 3 sessions

Baseball Experiment: Design

- Part 1: Test of inter-judge correlation model
  - Reliability: Session 1 and 2 had no player names, and 20 repeats
  - Cues: ρ_c computed as average of correlation between baseball stats
  - Weights: Inferred from predictions (linear regression weights)
- Part 2: Impact of knowledge of player names
  - Session 3 included player names
  - Predictions for 30 players with and without names
- Part 3: Expert perception of prediction environment
  - Experts estimated the correlation between each pair of cues
  - Experts estimated the weight given to each cue

Part 1: Test of Inter-Judge Correlation Model

The Model's Prediction of the Inter-Judge Correlation:
The Baseball Example

Group	Design Parameters		Mean Inter-Judge Correlation		
	N	ρ_c	Predicted	Observed	Difference
1	5	0.61	0.72	0.66	0.05
2	4	0.49	0.85	0.77	0.08
3	4	0.74	0.92	0.88	0.05
4	5	0.27	0.60	0.61	-0.02

Part 1: Test of Inter-Judge Correlation Model

The Model's Prediction of Efficacy in the Baseball Example

Group	Design Parameters		Mean Efficacy		
	N	J	Predicted	Observed	Difference
1	5	4	1.12	1.16	0.04
2	4	5	1.07	1.11	0.04
3	4	4	1.03	1.05	0.02
4	5	3	1.17	1.10	-0.06

Part 2: Impact of Knowledge of Player Names

Does knowledge of player names impact prediction?

- Experts tend to over-estimate player value
- Knowing player names leads to...
  - higher evaluation for valuable and recognizable players
  - lower evaluation for less valuable players

Group Level

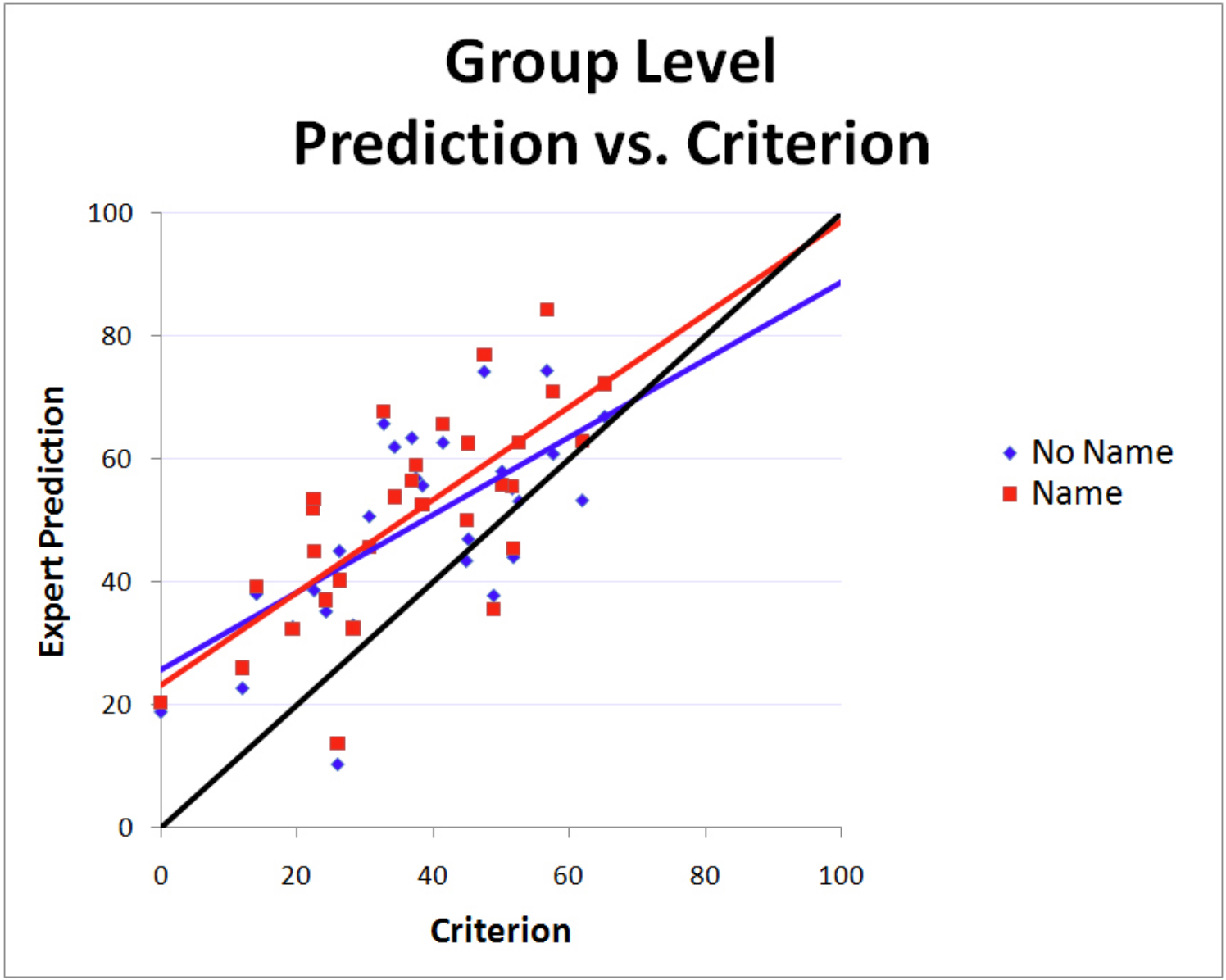
Criterion	Higher Name	Higher No Name	Total
High	12	3	15
Low	8	7	15
Total	20	10	30

