Effect of Within-Category Cue-Cue Correlations on the Accuracy of Relative Weight Measures in Logistic Lens Model Analyses

Robert M. Hamm,[1]

Esther Kaufmann,[2] Emmanuel Bottieau,[3] Jef Van den Ende[3]

(1) University of Oklahoma Health Sciences Center, Oklahoma City, OK, USA

(2) University of Zurich, Switzerland

(3) Institute of Tropical Medicine, Antwerp, Belgium

Overview

- Review Brunswik Lens Model
  - Principle – compare judgment model to environment model
  - Lens Model Equations
    - Linear and Logistic
- Specify problems due to cue-cue correlations
  - Characterize effect of changing category proportion in stimulus set
  - Show predictive formulas
- Observe cue-cue correlations occurring in an environment – diseases causing fever in a clinic.
  - Assess effect of changing category proportion.

The Brunswik Lens Model

- Model the two sides, the world and the person's thinking
  - In common: the observable "proximal" cues
  - Criterion and judgment
  - Same general form of model

    Criterion = f(cues)

    Judgment = f(cues)

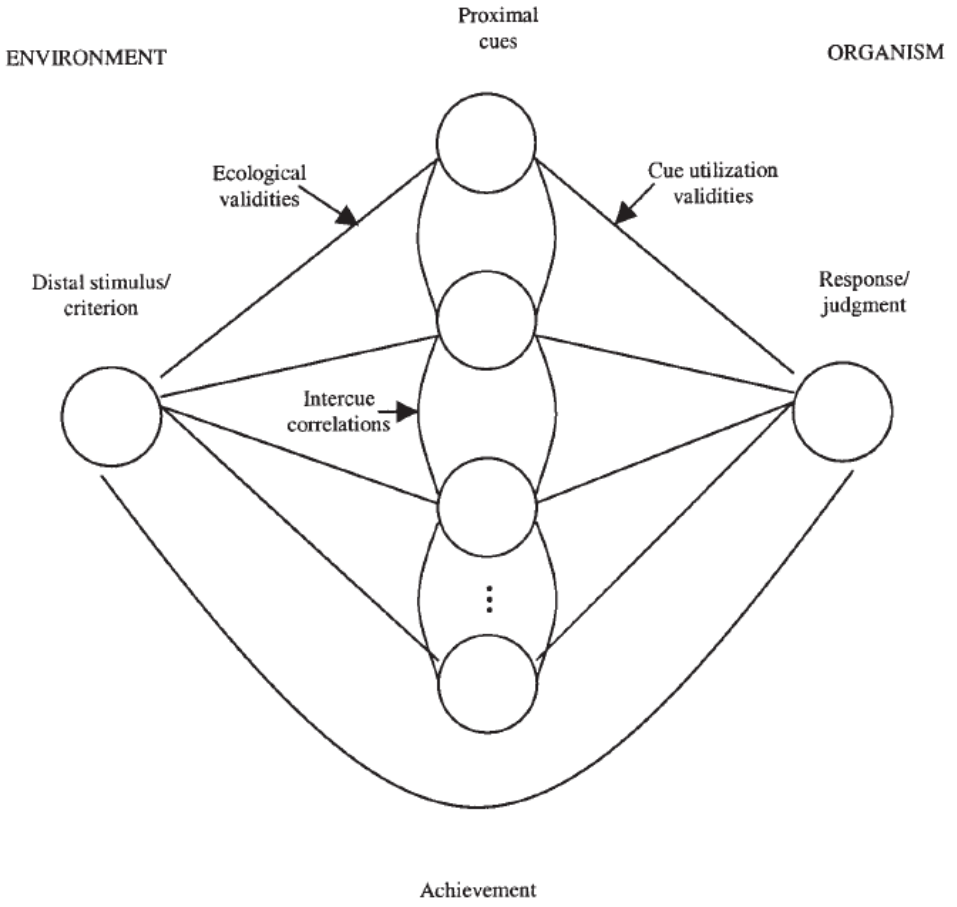


Figure 1. Adapted lens model. From "The conceptual framework of psychology." In *International Encyclopedia of Unified Science* (p. 678), by E. Brunswik, 1952, Chicago: University of Chicago Press. Copyright 1952 by the University of Chicago Press.

Regression Model Equations

- Linear regression model for continuous dependent variable (criterion or judgment).
  - $Y_e = B_0 + B_1X_1 + B_2X_2 + B_3X_3$
- Logistic regression model for categorical dependent variable.
  - Categories happen with cutoffs on a general dimension, for category membership, e.g., “at risk” or not.
  - Or when the categories are different things, like animals or diseases. No camelephants.
  - $Ln\left(\frac{p(Y_e)}{1-p(Y_e)}\right) = B_0 + B_1X_1 + B_2X_2 + B_3X_3$
  - $\frac{p(Y_e)}{1-p(Y_e)} = e^{B_0 + B_1X_1 + B_2X_2 + B_3X_3}$
  - $p(Y_e) = \frac{e^{B_0 + B_1X_1 + B_2X_2 + B_3X_3}}{1+e^{B_0 + B_1X_1 + B_2X_2 + B_3X_3}} = \frac{1}{1+e^{-(B_0 + B_1X_1 + B_2X_2 + B_3X_3)}}$

Lens Model Formulas

- When the regression models have the same form, except for parameters, the judgmental achievement can be "decomposed" into meaningful components.
  - Linear Lens Model Equation

