

Mammogram stimuli: An initial set of images was obtained from the DDSM volumes of normal and cancer cases. The two authors and a third non-naïve colleague rated the extent to which each image in these cases exemplified the given target present or absent category on a scale of 0 (poor example) to 100 (excellent example) in terms of abnormality (in the cancer cases) or normalcy (in the normal cases). The averages of the ratings were used to rank the images in each target present/absent category. Then, the top-ranked 270 target present images and 675 target absent images were selected for use in Experiment 1, and any superfluous background marks (e.g., writing or other visual anomalies) were painted over using a custom-made MATLAB program. The top-rated 220 target present images and 580 target absent images were used in the experimental conditions, whereas the remaining 50 target present and 95 target-absent images were used in the training and practice parts of the experiment. Twenty target present images for the 5% condition were obtained by choosing every 10th image from the list of ranked target present images. The remaining 200 target present images were used for the 50% condition. The 580 rank-ordered target absent images were divided into quartiles, with the order of the 145 images in each quartile randomized. The first 95 images in each quartile were used in the 5% condition and the last 50 images in each quartile were used in the 50% condition. The entire set of 945 mammograms used in Experiment 1 (in jpg format) and participants' corresponding slider estimates (mean and SD) for each image are available here: <https://osf.io/uqg5m/#>).

Training & Practice: To familiarize participants with all experimental tasks, further ensure that they were able to overcome the tendencies to use extreme or middle values, and guard against potential response biases in the different prevalence conditions, participants were required to pass a practice block of 60 trials in each of the 5% and 50% Mammos conditions in Experiment 1, a

practice block of 64 trials in each of the Local Gabors and Global Gabors conditions, and a practice block of 60 trials in each of the 5% and 50% Tools conditions in Experiment 2. The practice blocks were identical to the main tasks with the exception that none of the images used in the practice conditions were used in the experimental conditions.

Each participant also performed a training task before beginning the first Tools condition to ensure they could identify and correctly categorize all of the 60 items as tools or non-tools. On each training trial, one of the 60 possible items was presented in the center of the display at full contrast on a white background. Participants' task was to decide to which of five categories the item belonged: Clothes, Fruit, Toiletries, Toys, or Tools. The name of the item was centered above the image of the object, and the five choices and corresponding numbers were centered below the object. The display remained visible until a response was made on each trial. When participants failed to correctly classify an object, an error warning was shown and this stimulus was repeated at the end of the block.

At the start of the training blocks, participants were shown written, illustrated instructions on the computer screen describing the task. At the start of each practice block, participants were presented with detailed instructions describing their task as a "radiologist," or "baggage screener." Importantly, participants were informed of the 5% or 50% target prevalence at the start of each practice and experimental block. Trials were self-paced to allow participants to take short breaks during blocks, there were mandatory breaks (about 1 minute) between blocks, and a longer 5-minute break halfway through each experimental session.

Table S1. Constructing all possible unique arrangements of observers into dyads. In this example, the 18 participants in Experiment 1 were grouped into nine dyads such that each arrangement included all observers but each possible dyad was present only once in the resultant 17 arrangements. The analyses in Figure 3 were then conducted using the conditional means from the 17 possible arrangements. The same process was repeated for the 16 observers in Experiment 2, resulting in 15 possible arrangements of all 16 participants into eight dyads.

