Improving Accuracy in Bayesian Inference Problems through Training

Alaina N. Talboy, M.A. and Sandra L. Schneider, Ph.D.

Judgment and Decision Making Lab, Department of Psychology, University of South Florida, Tampa, FL, USA

Abstract

Building on classic medical Bayesian inference problems, we asked participants to determine the positive predictive value of medical, legal, and sports inference problems (i.e., likelihood that positive tests correctly indicate presence of condition of interest). Training was provided either using graphs or tables, whereas a third control group received no training. Both training groups provided accurate estimates more often than those in the control group, though accuracy rates overall were low (45-50% for training and 32% for control groups). Table training improved performance only for problems presented as tables, whereas graph training improved performance in both table and graph formats.

Introduction

Improving accuracy on Bayesian inference problems can be achieved through use of natural frequencies instead of probabilities (Galesic et al., 2009; Gigerenzer et al., 2008) or inclusion of visual aids (Galesic & Garcia-Retamero, 2011). However, accuracy has not exceeded 62% in previous studies.

Researchers allude to the possible benefits of training, but it has not been tested prior to this study. The current study utilizes natural frequencies throughout the inference problems, then uses training to teach participants how to convert those to probabilities to determine if accuracy can exceed 62%.

Hypotheses: Those who receive training will perform better than those in the control condition. Performance will be better on graph problems than table problems because graphs problems include both numerical and spatial information about the quantities involved.

Inference Problems

Topic	Sample	BR	(Test +, Act +)	(Test +, Act -)
Mammogram	10,000	50	80 | 100	990 | 9,900
Diabetes	10,000	100	48 | 50	4,975 | 9,950
Polygraph	1,000	50	47 | 50	47 | 950
Recidivism	1,000	156	130 | 156	220 | 844
Tennis	10,000	2,800	2,000 | 2,800	1,100 | 7,200
Baseball	146	99	79 | 99	9 | 47

Note. BR. = base rate. (Test +, Act +) = the number of people who test positive correctly out of the number of people who are actually positive. (Test +, Act -) = the number of people who test positive erroneously out of the number of people who are actually negative.

Design and Variables

Design: 3 x 2 x 3 Training x Problem Format x Domain

Participants: 208 undergraduates (155 female)

Independent Variables

- **Training Condition**
  - Graph Training
  - Table Training
  - Control
- **Problem Format**
  - Graph Problems
  - Table Problems
- **Domain**
  - Medical
  - Legal
  - Sports

Dependent Variable

- Number of inference problem responses within +/-10% of correct PPV response

Individual Difference Measures

- Subjective Numeracy Scale (SNS)
- Rasch-based Numeracy Scale (RBN)
  - Includes items from the Objective Numeracy Scale and the Cognitive Reflection Test
- Graph Literacy Scale (GL)

Individual Difference Measures

	Graph Problems	Table Problems	SNS	RBN	GL
Graph Problems	1	**0.361****	0.131	**0.340****	0.091
Table Problems		1	**0.149***	**0.322****	0.124
SNS			1	**0.393****	0.114
RBN				1	**0.277****
GL					1

** significant at $p < .01$ (2-tailed); * significant at $p < .05$ (2-tailed)

Key References

- Galesic, M., & Garcia-Retamero, R. (2011). Graph Literacy A Cross-Cultural Comparison. *Medical Decision Making, 31*(3), 444-457.
- Galesic, M., Gigerenzer, G., & Straubinger, N. (2009). Natural frequencies help older adults and people with low numeracy to evaluate medical screening tests. *Medical Decision Making, 29,* 368-371.
- Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review*, *102*(4), 684.

Primary Results

Planned Comparison (Main Effects):

- Training conditions (M = 1.43, SD = .88) provided correct responses more often than Control condition (M = 0.95, SD = .64), $F(1, 205) = 16.07$, $p < .001$
- Accuracy was higher on table problems (M = 1.35, SD = .31) than on graph problems (M = 1.19, SD = .33), $F(1, 205) = 4.37$, $p = .038$, $\eta^2 = .02$