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

  - Logistic Lens Model Equation

$$\mathrm{r}_a = G\frac{\sigma_{\tilde{Y}_e}}{\sigma_{Y_e}}\frac{\sigma_{\tilde{Y}_s}}{\sigma_{Y_s}} + C_1\frac{\sigma_{\tilde{Z}_e}}{\sigma_{Y_e}}\frac{\sigma_{\tilde{Z}_s}}{\sigma_{Y_s}} + C_2\frac{\sigma_{\tilde{Y}_e}}{\sigma_{Y_e}}\frac{\sigma_{\tilde{Z}_s}}{\sigma_{Y_s}} + C_3\frac{\sigma_{\tilde{Z}_e}}{\sigma_{Y_e}}\frac{\sigma_{\tilde{Y}_s}}{\sigma_{Y_s}}$$

  - The latter is quite general, and can handle use of
    - more predictors in one model than another,
    - or different model forms.

Conflicting Desirable Features of a Research Design

- Representative sample of cases
  - Often the cues are correlated
  - Often important categories occur rarely
- Feasible task for participants
  - Small number of stimulus cases
  - Small number of cues
- Manageable analysis. It is easier to fit, interpret, and generalize models when stimulus set has
  - Well distributed criterion (not skewed)
  - Small or 0 intercorrelations among cues
    - E.g., fractional factorial design.

A Study of Judgments about a Rare Category

- Yang et al 2013. To ask participants to judge a small set of cases, boosted the category to 50% --10 out of 20 cases.
  - Random draw from cases
  - Hard to get unique fits for linear regression
  - High cue intercorrelation – different from what was expected.

Table A2. Cue and criterion intercorrelations and multicollinearity diagnostics (tolerance). Above diagonal: Sample (n = 20). Below diagonal: Subbe data (N = 555).

Cues	Systolic BP	Heart rate	Resp. rate	Temper-ature	Conscious-ness	Ecology Criterion	Tolerance
Systolic BP	~	.072	.032	.558*	-.616**	-.147	.25
Heart rate	.004	~	.507*	.493*	.485*	.455*	.47
Respiration rate	.014	.284***	~	.353	.542*	.594**	.45
Temperature	-.047	.223***	.126**	~	.007	.023	.43
Consciousness	-.145**	.069	.116**	.088*	~	.523*	.18
Ecology Criter.	-.171***	.124**	.179***	-.031	.161***	~	
Tolerance	.98	.88	.91	.94	.96		

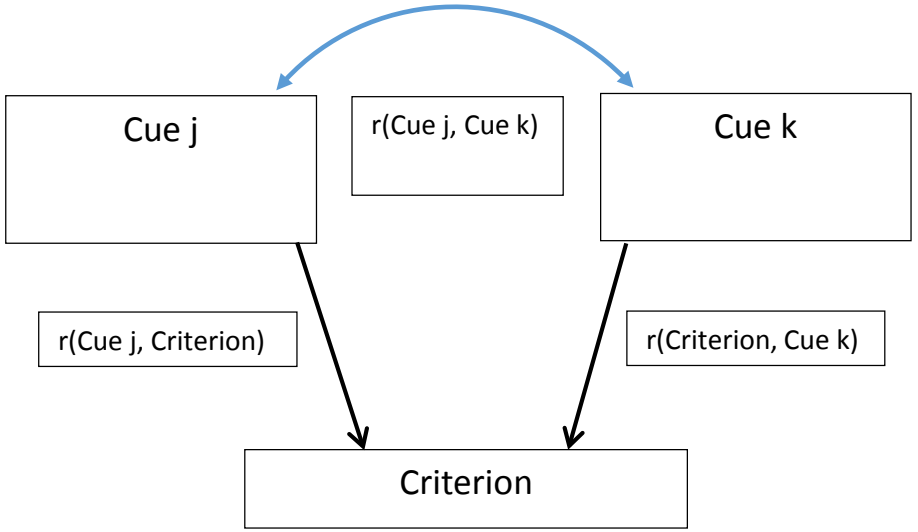
Correlation is significant at: * the 0.05 level; ** the 0.01 level; *** the 0.001 level (all 2-tailed).
Note: The ecology criterion is whether the patient is "at risk," as defined in the text. Tolerance is $1 - R^2$, the complement of the proportion of variance explained in a multiple linear regression predicting the cue from all the other cues.

Why did cue-cue correlation change?

- These ideas emerged:
  - Expectation of cue-cue correlations due to cue-criterion correlations.
  - Cue-cue correlations within a category may differ for different categories.
  - The mixture proportion, category versus non-category, can affect these correlations, and hence
    - Possibility of fitting a good regression model
    - Size of regression coefficients, relative weights.

Cue-cue correlations through criterion.

