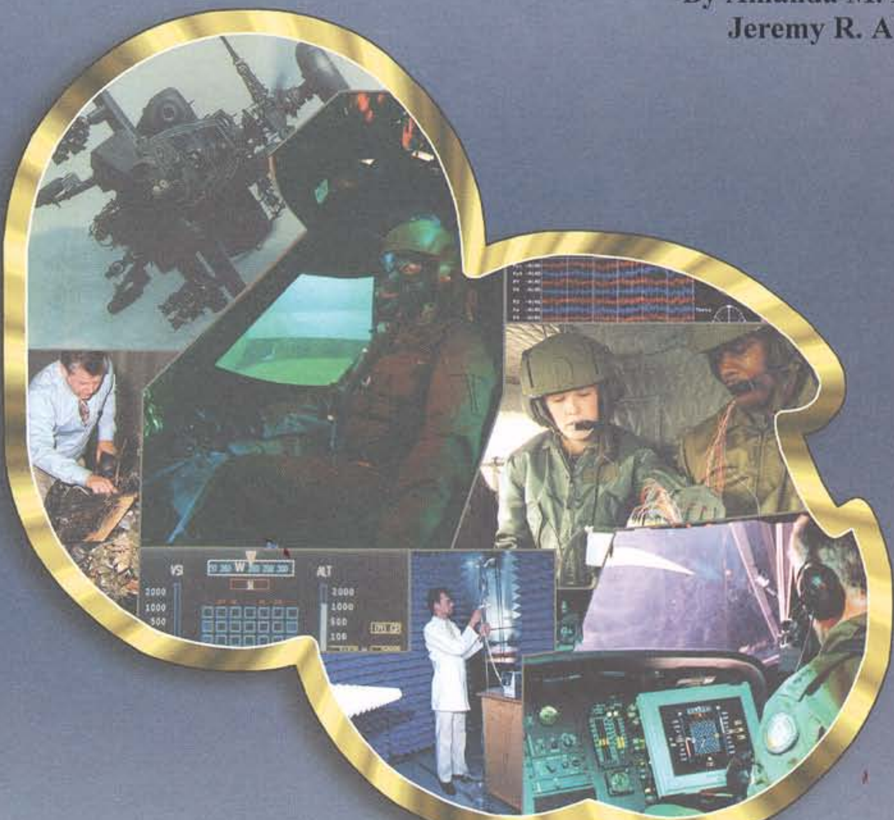
USAARL Report No. 2009-16

The Effects of Observation and Intervention on the Judgment of Causal and Correlational Relationships

By Amanda M. Kelley
Jeremy R. Athy



Warfighter Performance and Health Division

July 2009

Approved for public release, distribution unlimited.

USAARL

U.S. Army
Aeromedical Research
Laboratory

Notice

Qualified requesters

Qualified requesters may obtain copies from the Defense Technical Information Center (DTIC), Cameron Station, Alexandria, Virginia 22314. Orders will be expedited if placed through the librarian or other person designated to request documents from DTIC.

Change of address

Organizations receiving reports from the U.S. Army Aeromedical Research Laboratory on automatic mailing lists should confirm correct address when corresponding about laboratory reports.

Disposition

Destroy this document when it is no longer needed. Do not return it to the originator.

Disclaimer

The views, opinions, and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy, or decision, unless so designated by other official documentation. Citation of trade names in this report does not constitute an official Department of the Army endorsement or approval of the use of such commercial items.

Human use

Human subjects participated in these studies after giving their free and informed voluntary consent. Investigators adhered to AR 70-25 and USAMRMC Reg 70-25 on Use of Volunteers in Research.

REPORT DOCUMENTATION PAGE	*Form Approved* *OMB No. 0704-0188*

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
28-07-2009	Final	

4. TITLE AND SUBTITLE	
The Effects of Observation and Intervention on the Judgment of Causal and Correlational Relationships	5a. CONTRACT NUMBER
	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S) Amanda M. Kelley and Jeremy R. Athy	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)	8. PERFORMING ORGANIZATION REPORT NUMBER
U.S. Army Aeromedical Research Laboratory P.O. Box 620577 Fort Rucker, AL 36362-0577	USAARL 2009-16
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	**10. SPONSOR/MONITOR'S ACRONYM(S)**
U.S. Army Medical Research and Materiel Command 504 Scott Street Fort Detrick, MD 21702	USAMRMC
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release, distribution unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Recent theories of causal judgment describe it as a two-stage process involving a heuristic stage and an analytic stage. The present study evaluated discrimination of causal and correlational relationships using observation and intervention tasks. Results show that participants' causal judgments reflected the objective sample correlations in the observation tasks rather than the probabilities in the intervention tasks. This suggests that people are more sensitive to objective correlations than underlying causal probabilities.

15. SUBJECT TERMS

causal judgment, correlation detection

16. SECURITY CLASSIFICATION OF: a. REPORT	b. ABSTRACT	c. THIS PAGE	17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
					Loraine Parish St. Onge, PhD
UNCLAS	UNCLAS	UNCLAS	SAR	62	**19b. TELEPHONE NUMBER** *(Include area code)* 334-255-6906

Reset

Acknowledgements

The authors would like to express their sincere gratitude to the following people for their contributions to this project.

- Ms. Elizabeth Stokes for help with administrative matters.
- Dr. Loraine St. Onge for her editorial assistance.
- Ms. Melinda Vasbinder, Ms. Lana Milam, SGT Robert MacNeill, and SGT William Schober for assistance with the data collection.

Table of contents

Page

Table of contents (continued)

Page

Page

List of figures

List of figures (continued)

Page

List of tables

Table of contents (continued)
List of tables (continued)

Page

Introduction

Covariation detection is the ability to perceive relationships in the environment and is widely acknowledged as a fundamental aspect of everyday life (e.g., Busemeyer, 1990; Chapman, 1967; Crocker, 1981; Kareev, 2005). This ability underlies higher order processes such as determining cause and effect relationships (see Shanks, 2004). Also, covariation detection is essential for “explaining the past, controlling the present, and predicting the future” (Crocker, 1981). The literature is abundant with theories and models describing how people determine correlational (two events systematically vary together) and causal (one variable generates change in another variable) relationships. There is minimal discussion, however, of how people integrate conflicting information regarding the two relationship types.

The current study evaluated correlation detection and causal judgments using two laboratory tasks; one of which measures correlation detection and one of which is structured to elicit the causal judgment process. Previous researchers have used these tasks in isolation to study either correlation detection or causal judgment but the current study is the first, as far as the authors know, to integrate the two tasks in an effort to evaluate how correlation information is integrated with causal information. Ongoing research at the United States Army Aeromedical Research Laboratory (USAARL) at Fort Rucker, Alabama is being conducted to investigate biases and errors in causal judgment by Soldiers after periods of sleep deprivation. This work utilizes the same measures as those in the current study. Thus, the results of the current study suggested further the validity of these measures (particularly when used in tandem) and supports use of them in the ongoing work.

Military significance

The ability to detect relationships in the environment, both correlational and causal, is a fundamental cognitive process. Accuracy in detection of causal and correlational relationships in the natural environment is essential to other cognitive functions such as learning and categorization. Correlation detection plays a role in military settings such as intelligence analysis (Heuer, Jr., 1999). Likewise, the ability to identify cause and effect relationships is important in military settings. For instance, if these abilities are compromised in a combat environment, then Soldiers are more likely to make errors in calculating future events, forming accurate expectations, and, subsequently, taking appropriate precautionary actions. Our actions are based on beliefs about how things in our environment are related to each other. One way that these beliefs are formed is through correlation detection, causal judgment, and associative learning. Research and experience shows that these beliefs are very difficult to change or alter once they are formed and reinforced. Therefore the formation of beliefs is crucial. When simple heuristics, or rules used as shortcuts to aid in decision making, are used in correlation detection (typically under conditions of high cognitive load), the likelihood of the accuracy of the perception is compromised. If an inaccurate perception translates to an inaccurate belief, then the perceiver is likely to behave inappropriately and make inaccurate judgments related to that belief. For example, if a pilot is cognitively overloaded and inaccurately perceives the actual relationship between an action and an outcome, then his reaction to a display message warning of the outcome may be inaccurate resulting in potential disaster. The present study helps to identify extraneous information in making these judgments which may aid in reduction of information overload. The current study addressed the perception of correlational and causal relationships in a laboratory setting, the integration of correlation information with causal information, and set the groundwork for future studies evaluating correlation and causation detection in specific military operations and under conditions of operational stress.