Chi-Square $= 2.4$, $p = 0.12$
(Fisher's Exact Test, $p = 0.25$)

Individual Level

Criterion	Higher Name	Higher No Name	Total
High	128	67	195
Low	85	110	195
Total	213	177	390

Chi-Square $= 19.13$, $p < 0.0001$


Group Level
Prediction vs. Criterion
Expert Prediction
Criterion
0
20
40
60
80
100
No Name
Name

Part 2: Impact of Knowledge of Player Names

Tucker's (1964) Formulation of the Lens Model

$$r_a = GR_sR_e + C\sqrt{1 - R_s^2}\sqrt{1 - R_e^2}$$

r_a	$=$	Achievement:	Correlation between judgment and criterion
G	$=$	Knowledge:	Correlation between the linear models of the judge and the environment
R_s	$=$	Environmental Predictability:	Multiple correlation between the environment and the cues
R_e	$=$	Participant's Consistency:	Multiple correlation between the cues and the judgment
C	$=$	Unmodeled knowledge:	Correlation between the residuals of both models

Part 2: Impact of Knowledge of Player Names

Lens Model Parameters

Mean Parameter Estimates

	r_a	R_s	R_e	G	C
No Name	0.67	0.89	0.78	0.93	0.11
Name	0.63	0.88	0.79	0.85	0.12
Signed Rank	36.5	-6	-15	53.5	-3.5
p	0.059	0.78	0.19	0.0009	0.87

- We find a significant decrease for G with the inclusion of player names by sign rank test, $p < .001$.
- The remaining parameters are not affected by knowledge of player names

Part 3: Expert Perception of Prediction Environment

Do experts understand the correlation structure underlying the cues?

- Experts assigned a number between -10 and 10 to each pairwise combination of cues (responses scaled to -1 and 1)
- Experts rank ordered correctly the various pairs of cues, i.e., identify which are more highly correlated (median $\tau = 0.52$)
- Experts tended to underestimate the magnitude of cue correlations (sign test, $p < .001$)

Part 3: Expert Perception of Prediction Environment

Do experts understand their own weighting structure?

- Experts allocated 100 points to the available cues
- Subjective weights were correlated with the estimated weights (median $r = 0.37, \tau = 0.18$)
- Subjective weights tended to be less differentiated than the inferred weights
  $\sigma^2_{subjective} < \sigma^2_{regression}$

Conclusions (1)

The Model of Inter-Judge Correlation

Role of Parameters

- Increase Inter-judge Correlation
  - Cue Similarity (ρ_c)
  - Number of Cues (N)
  - Weight Similarity (ρ_w)
- Decrease inter-judge correlation
  - Cue Differentiation (σ_w^2)
  - Unreliability (δ)
- Cue Similarity has more influence on the inter-judge correlation than weight similarity and cue differentiation

Conclusions (2)

The Model of Inter-Judge Correlation

Applications

- Restrictions imposed by validity do not explain empirical findings
- Additional judges increase efficacy at a greater rate than do additional cues
  - The reverse can be true for certain cues selected to reduce inter-cue correlation.
- Analysis of real experts confirms the predictive ability of the model.
- Assumptions underlying the model should always be considered before application.

Conclusions (3)

Summary of Data Analysis

- The impact of player names is minimal
  - Knowledge of names influences over (under) estimation for (un)recognizable players
  - Lens model parameters are stable (including validity)
  - Weighting patterns are highly similar
- Analysis of expert perception of the environment (cues) reveals ...
  - Accurate representation of rank order cue dependencies
  - Experts underestimate the magnitude of inter-cue correlation
- Analysis of expert perception of decision weights reveals ...
  - Similarity between subjective and inferred weights
  - Results suggest that experts may underestimate cue differentiation

Thank You

Thank you!