- The dependency of the cue-cue correlations on
  - Cue-criterion correlations, adapting a model in which the cues cause the criterion.



$$r(Cue_j, Cue_k) = r(Cue_j, Crit) * r(Crit, Cue_k)$$

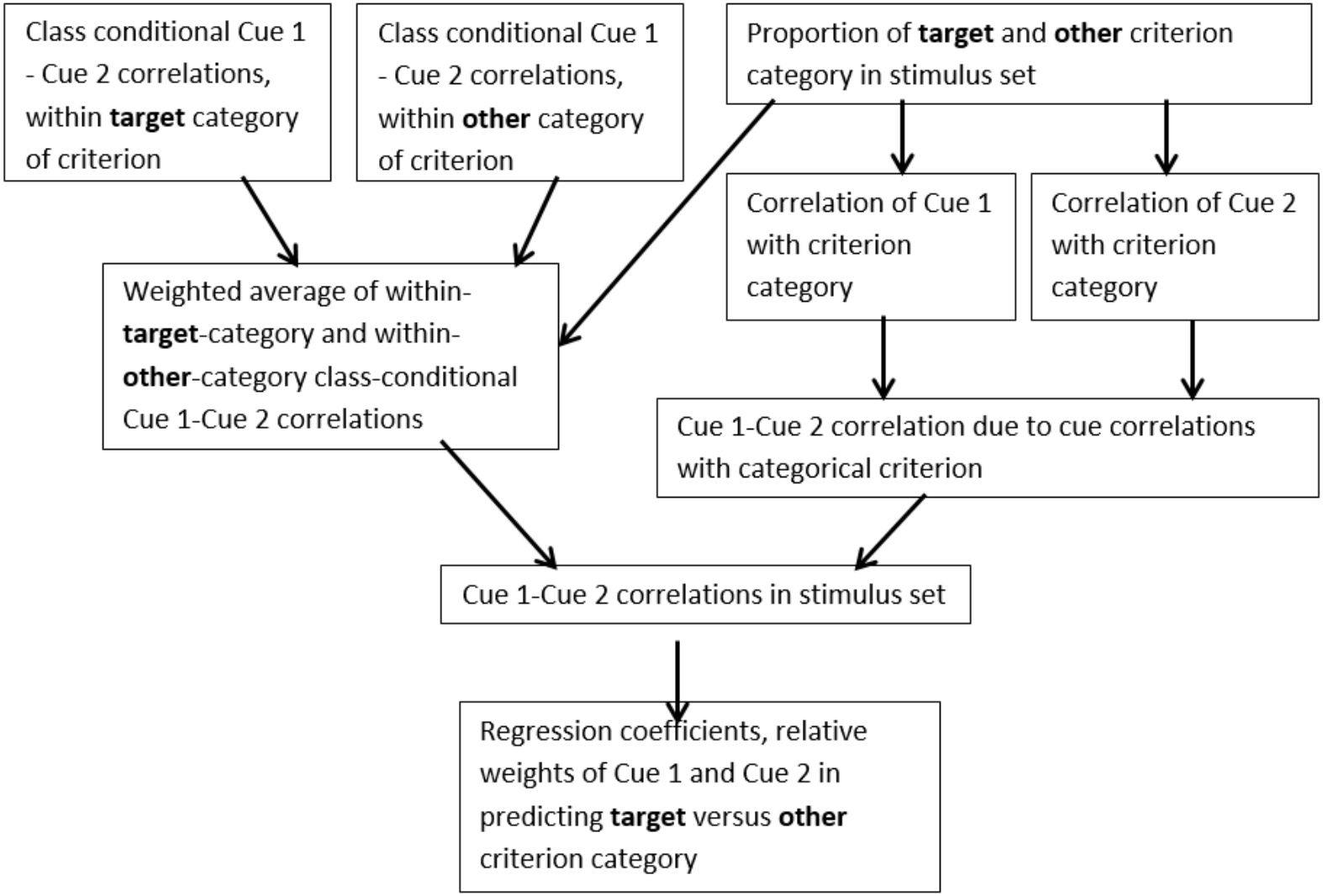
- The dependency of the overall cue-cue correlation on
  - Cue-cue correlations within each category
  - Proportions of each (of two) categories
    - Especially when there are different class-conditional cue-cue correlations in the one class versus in the other class.

$$\frac{p(C^+) * \left(r(S_1, S_2|C^+)\right) + p(C^-) * \left(r(S_1, S_2|C^-)\right)}{p(C^+) + p(C^-)}$$

- Combined

$$r(S_1, S_2) = \lambda * [r(S_1, C) * r(S_2, C)] + (1 - \lambda) * \frac{p(C^+) * \left(r(S_1, S_2|C^+)\right) + p(C^-) * \left(r(S_1, S_2|C^-)\right)}{p(C^+) + p(C^-)}$$

General model of influence of cue-cue correlation, cue-criterion correlation, and category proportion on relative weights.



Direction of change in cue-cue correlation as category proportion changes.

Whole Set with Original Proportions

	C^+	C^-	$Total$
S_1^+	a	b	e
S_1^-	c	d	f
$Total$	g	h	n

Whole Set with Adjusted Proportions

	C^+	C^-	$Total$
S_1^+	ka	b	e'
S_1^-	kc	d	f'
$Total$	kg	h	n'

$$\varphi(S_1, C)_{Original} = \frac{ad - bc}{\sqrt{efgh}} = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}$$

$$\varphi(S_1, C)_{Adjusted} = \frac{kad - kbc}{\sqrt{(ka+b)(kc+d)(ka+kc)(b+d)}}$$

- The $\varphi(S_1, C)$ in the original stimulus set and in the set with 50% of the target category would differ by a factor of

$$\frac{\sqrt{(a+b)(c+d)}}{\sqrt{\left(a+\frac{b}{k}\right)(kc+d)}}$$

- And its direction can not be predicted from knowing *k*.

Overall and within-category cue-cue correlations.

- Direction of change in overall cue-cue correlation due to within-category cue-cue correlations and proportion changes.