Background

Models of correlation detection and causal judgment are abundant in the literature. Numerous models of causal judgment describe the process in two stages and implicate correlation detection as the first stage (e.g., Anderson & Sheu, 1995; Baker, Murphy, & Vallee-Tourangeau, 1996; Cheng, 1997). This is not to imply that correlation is necessary to determine causation but that rather can serve as an indication of a causal relationship. By a two stage process, correlation is first assessed between potential causal candidates and the observed effect in an observational format (i.e., participant views "data" and makes determination of correlational relationship). From this assessment, a set of possible candidate causes is determined and the second stage, the analytic stage, is begun. In this stage, more cognitive resources are expended and the set of possible causes becomes smaller. Limited research has been done on what processes are at work in the analytic stage.

For the purposes of the present study, discussion will be limited to a single causal candidate and single effect. In studies of correlation detection and causal judgment, typically, participants are given a cover story which presents a hypothetical situation in which the relationship between two variables must be determined. For example, a commonly employed cover story describes a hypothetical situation in which the participant must imagine that he/she is an agricultural scientist and is investigating the effectiveness of a newly determined chemical compound as a fertilizer. In order to accomplish this task, participants are next presented with a set of data points (or a sample of data). Each data point indicates whether the causal candidate and whether the effect variable is present or absent. Following the fertilizer cover story, a data point in the sample would indicate whether a plant had been sprayed with the chemical compound and whether it had grown a significant amount. In the laboratory task, this set of data points is presented either sequentially or in a list format. This data can then be summarized in a contingency table which has four cells; Cell A is the frequency of observations where the causal candidate and effect variables are both present (e.g., plant received chemical compound and grew), Cell B is the frequency of observations where the causal candidate is present and the effect is absent (e.g., plant received chemical compound and did not grow), Cell C is the frequency of observations where the causal candidate is absent and the effect is present (e.g., plant did not receive chemical compound but did grow), and Cell D is the frequency of observations where both variables are absent (e.g., plant did not receive chemical compound and did not grow; figure 1).

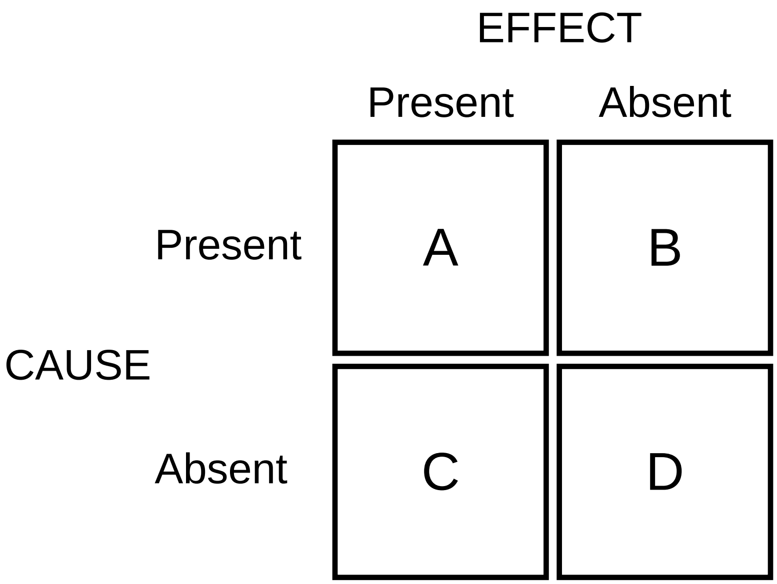


Figure 1. A contingency table. The cell entries are used to calculate the generally accepted measure of contingency between binary variables; ΔP. The cell entries A through D denote the number of observations in each cell of the contingency table.

Covariation detection

Research shows that people are very accurate at judging correlational relationships from scatterplots and raw numbers when the relationship is moderate to very strong ($\Delta P \geq 0.5$) (e.g., Doherty, Anderson, Angott, & Klopfer, 2007). The generally accepted measure of contingency is ΔP (Jenkins & Ward, 1965; Ward & Jenkins, 1965). This measure of contingency is appropriate for binary variables whereas other well known measures of correlation such as Pearson's *r* are appropriate for use with continuous variables. The expression, in terms of the cells of a contingency table, is presented in Equation 1:

$$\Delta P = (A/(A+B)) - (C/(C+D)) \qquad (1)$$

Delta P can also be expressed as the difference between the probability of the effect in the presence of the cause and the probability of the effect in the absence of the cause. The greater the difference between the two probabilities (i.e., the greater the value of ΔP), the greater the likelihood that the causal candidate produced the outcome in question. A number of models and theories have been based on ΔP such as Cheng's (1997) Power PC theory. This theory states that people's inferences about causality do not rely solely on an assessment of correlational strength, (e.g., ΔP), but rather that the observer is trying to infer a candidate's unobservable *causal power*. Rather, *causal power* refers to the probability that the causal candidate generated or inhibited the effect. Power PC defines *causal power*, *p*, in Equation 2 when the effect, *e*, is generated by cause, *c*, (positive relationship):

$$p = \Delta P / (1 - P(e|{\sim}c)) \tag{2}$$

Power PC defines *causal power*, *p*, in Equation 3 when the effect, *e*, is inhibited by cause, *c*, (negative relationship):

$$p = - \Delta P / P(e|{\sim}c) \tag{3}$$

A measure such as ΔP is insufficient for measurements of causation as it does not specify directionality (e.g., X causes Y or Y causes X).

Most recently, Hattori and Oaksford (2007) proposed the dual factor heuristic (rule) as a model of covariation detection which is expressed in Equation 4:

$$H = a/((a + b)(a + c))^{1/2} \tag{4}$$

This model differs from previous models of covariation and causal detection, such as ΔP and the Power PC theory, in that it describes covariation detection while ignoring Cell D observations (i.e., observations of the absence of both variables). A number of studies show that people weigh information from the four cells of the contingency table unequally such that A > B ≥ C > D (e.g., Kao & Wasserman, 1993; Levin et al., 1993; Mandel & Lehman, 1998; Wasserman et al., 1990). Hattori and Oaksford propose that at the heuristic stage by ignoring observations of non-occurrence (e.g., observations where the cause did not occur and the event did not occur) a judgment of covariation can be determined by an index which is equivalent to the geometric mean of the probability of the cause given the effect and the probability of the effect given the cause. Arguably, the number of observations where both variables are absent (non-occurences) is quite large given that the probability of the cause (P(C)) and the probability of the effect (P(E)) are both small. Previous research suggests that Cell D observations (observations of joint non-occurrence) are relatively less informative than Cell A observations (observations of joint occurrence) given probabilistic rarity of the presence of a variable relative to the probability of the absence of a variable (e.g., Anderson, 1990; McKenzie & Mikkelson, 2007). Limited evidence shows that Cell D observations may be perceived as most informative or "preferred" under some conditions (McKenzie & Mikkelson, 2007). In addition to testing the model fit for the dual factor heuristic, Hattori and Oaksford (2007) performed a meta-analysis including an additional 40 models of causal judgment and correlation detection comparing against data established in the literature as well as experimental data they collected. Through this analysis, they found support for their proposed dual factor heuristic model.