	Dyad 1	Dyad 2	Dyad 3	Dyad 4	Dyad 5	Dyad 6	Dyad 7	Dyad 8	Dyad 9	Conditional means used in Figure 3 analyses									
Arrangement 1	1	18	2	17	3	16	4	15	5	14	6	13	7	12	8	11	9	10	Average of Arrangement 1
Arrangement 2	18	10	11	9	12	8	13	7	14	6	15	5	16	4	17	3	1	2	Average of Arrangement 2
Arrangement 3	2	18	3	1	4	17	5	16	6	15	7	14	8	13	9	12	10	11	Average of Arrangement 3
Arrangement 4	18	11	12	10	13	9	14	8	15	7	16	6	17	5	1	4	2	3	Average of Arrangement 4
Arrangement 5	3	18	4	2	5	1	6	17	7	16	8	15	9	14	10	13	11	12	Average of Arrangement 5
Arrangement 6	18	12	13	11	14	10	15	9	16	8	17	7	1	6	2	5	3	4	Average of Arrangement 6
Arrangement 7	4	18	5	3	6	2	7	1	8	17	9	16	10	15	11	14	12	13	Average of Arrangement 7
Arrangement 8	18	13	14	12	15	11	16	10	17	9	1	8	2	7	3	6	4	5	Average of Arrangement 8
Arrangement 9	5	18	6	4	7	3	8	2	9	1	10	17	11	16	12	15	13	14	Average of Arrangement 9
Arrangement 10	18	14	15	13	16	12	17	11	1	10	2	9	3	8	4	7	5	6	Average of Arrangement 10
Arrangement 11	6	18	7	5	8	4	9	3	10	2	11	1	12	17	13	16	14	15	Average of Arrangement 11
Arrangement 12	18	15	16	14	17	13	1	12	2	11	3	10	4	9	5	8	6	7	Average of Arrangement 12
Arrangement 13	7	18	8	6	9	5	10	4	11	3	12	2	13	1	14	17	15	16	Average of Arrangement 13
Arrangement 14	18	16	17	15	1	14	2	13	3	12	4	11	5	10	6	9	7	8	Average of Arrangement 14
Arrangement 15	8	18	9	7	10	6	11	5	12	4	13	3	14	2	15	1	16	17	Average of Arrangement 15
Arrangement 16	18	17	1	16	2	15	3	14	4	13	5	12	6	11	7	10	8	9	Average of Arrangement 16
Arrangement 17	9	18	10	8	11	7	12	6	13	5	14	4	15	3	16	2	17	1	Average of Arrangement 17

Table S2. Results of planned comparisons between dyads and individuals and dyads and the better of the two individuals for d' , proportion of misses, and proportion of false alarms in each of the Mammos, Gabors, and Tools conditions. For each comparison, the means of the two factors, the percent increase in the dependent measure from Factor A to Factor B, the t-value of the comparison, the degrees of freedom (df), the corresponding p-value, and Cohen's d are reported. Asterisks represent significant differences, with $\alpha = 0.0125$ using the Bonferroni correction for four multiple comparisons.