Disease (C^+)

	S_2^+	S_2^-	$Total$
S_1^+	a_1	b_1	e_1
S_1^-	c_1	d_1	f_1
$Total$	g_1	h_1	n_1

No Disease (C^-)

	S_2^+	S_2^-	$Total$
S_1^+	a_2	b_2	e_2
S_1^-	c_2	d_2	f_2
$Total$	g_2	h_2	n_2

Whole Set with Original Proportions

	S_2^+	S_2^-	$Total$
S_1^+	$a_1 + a_2$	$b_1 + b_2$	$e_1 + e_2$
S_1^-	$c_1 + c_2$	$d_1 + d_2$	$f_1 + f_2$
$Total$	$g_1 + g_2$	$h_1 + h_2$	$n_1 + n_2$

Whole Set with Adjusted Proportions

	S_2^+	S_2^-	$Total$
S_1^+	$ka_1 + a_2$	$kb_1 + b_2$	$ke_1 + e_2$
S_1^-	$kc_1 + c_2$	$kd_1 + d_2$	$kf_1 + f_2$
$Total$	$kg_1 + g_2$	$kh_1 + h_2$	$kn_1 + n_2$

Corresponding formulas

- Original category proportions

$$\varphi(S_1, S_2)_{Original} = \frac{(a_1 + a_2)(d_1 + d_2) - (b_1 + b_2)(c_1 + c_2)}{\sqrt{(e_1 + e_2)(f_1 + f_2)(g_1 + g_2)(h_1 + h_2)}}$$

- Adjusted category proportions

$$\varphi(S_1, S_2)_{Adjusted} = \frac{(ka_1 + a_2)(kd_1 + d_2) - (kb_1 + b_2)(kc_1 + c_2)}{\sqrt{(ke_1 + e_2)(kf_1 + f_2)(kg_1 + g_2)(kh_1 + h_2)}}$$

- As k is in both numerator and denominator, direction of change is not predictable.

$$\varphi(S_1, S_2)_{Adjusted} = \frac{k^2(a_1d_1 + b_1c_1) + k(a_1d_2 + a_2d_1 - b_1c_2 - b_2c_1) + (a_2d_2 + b_2c_2)}{\sqrt{\begin{matrix} k^4e_1f_1g_1h_1 + k^3(e_1f_1(g_1h_2 + g_2h_1) + (e_1f_2 + e_2f_1)g_1h_1) \\ +k^2(e_1f_1g_2h_2 + (e_1f_2 + e_2f_1)(g_1h_2 + g_2h_1) + e_2f_2g_1h_1) \\ +k\big((e_1f_2 + e_2f_1)g_2h_2 + e_2f_2(g_1h_2 + g_2h_1)\big) + e_2f_2g_2h_2 \end{matrix}}}$$

What cue-cue relations occur in actual environmental data?

- This is a statistical feature of the environment, not of people's judgment. No psychology in this talk!
- Fever data set, 2200 patients with fever at a clinic.
  - Restrict to clinical cues – signs and symptoms.
  - Look at the most common diseases, >= 50 patients.
- Look only at symptoms that predict the target disease and that have class-conditional cue-cue correlations within the disease.
  - Additional filter – and the clinicians recognize the correlation as clinically important, understandable.
  - But still, not necessarily looking at the most useful cues;
    - only the most correlated pairs of useful cues.
- Multiple analyses.
  - A) **against set of other common diseases**, or against all others.
  - B) with the ecological prevalence, or with the prevalence of the target disease inflated/deflated to be 50%

Criteria and cues.

- 5 diagnoses for which there were more than 50 cases (> 2.5% of cohort).
  - Malaria Falciparum (N=473), Dengue Fever (N=69), Pneumonia (N=67), Rickettsiosis (N=62), Tonsillitis (N=50).
- Twenty symptoms. All yes/no.
  - SkinRash, SkinErupt, SkinUlcers, CoughProd, LungAusc, CoughIrrit, Vomiting, Nausea, SymptENT, LocLymph, Headache, Myalgia, Fever39, BloodDiarrh, AbdoPain, Diarrh, SplenoMeg, Jaundice, HepatoMeg, GenLymph
- 20 symptoms → 20*19/2 = 190 pairs. No!

Make a display of how many pairs for each of the 5 diseases.

- 9, 10, 6, 8, 2 pairs per disease, in order M, D, P, R, T.

Example Symptom Pairs for Malaria

- Not much difference in overall correlations

Min	-0.029
Mean	0.006
Max	0.041

Sympt 1	Sympt 2	$R(S_1,S_2)$ w/in **Target** Disease	$R(S_1,S_2)$ w/in **Other** Diseases	Correls in Original Set	Correls in 50% Set
Skin Eruption	Skin Ulcers	0.312	0.440	0.505	0.498
Vomiting	Nausea	0.658	0.555	0.661	0.654
Headache	Myalgia	0.336	0.209	0.294	0.274
SplenoMeg	Jaundice	0.193	0.190	0.225	0.236
SplenoMeg	HepatoMeg	0.167	-0.025	0.160	0.147
Jaundice	HepatoMeg	0.125	-0.010	0.130	0.127
CoughProductive	LungAusc	0.104	0.583	0.518	0.554
CoughProductive	CoughIrritated	0.174	0.418	0.364	0.395
LungAusc	CoughIrritated	0.088	0.380	0.312	0.347