<u>Causal judgment</u>

As mentioned above, some models of causation are based on the normative measure of contingency, ΔP (e.g. Cheng, 1997, Cheng & Novick, 1992). This class of models will be further referred to as normative models of causation. A second type of model, which are based on Pavlovian classical conditioning, is associative learning models (e.g, Rescorla & Wagner, 1972). Finally, an emergent perspective in the literature is models which take an inferential approach (i.e., people do not summarize data but rather use data to infer the likelihood that a relationship exists). Griffiths and Tenenbaum (2005) proposed the causal support theory, an inferential model, which is based on Bayesian causal nets. This model describes how people infer causal structure (i.e., whether a relationship exists) rather than causal strength (i.e., the degree of strength between a cause and an effect). Unlike normative and associative learning models, inferential models are capable of accounting for experimental phenomena such as correlational sample indeterminacy and sample size. Kelley (2007) investigated the extent to which people were sensitive to the likelihood of a sample being drawn from a correlated versus an uncorrelated population. In this study, objective probabilities were calculated for all sample types (correlationally determinate and correlationally indeterminate) and were subsequently compared to participants' performance on a ranking task. The results showed that under most conditions, participants' behavior was reflective of the objective probabilities, thus providing further support for an inferential approach to causal judgment. An inferential approach is cooperative with arguments for a two-stage process of causal judgment such that in the first stage covariation is assessed. An inferential approach may be appropriate for further evaluation of the processes involved in the analytic stage. The present study does not focus on this assertion but is an important area for future directions.

Distinctions between covariation and causation

Of course covariation does not equal causation and distinctions between covariation detection and causal judgment are abundant in the literature. However, in some cases, it is difficult to separate the two concepts. Some causal models do not account for causal scenarios where a correlational relationship does not exist. For example, as previously mentioned, Cheng (1997) developed a normative perspective of causality, the Power PC theory. By this account, causality is determined by the base rate of the effect and the difference between the probability of the effect in the presence and absence of the causal candidate. If ΔP is equal to zero (meaning there is no correlational relationship), then the numerator is zero and causal power is zero. In other words, by this account, one cannot determine causality in the absence of a correlational relationship. In the present study, participants were asked to make judgments about covariation as well as judgments of causality using two fundamentally different tasks. They were also asked to incorporate information from both tasks into one cohesive judgment. Arguably, participants should be sensitive to the different relationships and conditions in which the absence of evidence supporting one relationship type does not negate the possibility of the other relationship type.

It is worth noting here that people, only a portion of the time, draw conclusions about causality or covariation based on observed data alone. Prior expectancies and beliefs play a large role in this process as is discussed most recently by McKenzie and Mikkelson (2007). In the present study, consideration is given to the cover stories chosen such that the variables included are causally plausible but nonspecific. The present study employed cover stories typically used in the covariation detection and causal judgment literature including fertilizers and plant growth; gene expression and physical traits; drug administration and symptom relief; and treatment and weight loss.

Sample correlational relationships

Typically, studies of covariation detection and causal judgment focus on samples in which the relationship is positive/generative or zero. Relatively few studies incorporate samples with negative/preventative relationships and even fewer include samples with indeterminate correlational relationships (e.g., Griffiths & Tenenbaum, 2005; Kelley, 2007; Kelley, Anderson, & Doherty, 2007; White, 2000). For the present purposes, indeterminate correlational relationships result when one variable does not vary. The present study incorporated all four types of correlational relationships (positive, negative, zero, and indeterminate).

Paradigms

The distinction between the heuristic and analytic stages in causal judgment has been paralleled with the distinction between *observation* (discrete) and *intervention* (continuous) tasks (Hattori & Oaksford, 2007). Specifically, an *observation* task presents the participant with a summary of data from which they are to make a judgment about the relationship, typically assigning a rating of the relationship between 0 (no relationship) and 100 (perfect relationship). Since the judgment is solely based on observations, then only the heuristic stage of the process should be activated. In an *intervention* task, participants are told that they can either administer the causal candidate or not, and then observe whether the effect occurs. The occurrence/nonoccurrence of the effect is determined by the probabilities set by the experimenter. Hattori and Oaksford (2007, p.768) hypothesized that the active role of the participant thus engages the analytic stage of the process and the act of "intervening between events that covary is essential to learning whether one of these events is the cause of the other." Recent research shows support for these distinctions (e.g., Hattori & Oaksford, 2007; Lagnado & Sloman, 2004; Steyvers, Tenenbaum, Wagenmakers, & Blum, 2003).

Research questions

While correlational and causal relationships may overlap, one type of relationship does not imply the other. Research suggests that causal judgment is a two-stage process that begins with correlation perception followed by an analytic stage resulting in a judgment. However, it is unclear how people integrate conflicting correlational and causal information. To what extent are judgments relatively influenced by these two types of information? In addition to these questions, the present study explored causal judgment in relation to negative or preventative relationships as well as correlationally indeterminate relationships.

Research objective and hypotheses

In the present study (composed of 3 experiments), participants made causal and covariation judgments in observation and intervention tasks. The broad scope of the set of experiments was to evaluate the extent to which judgments were influenced by observed correlational information and generated causal information. In other words, the present set of experiments focused on the relative influence of two types of information (observed and generated) on causal judgments. Correlational relationships are relationships where the occurrence/non-occurrence of one factor changes systematically with the occurrence/non-occurrence of another factor (e.g., as variable x increases, variable y increases) and causal relationships are relationships where one factor is used to predict, explain or describe the occurrence/non-occurrence of an event or factor (e.g., the occurrence of variable x causes the occurrence of variable y). There were three main study objectives:

Objective 1

To evaluate the extent to which causal judgments were influenced by observed correlational information and generated causal information using positive/generative sample relationships and uncorrelated (zero correlation) samples (experiment 1), negative/preventative sample relationships and uncorrelated (zero correlation) samples (experiment 2), and indeterminate sample correlational relationships and uncorrelated (zero correlation) samples (experiment 3).

Objective 2

To determine whether people select samples with an indeterminate correlational relationship and if the judgments based on those self-selected samples were reflective of the probability of the sample being drawn from a correlated versus an uncorrelated population (experiment 3).

Objective 3

To evaluate the extent to which final judgments of causality were reflective of observed correlational information and actively generated causal information under conditions when the two types of information are conflicting and when they are in agreement (experiments 1, 2, and 3).

Hypothesis 1

Given that research shows people are accurate at detecting covariation when the relationship is moderate to strong, it was expected that participants would accurately rate the sample correlational relationship in the observation task. In the intervention task, it was expected that participants' causal relationship ratings would reflect the defined causal probabilities rather than the objective correlational value of the data they generated. If observation tasks elicited the heuristic stage of causal judgment, then participants' judgment should reflect the correlational relationship in the observed sample. If intervention tasks elicited the analytic stage of causal judgment, then judgments should reflect the probability of the effect in the presence of the cause and the probability of the effect in the absence of the cause. In other words, it was predicted that correlational relationship ratings would reflect the objective sample correlations in the observation task and the causal relationship ratings would reflect the probabilities set in the intervention tasks. This hypothesis was the same for both positive/generative relationships and negative/preventative relationships.

For the observation task, it was predicted that the mean ratings would reflect the objective correlational value of the sample presented.

For the intervention task, it was predicted that the mean ratings would reflect the objective causal probabilities for the data generated.

Hypothesis 2

When presented with samples with an indeterminate correlational relationship, participants' responses would reflect that the probability of the sample being drawn from a correlated population is greater than that from an uncorrelated population. Participants would be able to make a causal judgment reflective of the causal probabilities in the absence of defined sample correlation data and those judgments would be greater for indeterminate samples than those for uncorrelated samples.