MAMMOS	Factor A	Factor B	Mean A	Mean B	% increase	t	df	p	Cohen's d
<u>d'</u>	5% Dyad	5% Indiv	2.94	2.60	12.89	16.32	16	< 0.001*	5.60
	5% Dyad	5% Better	2.94	2.87	2.32	3.25	16	0.005*	0.98
	50% Dyad	50% Indiv	2.37	2.14	10.81	23.71	16	< 0.001*	8.14
	50% Dyad	50% Better	2.37	2.31	2.31	4.57	16	< 0.001*	1.47
<u>Prop. Misses</u>	5% Dyad	5% Indiv	0.16	0.20	-17.48	-6.33	16	< 0.001*	2.17
	5% Dyad	5% Better	0.16	0.14	20.97	4.70	16	< 0.001*	1.42
	50% Dyad	50% Indiv	0.08	0.11	-27.31	-23.44	16	< 0.001*	8.04
	50% Dyad	50% Better	0.08	0.09	-15.58	-7.71	16	< 0.001*	2.42
<u>Prop. FAs</u>	5% Dyad	5% Indiv	0.04	0.07	-44.92	-30.34	16	< 0.001*	10.41
	5% Dyad	5% Better	0.04	0.08	-50.22	-15.02	16	< 0.001*	5.26
	50% Dyad	50% Indiv	0.19	0.20	-9.11	-11.13	16	< 0.001*	3.82
	50% Dyad	50% Better	0.19	0.18	4.23	2.52	16	0.023	0.90
GABORS	Factor A	Factor B	Mean A	Mean B	% increase	t	df	p	Cohen's d
<u>d'</u>	Local Dyad	Local Indiv	3.67	3.01	21.80	30.96	14	< 0.001*	11.30
	Local Dyad	Local Better	3.67	3.51	4.26	5.03	14	< 0.001*	1.49
	Global Dyad	Global Indiv	1.31	1.13	16.06	14.59	14	< 0.001*	5.33
	Global Dyad	Global Better	1.31	1.28	2.13	2.87	14	0.012*	0.67
<u>Prop. Misses</u>	Local Dyad	Local Indiv	0.03	0.09	-64.35	-33.58	14	< 0.001*	12.27
	Local Dyad	Local Better	0.03	0.05	-34.61	-8.21	14	< 0.001*	1.92
	Global Dyad	Global Indiv	0.25	0.31	-18.55	-20.66	14	< 0.001*	7.54
	Global Dyad	Global Better	0.25	0.28	-10.36	-5.92	14	< 0.001*	2.03
<u>Prop. FAs</u>	Local Dyad	Local Indiv	0.06	0.08	-30.14	-13.33	14	< 0.001*	4.87
	Local Dyad	Local Better	0.06	0.05	12.86	6.05	14	< 0.001*	0.85
	Global Dyad	Global Indiv	0.27	0.28	-1.82	-1.14	14	0.273	0.42
	Global Dyad	Global Better	0.27	0.25	7.58	5.56	14	< 0.001*	1.19
TOOLS	Factor A	Factor B	Mean A	Mean B	% increase	t	df	p	Cohen's d
<u>d'</u>	5% Dyad	5% Indiv	2.17	1.78	22.10	24.66	14	< 0.001*	9.00
	5% Dyad	5% Better	2.17	2.01	8.12	12.08	14	< 0.001*	3.30
	50% Dyad	50% Indiv	1.95	1.59	22.14	28.18	14	< 0.001*	10.29
	50% Dyad	50% Better	1.95	1.83	5.89	8.74	14	< 0.001*	2.00
<u>Prop. Misses</u>	5% Dyad	5% Indiv	0.36	0.43	-15.02	-14.88	14	< 0.001*	5.43
	5% Dyad	5% Better	0.36	0.40	-10.20	-6.50	14	< 0.001*	2.06
	50% Dyad	50% Indiv	0.16	0.25	-34.73	-38.96	14	< 0.001*	14.22
	50% Dyad	50% Better	0.16	0.23	-28.34	-23.72	14	< 0.001*	5.44
<u>Prop. FAs</u>	5% Dyad	5% Indiv	0.04	0.07	-41.46	-24.70	14	< 0.001*	9.03
	5% Dyad	5% Better	0.04	0.04	-8.67	-2.56	14	0.004*	0.71
	50% Dyad	50% Indiv	0.18	0.20	-7.80	-3.48	14	0.023	1.27
	50% Dyad	50% Better	0.18	0.15	19.66	6.72	14	< 0.001*	1.42