Example: Symptom Pairs for Rickettsia Africae

- Changes in correlation

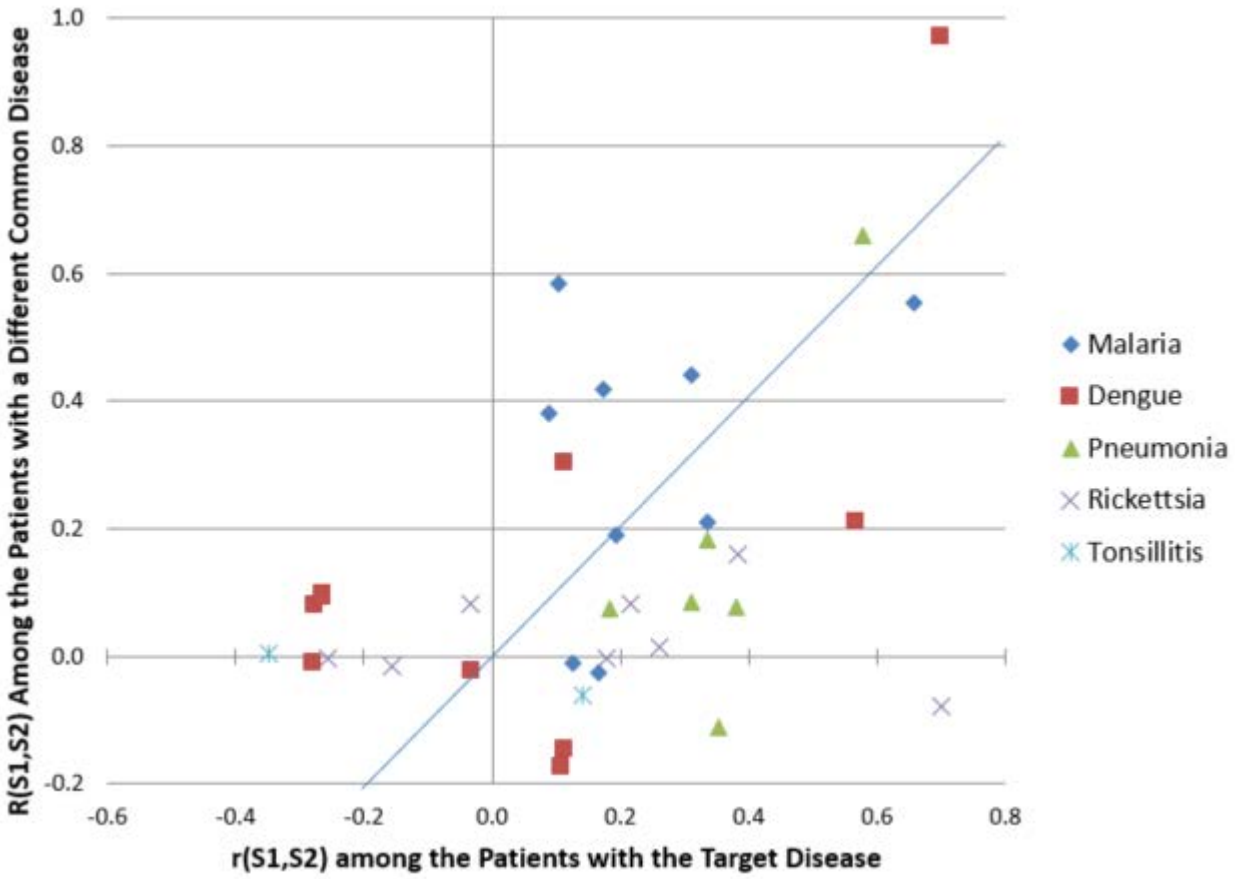
Min	-0.159
Mean	-0.019
Max	0.226

- Bigger differences than with malaria

Sympt 1	Sympt 2	$R(S_1,S_2)$ w/in **Target** Disease	$R(S_1,S_2)$ w/in **Other** Diseases	Correls in Original Set	Correls in 50% Set
SymptENT	SplenoMeg	0.701	-0.079	-0.056	0.103
Diarrhea	Jaundice	-0.034	0.082	0.089	0.101
SkinErupt	Nausea	-0.255	-0.004	-0.139	-0.365
SkinErupt	AbdPain	-0.156	-0.017	-0.073	-0.185
Nausea	AbdPain	0.384	0.161	0.183	0.272
CoughIrr	Fever	0.178	-0.003	0.016	0.100
CoughIrr	HepatoMeg	0.261	0.014	0.028	0.116
Fever	HepatoMeg	0.217	0.082	0.093	0.148