Hypothesis 3

For purposes of clarity, trials where the observation task and intervention task parameters indicate the same relationship structure will be further referred to as congruent trials. Trials where the observation task and intervention task parameters do not indicate the same relationship structure will be referred to as incongruent trials. For congruent trials where the structure was no relationship, the final rating would be close to zero (indicating no relationship between the variables). For congruent trials where the structure was a relationship, the final rating would be close to 10 (indicating a perfect relationship between the variables). It was suspected that on incongruent trials, participants' final ratings would reflect the parameters set in the intervention task rather than the observation task.

Methods

General

The study protocol was approved in advance by USAARL's Human Use Committee (HUC). The study attempted to evaluate the extent to which causal judgments were influenced by observed correlational information and generated causal information as evidenced by the integration of the two information types. The study employed a 2 (*observation task correlational relationship* ΔP *value*) X 2 (*intervention task causal probabilities*) within-subjects design thus yielding 4 conditions. Note that the independent variables are italicized in this document for purposes of clarity. Participants completed 4 trials of each condition resulting in a total of 16 trials. Three experiments were implemented to evaluate the research questions, namely positive correlated and uncorrelated relationships (experiment 1), negatively correlated and uncorrelated relationships (experiment 2), and indeterminate and uncorrelated relationships (experiment 3). The conditions for experiments 1, 2, and 3 are presented in tables 1, 2, and 3, respectively. A limited number of studies have included negative, preventative relationships and even fewer studies have focused on indeterminate correlational relationships; thus the inclusion of these aspects enriches the scope of this study. The main difference between the three experiments was the levels of the independent variables. Specifically, for each experiment, the independent variables were the same, however, the levels of those variables changed. There were three main rationales for this separation. First, if one experiment were to incorporate positive, negative, zero, and indeterminate correlational relationships in a 4 (sample correlation type) X 3 (causal probabilities) repeated measures design, then the number of trials to be completed per participant would increase such that the time per test session may be greater than one hour potentially limiting the number of available participants as well as introducing potential fatigue effects. Second, a 4 X 3 repeated measures design would result in a large number of comparisons, most of which would not be hypothesis driven. Finally, such a design would result in 12 conditions, three of which would be classified as "congruent conditions" and nine of which would be classified as "incongruent conditions." Trials in which the level of *observation task* Δ*P value* and the level of *intervention task causal probabilities* are complementary and were considered "congruent" and trials in which the levels are contradictory were considered "incongruent." However, the important note here is the resulting inequality of conditions per classification or rather an imbalanced design. Thus, a 2 X 2 repeated measures design in three separate experiments was ideal for this study.

Table 1

Levels of the two independent variables in experiment 1 – Positive/generative relationships

	Relationship Type	
	Positive/Generative	Zero/No Relationship
Observation task ΔP value	0.5	0.0
Causal Probabilities		
P(E|C)	0.75	0.5
P(E|~C)	0.25	0.5

Note. The effect variable is labeled "E" and the causal candidate is labeled "C."

Table 2

Levels of the two independent variables in experiment 2 – Negative/preventative relationships

	Relationship Type	
	Negative/Preventative	Zero/No Relationship
Observation task ΔP value	-0.5	0.0
Causal Probabilities		
P(E|C)	0.25	0.5
P(E|~C)	0.75	0.5

Note. The effect variable is labeled "E" and the causal candidate is labeled "C."

Table 3

Levels of the two independent variables in experiment 3 – Indeterminate correlational relationships

	Relationship Type	
	Indeterminate Correlation	Zero/No Relationship
Observation task ΔP value	**	0.0
Causal Probabilities		
P(E|C)	0.75	0.5
P(E|~C)	0.25	0.5

Note. ** indicates an indeterminate correlational relationship such that the causal variable does not vary, thus the equation is undefined. The effect variable is labeled "E" and the causal candidate is labeled "C."

Participants

Participants were 60 United States Army Soldiers (20 per experiment). The mean age was 24.86 years and the mean education level was 15.71 years (e.g., 12 years = high school diploma). Participants with an advanced degree or a bachelor's degree in statistics were not eligible for participation in the study. Given the nature of the tasks in the present study, it is possible that potential participants with above average experience or knowledge about statistics may perform differently than those without additional experience (e.g., Doherty, Anderson, Kelley, & Albert, 2006). Volunteers did not receive any compensation for participation.

Procedure

Upon entering the laboratory, participants were briefed. After written consent was given, participants completed the Shipley Institute of Living Scale, to determine eligibility. The Shipley Institute of Living Scale is a widely used measure of general cognitive ability and participants who scored above the 95th percentile were to be excluded from the study given that research shows that people of superior intelligence perform differently on these tasks than those of average of intelligence (Stanovich, 1999). No participants were excluded based on their Shipley scores. Next, participants completed a demographics questionnaire on the computer after which the experiment began. Both the experiment and questionnaire were completed using Psychology Software Tools' experiment generator software package, E-prime 2.0. In each trial, participants were presented with a cover story describing the variables in question. Trial configurations are included in table 4. The instructions explained that the participant was to evaluate the relationships between the causal variable and the effect variable. First, a sample of data already "collected" was presented (observation task). After viewing the sample, the participant rated the relationship between the causal candidate and effect variables on a scale from 0 (no relationship) to 10 (perfect relationship). Next, the participant was told that to "collect" a sample of data. For eight observations, the participant chose to either administer the causal candidate, C, or not and subsequently observed the occurrence or non-occurrence of the effect, E, (intervention task). The participant was then prompted to give another set of ratings. At the end of the trial, the participant was asked to incorporate all the information presented and make a final forced choice binary judgment, a *recommendation response*, (e.g., “Given all the information you have seen and generated, do you recommend Chemical A as an effective plant fertilizer? 1. Yes 2. No”). Lastly, the participant was presented three options to justify the final assessment (*recommendation response)* of the causal candidate and effect variable, a *justification response*, (e..g, “Why did you decide to recommend or not recommend Chemical A? 1. Chemical A inhibits plant growth. 2. Chemical A doesn’t affect plant growth. 3. Chemical A enhances plant growth.). The participant gave a rating of his/her confidence level with each assessment (see Appendix A). The length of the test sessions ranged from 20 minutes to 45 minutes due to participant variability.

Table 4

Trial configurations including the experiment, correlational relationship ΔP value (corr. sample), the probability of the sample being drawn from correlated (P(Corr.)) and uncorrelated (P(Uncorr.)) populations and the observations in cells A, B, C, and D

Experiment	Corr. Sample	P(Corr.)	P(Uncorr.)	A	B	C	D
1	0.0	0.013	0.037	2	2	2	2
1	0.5	0.06	0.018	3	1	1	3
2	0.0	0.013	0.037	2	2	2	2
2	-0.5	0.001	0.018	1	3	3	1
3	0.0	0.013	0.037	2	2	2	2
3	Indeterminate	0.001	0.0003	6	2	0	0

Note. The probabilities for the indeterminate sample are extremely low because the probability of drawing an indeterminate sample is low (approximately 0.01; Kelley, 2007).

There were four conditions and participants completed four trials per condition; thus a total of 16 trials were complete. Four different cover stories were used. Participants were presented with one trial per condition per cover story. Overall, participants saw four trials per condition with each trial using a different cover story (see table 5). Examples of four of the 16 trials are provided in appendix A. The order was randomized and the letter assigned to label the causal candidate variable was arbitrary. All cover stories used have been also used in published studies in this field of study (see Shanks, 2004 for a review). It is important to note that causal relationships are unidirectional.