Interaction: See below, $F(2, 205) = 4.02$, $p = .019$, $\eta^2 = .04$

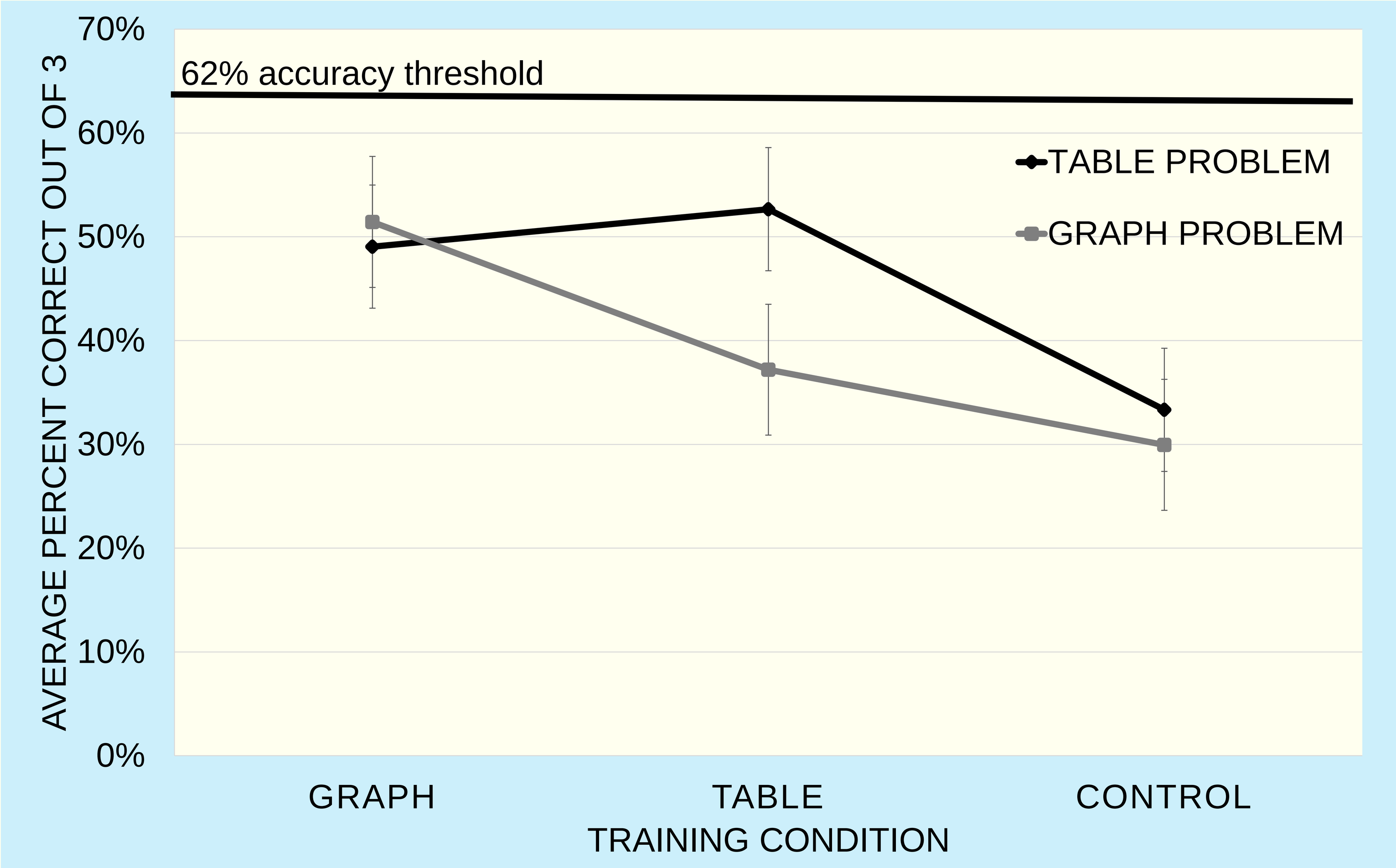

Figure 1: Interaction of Training and Problem Format η^2 = .04.
Those in graph training performed comparably on both problem formats, whereas those in the table training condition performed as well on table problems but significantly worse on graph problems. Those in the control condition performed poorly on both problem formats.

Conclusion

- Training improved performance. As predicted, only graph training benefits generalized across both formats.
- Nevertheless, accuracy rates were below levels established by previous studies.
- Numeracy, but not graph literacy, was predictive of performance.

Future Directions

- Re-configure, simplify phrasing in the inference problems.
- Evaluate participant confidence in relying on tests with low/high positive predictive values versus specificity.