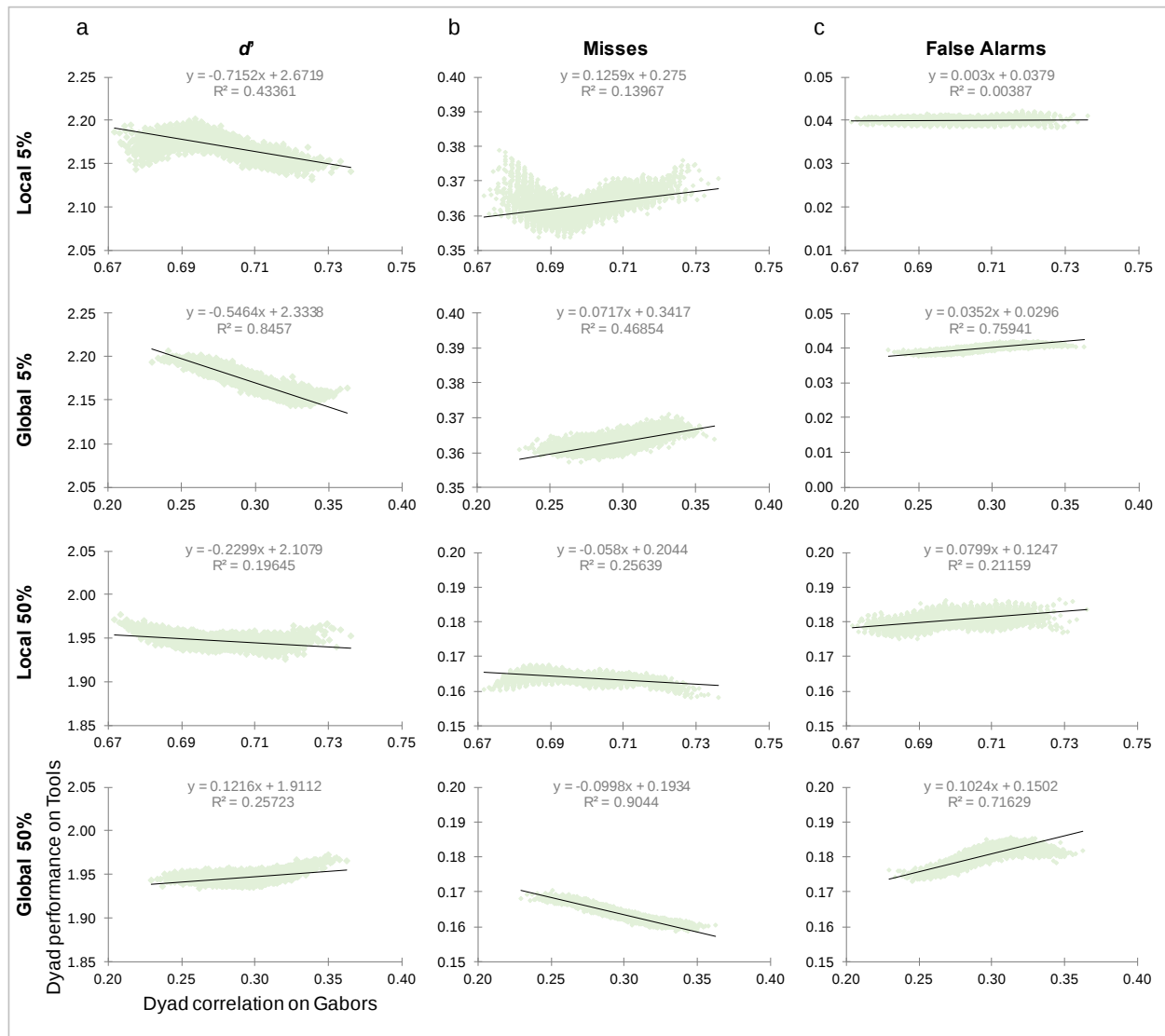


Figure S1. Binary decorrelated dyads results. **a. Detection performance (d'):** Dyads' d' performance on the Tools task was more negatively correlated with their correlated d' performance on the Gabors task when the continuous data was used to correlate performance on the Gabors task (Figure 4a) than when dichotomized data (responses $> 50 = 1$ /responses $< 50 = 0$) was used to correlate performance on the Gabors task in all conditions except the Local 5% condition where there was now a slight negative correlation using the dichotomized data. **b. Misses:** The slight improvement in d' in the 5% Tools task for the most decorrelated observers based on dichotomized versus continuous responses in the Local Gabors task was due to the slight positive correlation between dyads' hits in the 5% Tools task and their correlated dichotomized performance on the Local Gabors task compared to the reversed pattern of a slight negative correlation using continuous estimates in the Local Gabors task (Figure 4b). **c. False alarms:** Relative to the results obtained from using participants' continuous estimate in the Gabors tasks (Figure 4c), there were no improvements in false alarm rates in the Tools tasks from using participants' dichotomized data to decorrelate performance in the Gabors tasks. All plots depict the results of simple regression analyses with all $F_s(1,9008) > 34.99$ and all $p_s < 0.001$.