Table 5

All Cover Stories and Descriptive Variables Employed in the Present Study

Condition	Cover Story	Causal Candidate	Effect
1	Fertilizer	Chemical A	Plant Growth
2	Fertilizer	Chemical B	Plant Growth
3	Fertilizer	Chemical C	Plant Growth
4	Fertilizer	Chemical D	Plant Growth
1	Food Allergy	Food A	Reaction
2	Food Allergy	Food B	Reaction
3	Food Allergy	Food C	Reaction
4	Food Allergy	Food D	Reaction
1	Experimental Drug	Drug A	Pain Relief
2	Experimental Drug	Drug B	Pain Relief
3	Experimental Drug	Drug C	Pain Relief
4	Experimental Drug	Drug D	Pain Relief
1	Diet Plan	Plan A	Weight Loss
2	Diet Plan	Plan B	Weight Loss
3	Diet Plan	Plan C	Weight Loss
4	Diet Plan	Plan D	Weight Loss

Results

The data were analyzed using 2 (*observation task* ΔP *value*) X 2 (*intervention task causal probabilities*) repeated measures analyses of variance (ANOVAs). The ANOVAs were run for each experiment, for each dependent variable of interest (e.g., observation task relationship ratings, intervention task relationship ratings, final recommendation responses, justification responses, and respective confidence ratings). Subsequent paired sample *t*-tests were conducted for further comparisons when the ANOVA revealed significant effects. A multiple regression analysis was conducted to assess the predictive power of objective sample correlation and intervention task characteristics on the final relationship ratings. The statistical software package SPSS version 17.0 was used to conduct the analyses.

Experiment 1: Positively correlated and uncorrelated relationships

Observation task

Judgment accuracy in the observation task was assessed using a repeated measures ANOVA with *observation task* ΔP *value* set as the independent variable ($\Delta P = 0.0$ or $\Delta P = 0.5$) and relationship rating as the dependent measure. The analysis revealed a main effect of ΔP such that positively correlated samples were rated as higher than uncorrelated samples, $F(1, 19) = 50.64$, $p < .001$ (figure 2). The repeated measures ANOVA with confidence rating set as the dependent measure was not significant, thus indicating that the difference is ratings is not a reflection of confidence levels.

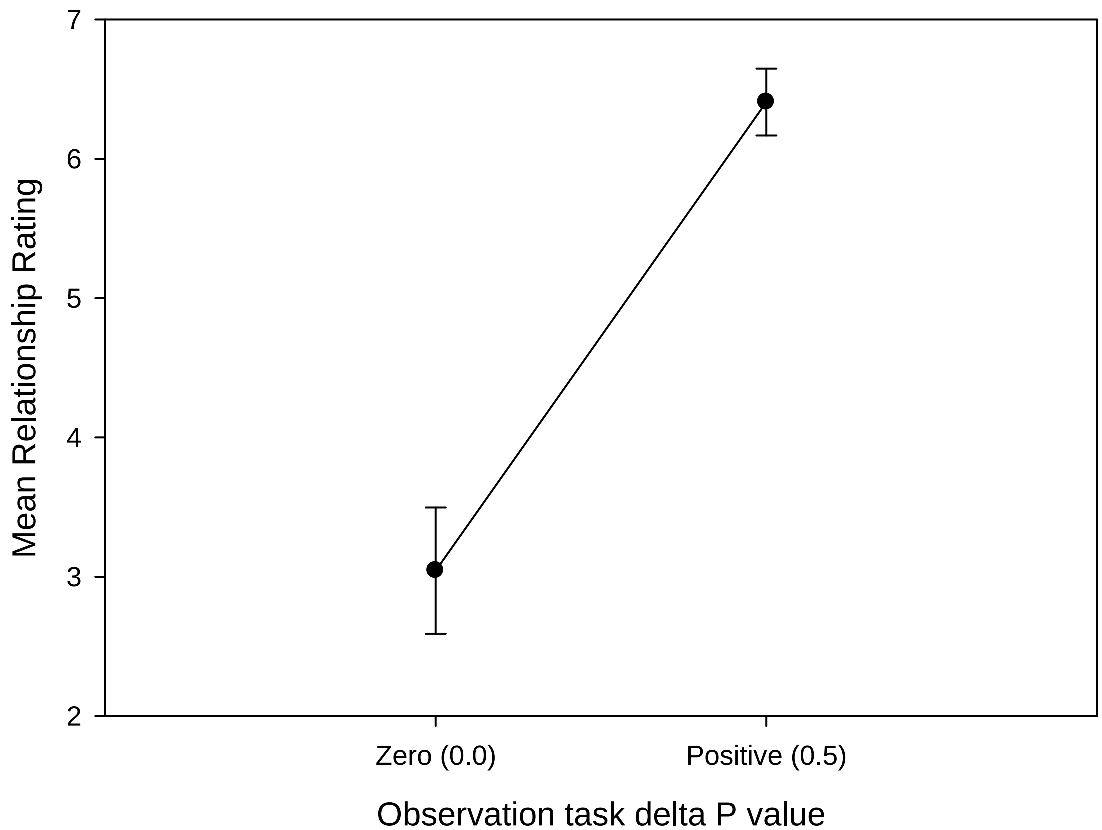


Figure 2. Observation task results: Mean relationship ratings by *observation task ΔP value* (0.5 and 0.0)

<u>Intervention task</u>

Judgment accuracy in the intervention task was assessed using a repeated measures ANOVA with *causal probability* set as the independent variable (($P(E|C) = .75$, $P(E|C) = .5$) which was non-significant. Given this result, the samples generated by the participants in the intervention task were assessed. Across all participants, 320 samples were generated of which 128 were correlationally indeterminate. Specifically, participants did not vary in their behavior choice (e.g., choose to administer drug on each trial) thus the level of the causal candidate (i.e., applied/not applied) did not vary, rendering the correlational relationship indeterminate. However, for purposes of analysis, samples were categorized with respect to the ratio of Cell A observations to Cell B observations (see figure 1), when the participant choose to apply the causal candidate on each observation, such that if the ratio was greater than one then the sample was categorized as positive, equal to one then the sample was no relationship, and less than one then the sample was negative. When the participant choose not to apply the causal candidate on each observation the generated samples were categorized with respect to the ratio of Cell C to Cell D observations, such that if the ratio was less than one then the sample was categorized as positive, equal to one then no relationship, and greater than one as negative. Further, this independent variable will be referred to as the *generated sample relationship type* which has three levels; positive, zero, and negative. A repeated measures ANOVA was run (one participant was excluded for incomplete data) setting the *generated sample relationship type* as the independent variable. This analysis showed a significant main effect, $F(2, 36) = 26.15$, $p < .001$. Paired samples t-tests revealed that when the generated sample indicated a positive relationship, the relationship was judged to be stronger than when it was zero or negative (figure 3; table 6). Confidence ratings were also analyzed in the same manner and revealed non-significant effects.