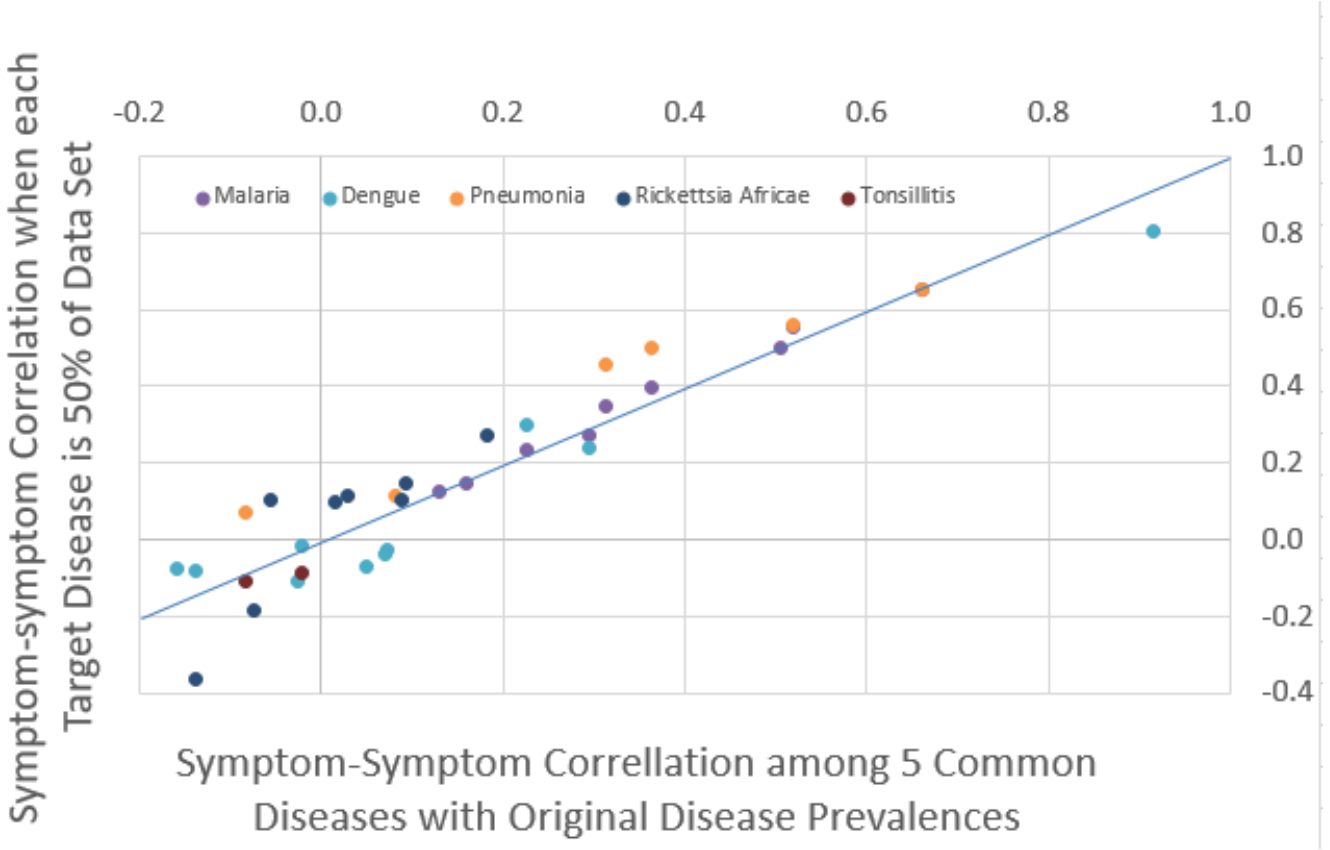
Relation Between the Disease-Conditional Symptom-Symptom Correlations Among the Target Disease Patients and the Patients with Other Diseases; Set of 5 Common Diseases

- Shows that cue-cue correlation within target category can differ from those in alternative category.



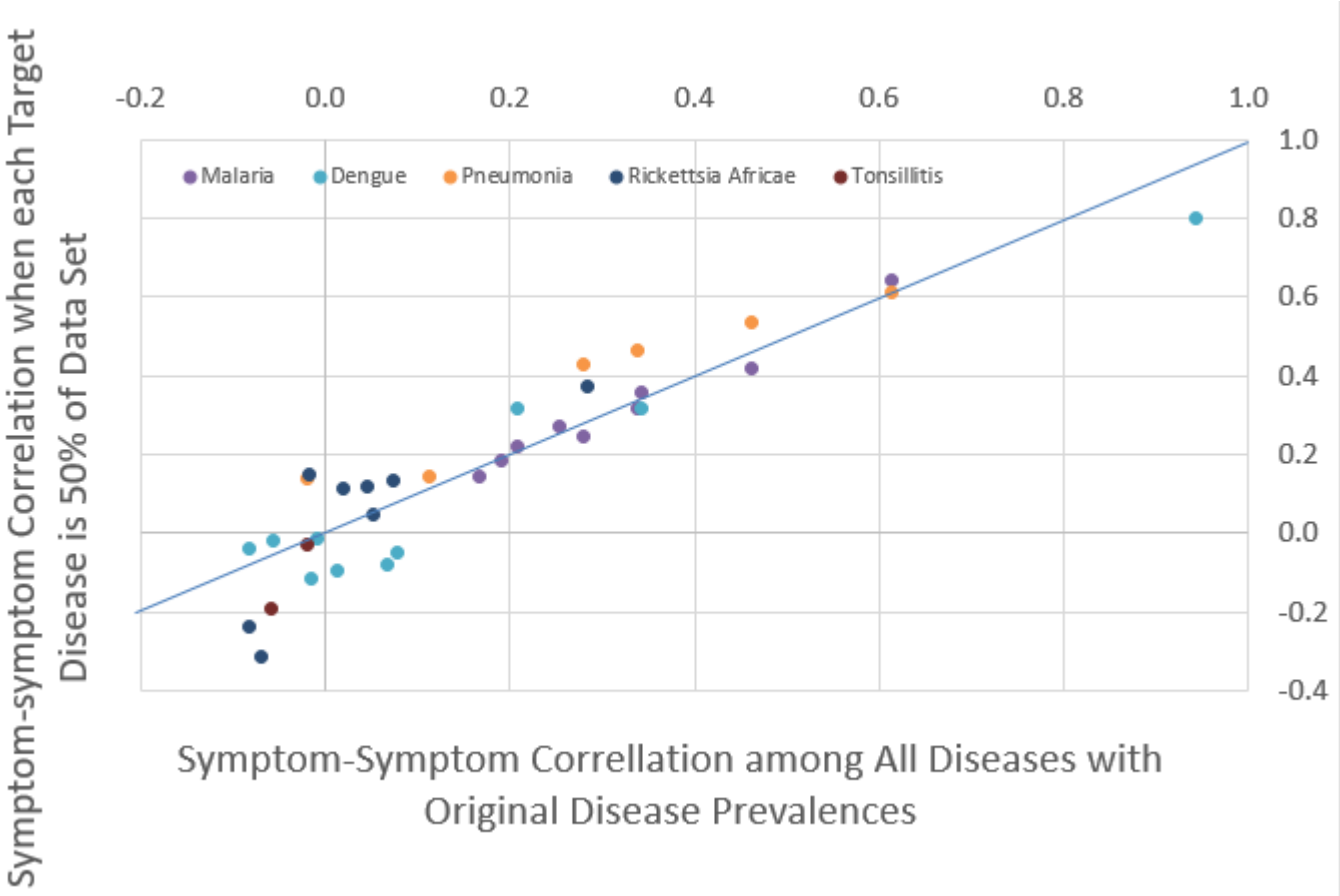
Comparison of Symptom-Symptom correlations in Original Data Set versus Adjusted Data Sets (50% Target Disease), for each of 5 Diseases, Compared against the Other 4.