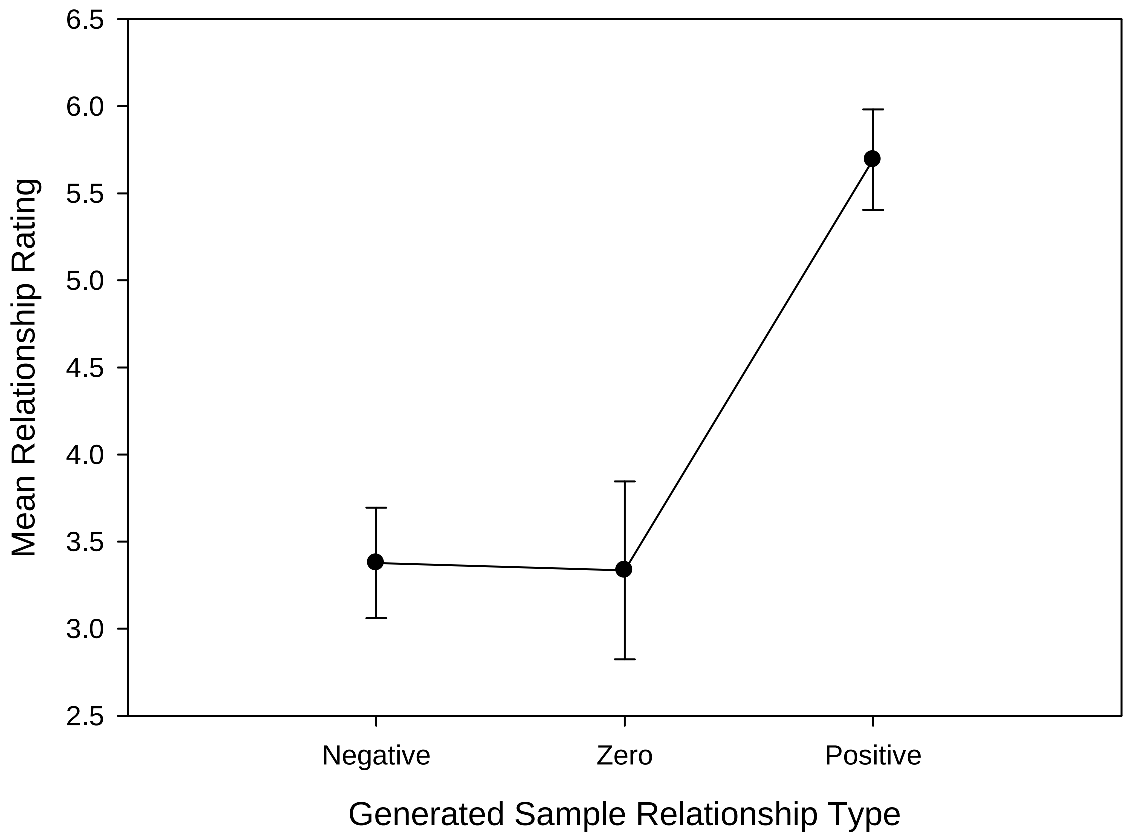


Figure 3. Intervention task results: Mean relationship ratings by *generated sample relationship types (positive, negative, zero)*

Table 6

Results of paired comparison *t*-tests for intervention task mean relationship ratings by *generated sample relationship types*

Paired Comparison	*N*	*t* value	*p* level
Positive Relationship Type – Negative Relationship Type	20	8.358	< .001*
Positive Relationship Type – No Relationship (Zero) Type	19	5.137	< .001*
Negative Relationship Type – No Relationship (Zero) Type	19	0.122	= .905

* significant

Final recommendation choice and justification

The final recommendation choices were analyzed using a 2 (*observation task* Δ*P value*) X 2 (*intervention task causal probabilities*) repeated measures ANOVA which revealed a significant main effect of *observation task* Δ*P value*, $F(1, 19) = 38.00$, $p < .001$, and a marginally significant main effect of *intervention task causal probabilities*, $F(1, 19) = 3.97$, $p = .061$ (figure 4, table 7). Given that no significant effect of *causal probabilities* was found in the analysis of the intervention task ratings, final recommendation choices were also analyzed with respect to *generated sample relationship type* as described above. It should be noted that six participants were excluded from the ANOVA for incomplete data (i.e., given that the task requires participants to generate their own samples, experimental control is sacrificed, thus some participants did not generate each possible trial type in the intervention task resulting in incomplete data). The 2 (*observation task* Δ*P value*) X 3 (*generated sample relationship type*) repeated measures ANOVA revealed a significant main effect of *observation task* Δ*P value*, $F(1, 13) = 25.18$, $p < .001$ and a significant interaction, $F(2, 26) = 7.53$, $p = .003$ (figure 5, table 8). Given the large number of comparisons, a Bonferroni correction was applied (corrected $\alpha = .05/15 = .003$).

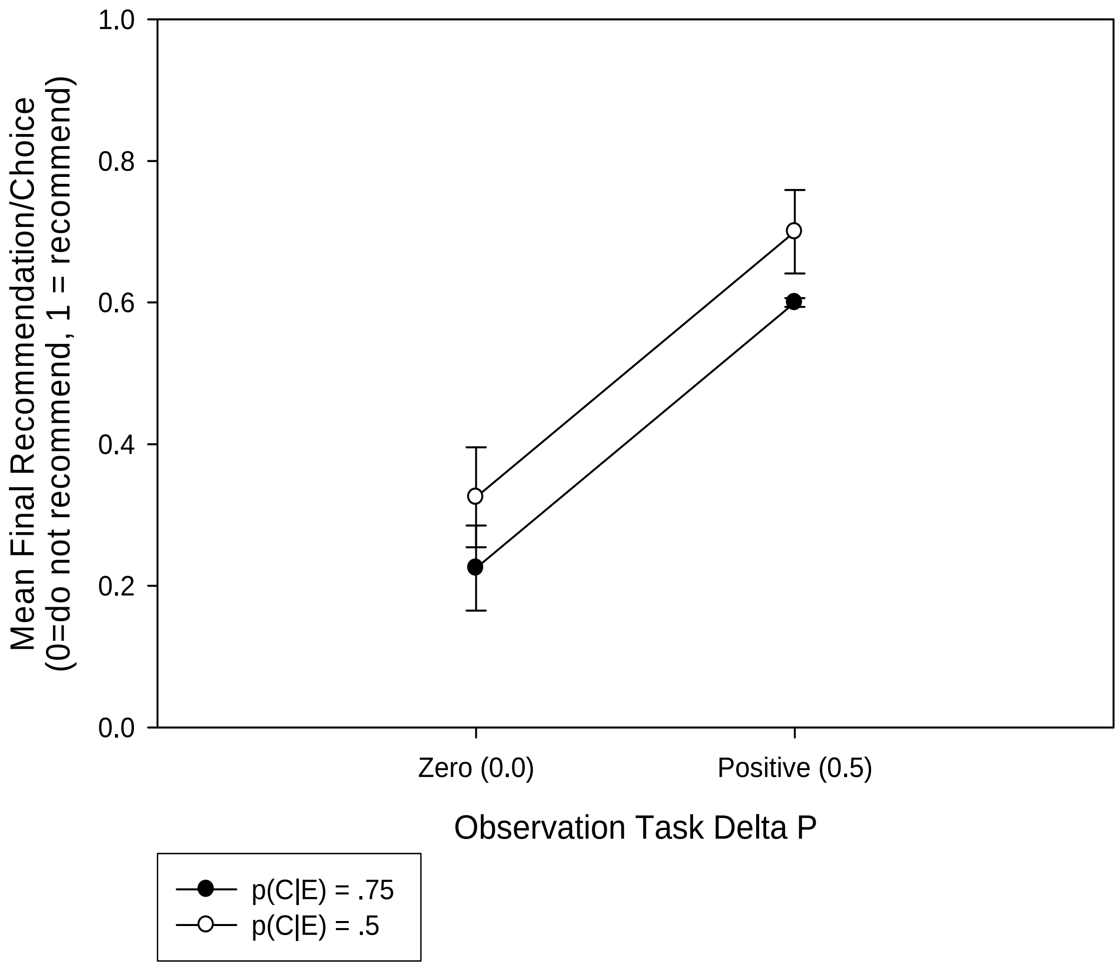


Figure 4. Final recommendation results: Mean recommendations by *observation task ΔP value (0.0 or 0.5)* and *intervention task causal probabilities (0.75 or 0.5)*

Table 7

Results of paired comparison *t*-tests for mean final recommendation choices by *observation task ΔP value (0.0 or 0.5)* and *intervention task causal probabilities (0.75 or 0.5)*