- Correlations are not identical, but similar.
- Is this large enough to be a problem for modeling correspondence of judgment policies to the environmental relations?



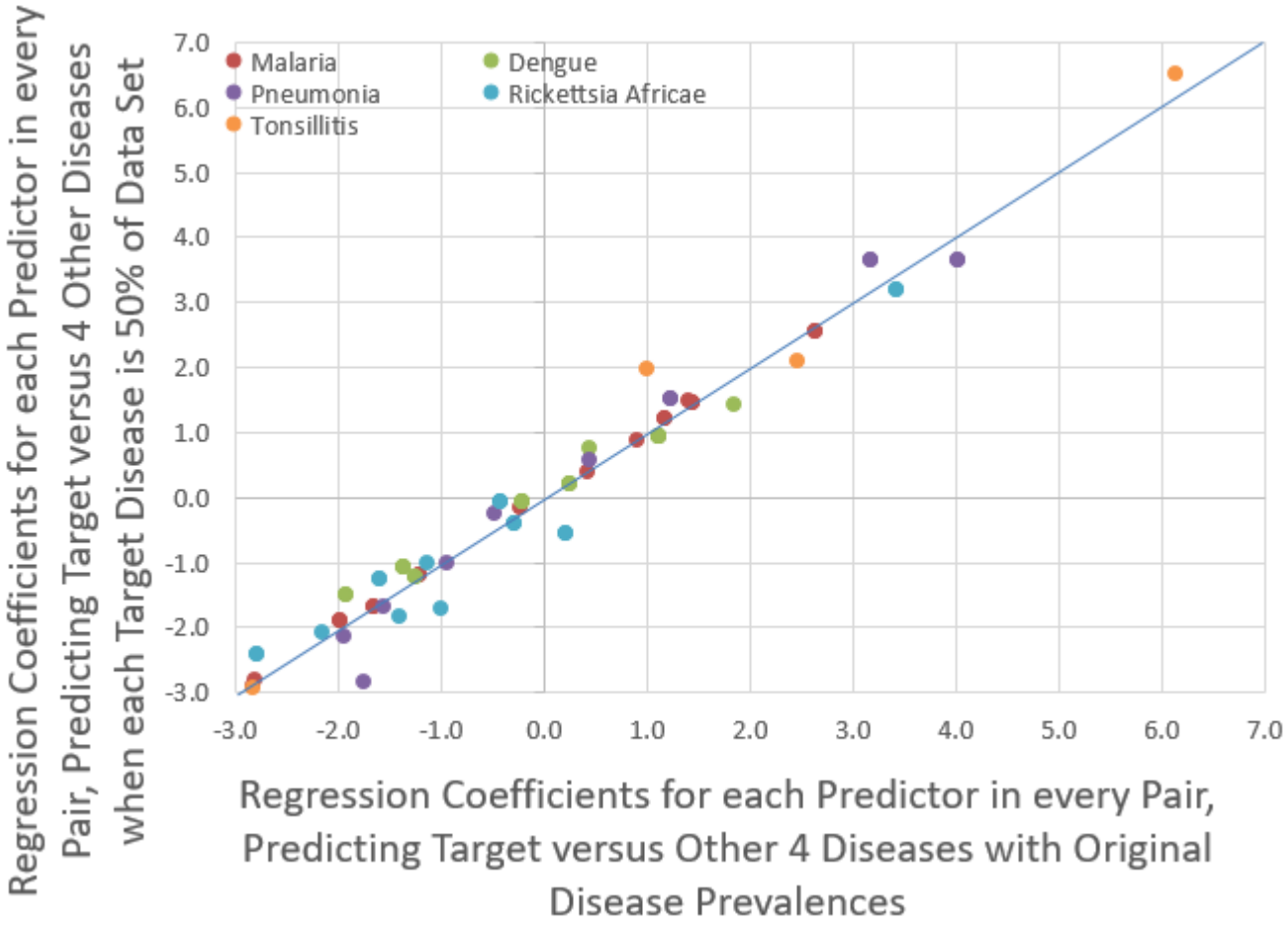
Comparison of Symptom-Symptom correlations in Original Data Set versus Adjusted Data Sets (50% Target Disease), for each of 5 Diseases, Compared against All Others.