Paired Comparison (ΔP value, p(C|E) value)	*N*	*t* value	*p* level
(0.0, 0.5) – (0.0, 0.75)	20	-1.71	= .104
(0.0, 0.5) – (0.5, 0.5)	20	4.18	= .001*
(0.0, 0.5) – (0.5, 0.75)	20	4.22	< .001*
(0.0, 0.75) – (0.5, 0.5)	20	5.25	< .001*
(0.0, 0.5) – (0.5, 0.75)	20	6.10	< .001*
(0.5, 0.5) – (0.5, 0.75)	20	-1.29	= .214

* significant

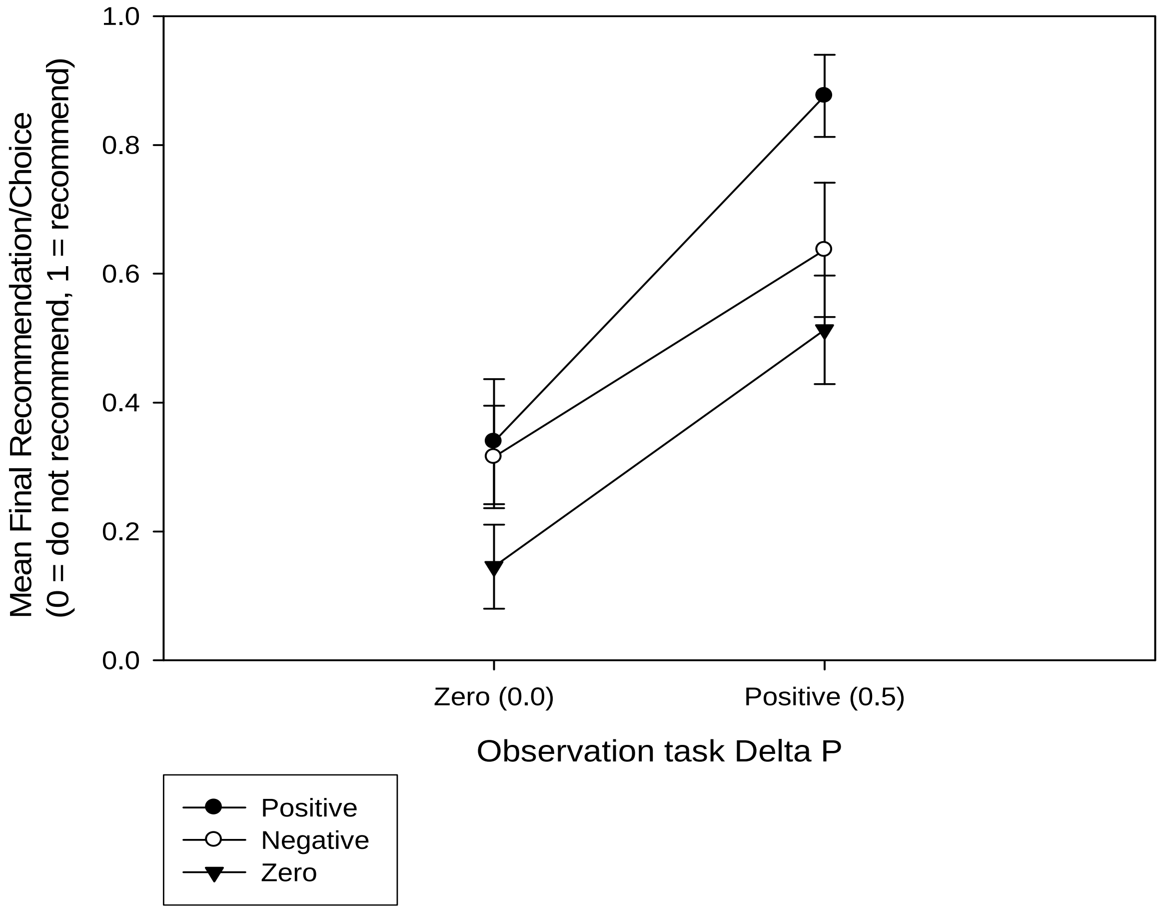


Figure 5. Final recommendation results: mean recommendation responses by *observation task* Δ*P value (0.0, 0.5)* and *generated sample relationship type (positive, negative, zero)*

<u>Table 8</u>

Results of paired comparison *t*-tests for mean final recommendation choices by *observation task ΔP value (0.0 or 0.5)* and *generated sample relationship type (positive, negative, zero)*

Paired Comparison (ΔP value, relationship type)	*N*	*t* value	*p* level
(0.0, negative) – (0.0, zero)	17	-2.56	= .021
(0.0, negative) – (0.0, positive)	20	-3.86	= .001*
(0.0, negative) – (0.5, positive)	20	-10.45	< .001*
(0.0, negative) – (0.5, zero)	16	-4.22	= .001*
(0.0, negative) – (0.5, negative)	20	-2.80	= .012
(0.0, zero) – (0.0, positive)	17	-0.50	= .623
(0.0, zero) – (0.5, positive)	17	-4.85	< .001*
(0.0, zero) – (0.5, zero)	14	-2.75	= .016
(0.0, zero) – (0.5, negative)	17	-1.26	= .226
(0.0, positive) – (0.5, positive)	20	-4.40	< .001*
(0.0, positive) – (0.0, negative)	20	-0.11	= .915
(0.0, positive) – (0.5, zero)	16	-2.45	= .027
(0.5, positive) – (0.5, zero)	16	1.99	= .064
(0.5, positive) – (0.5, negative)	20	4.63	< .001*
(0.5, zero) – (0.5, negative)	16	1.32	= .208

* significant

Participants' final judgment justifications were analyzed using a 2 (*observation task ΔP value*) X 2 (*intervention task causal probabilities*) repeated measures ANOVA which showed no significant main effects or interaction. Final recommendation choices were also analyzed with respect to *generated sample relationship type* as described above. Six participants were excluded from the ANOVA for incomplete data. The 2 (*observation task ΔP value*) X 3 (*generated sample relationship type*) repeated measure ANOVA revealed a significant main effect of *observation task ΔP value*, $F(1, 13) = 19.73$, $p = .001$; and a significant main effect of *generated sample relationship type*, $F(2, 26) = 14.08$, $p < .001$ (figure 6; table 9). Given the large number of comparisons, a Bonferroni correction was applied (corrected $\alpha = .05/15 = .003$).