- Same comparison, when target diseases are compared with all others in fever data set.



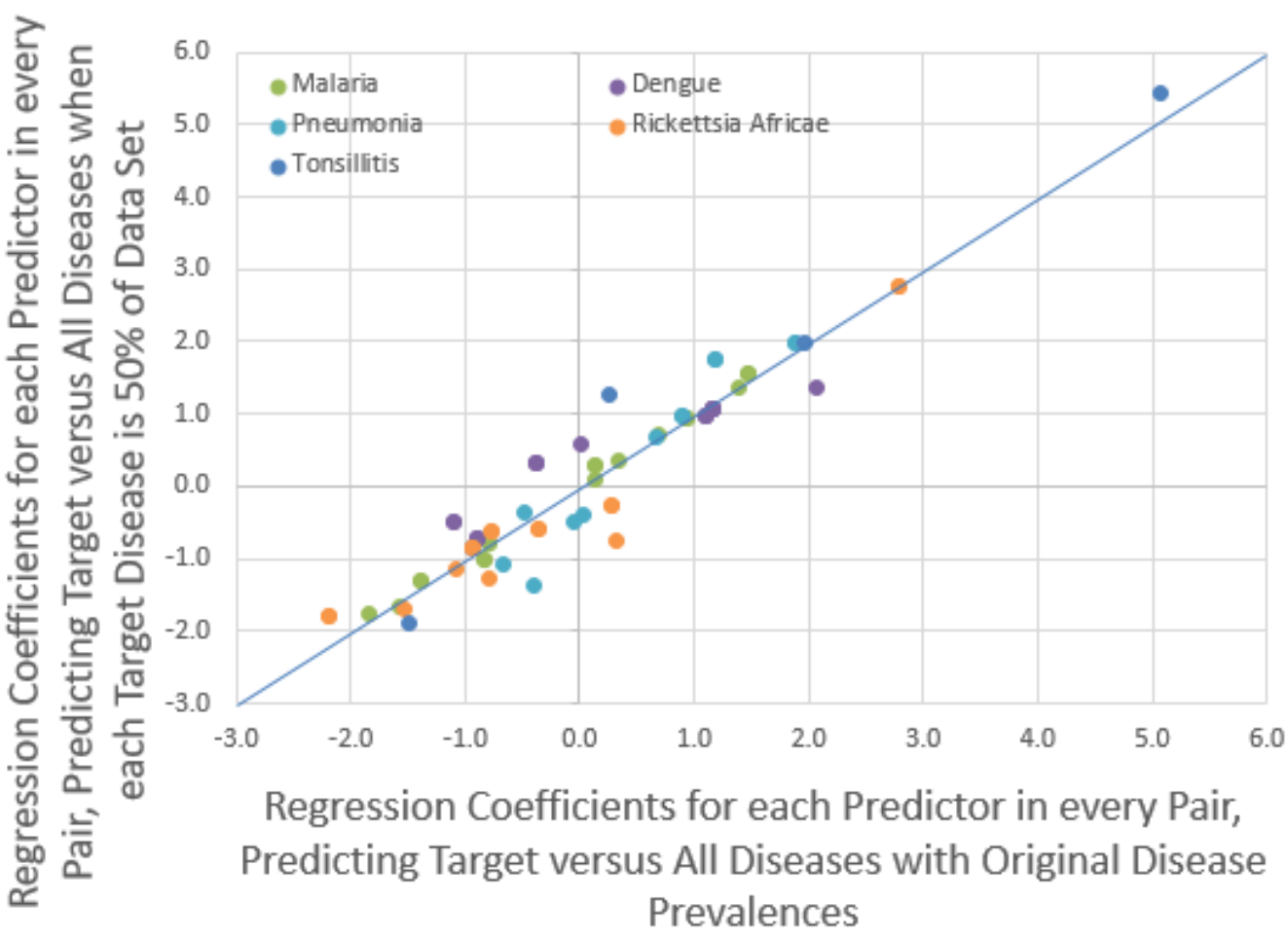
Comparison of Regression Coefficients of Symptoms Predicting Target Diseases, in Original Data Set versus Adjusted Data Sets (50% Target Disease), for each of 5 Diseases.

- The regression coefficients are pertinent to the interpretation of the judgment and ecology models in the Brunswik Lens Model.



Regression Coefficients in the Original Data Set Compared to the Adjusted (50%) Data Sets for Each Disease, in the Set of All Diseases

- Same with the full data set.
- Coefficients are smaller.



Conclusions.

- We've explained a problem
  - Cue-cue correlations can differ within category, and this may cause distorted summary models of judgment and lens models if the proportion of each category is altered.
- Analyzed its causes
  - Effects of within category cue-cue correlations and category proportions upon overall cue-cue correlations
- Shown the effect is not predictable from the magnitude of the adjustment in proportions
- Shown that it occurs in a real data set but the magnitude does not seem very large, in this domain.

Implications

- If designing a set of stimuli for a judgment study or a Brunswik Lens Model study
  - Yes, inspect the world and make the distribution of the features in your stimuli representative of their distributions in the world.
  - Yes, you may need to simplify by looking at only the more important features.
  - Yes, you may need to adjust the frequency of your rarely occurring categories, else you won't have any data to look at because your subjects will quit.
  - Although the cue-cue correlations may not have much of an effect on your conclusions (as in these data), they still might (as in Hamm and Yang), so you should make a point to if the cue-cue intercorrelations in your data set are similar to those in the world.