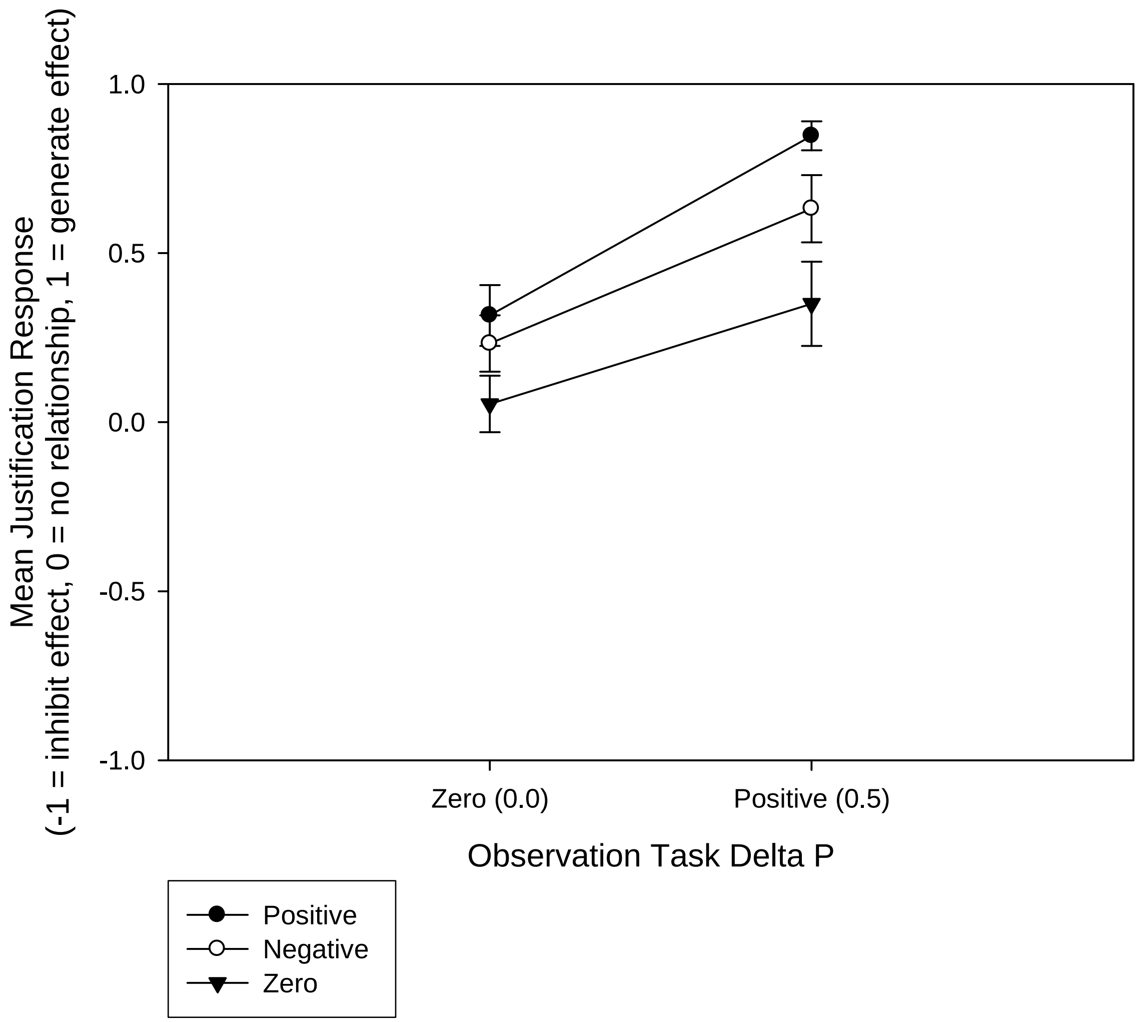


Figure 6. Final justification results: Mean justification responses by *observation task* *ΔP value (0.0, 0.5)* and *generated sample relationship type (positive, negative, zero)*

Table 9

Results of paired comparison *t*-tests for mean justification responses by *observation task* Δ*P value (0.0 or 0.5)* and *generated sample relationship type (positive, negative, zero)*

Paired Comparison (ΔP value, relationship type)	*N*	*t* value	*p* level
(0.0, negative) – (0.0, zero)	17	-1.75	= .098
(0.0, negative) – (0.0, positive)	20	-3.58	= .002*
(0.0, negative) – (0.5, negative)	20	-1.45	= .165
(0.0, negative) – (0.5, zero)	16	-4.50	< .001*
(0.0, negative) – (0.5, positive)	20	-7.13	< .001*
(0.0, zero) – (0.0, positive)	17	-1.66	= .117
(0.0, zero) – (0.5, negative)	17	-0.76	= .459
(0.0, zero) – (0.5, zero)	14	-2.82	= .014
(0.0, zero) – (0.5, positive)	17	-7.84	< .001*
(0.0, positive) – (0.5, negative)	20	1.18	= .252
(0.0, positive) – (0.5, zero)	16	-2.73	= .015
(0.0, positive) – (0.5, positive)	20	-3.56	= .002*
(0.5, negative) – (0.5, zero)	16	-2.20	= .044
(0.5, negative) – (0.5, positive)	20	-4.68	< .001*
(0.5, zero) – (0.5, positive)	16	-1.05	= .311

* significant

Effect of scenario/cover story

The potential effect of scenario/cover story was assessed using a repeated measures ANOVA with scenario (levels of which were fertilizer, food allergy, experimental drug, and diet plan) set as the independent variable and observation task rating set as the dependent variable. The analysis revealed non-significant effects, $F(3, 57) = .048$, $p = .986$.

Experiment 2: Negatively correlated and uncorrelated relationships

Observation task

Judgment accuracy in the observation task was assessed using a repeated measures ANOVA with *observation task* Δ*P value* set as the independent variable (ΔP = 0.0 or ΔP = -0.5) and relationship rating as the dependent measure. The analysis revealed no effect of ΔP such that negatively correlated samples were not rated differently than uncorrelated samples (figure 7).

The repeated measures ANOVA, with confidence rating set as the dependent measure, was also not significant.

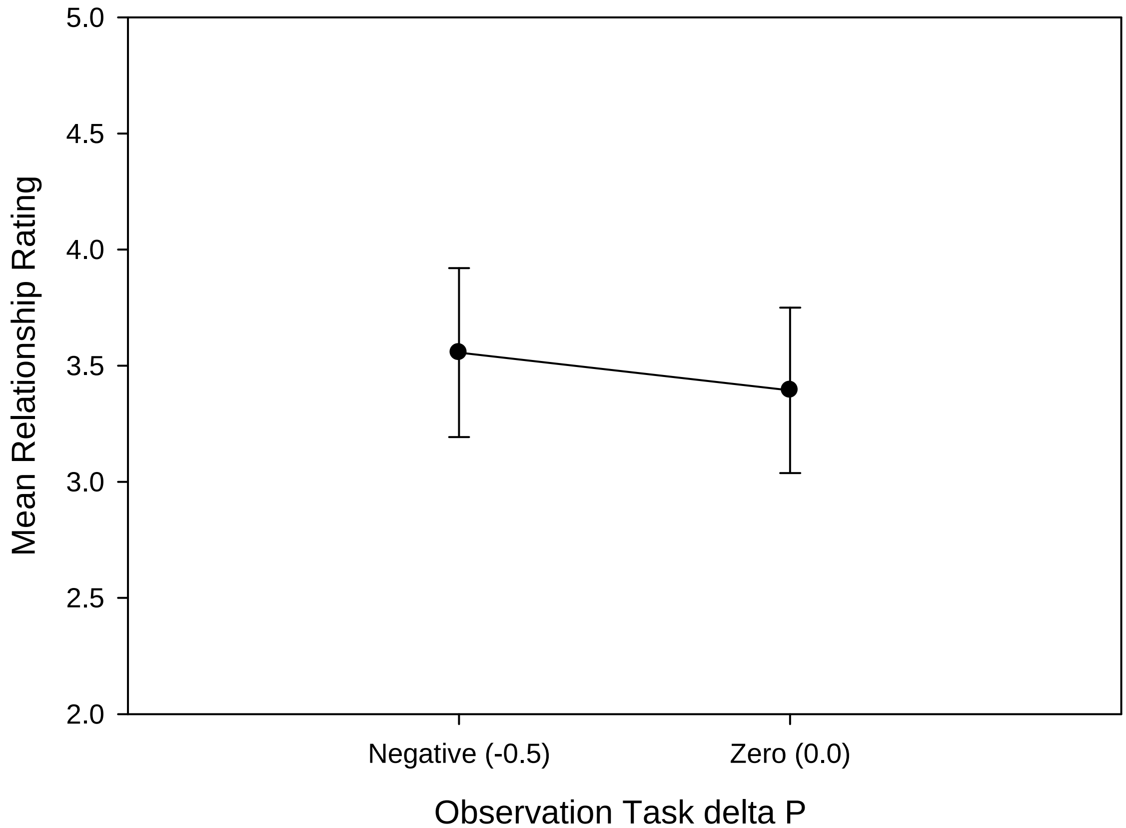


Figure 7. Observation task results: mean relationship ratings by *observation task* ΔP *value* (-0.5 and 0.0)

<u>Intervention task</u>

Judgment accuracy in the intervention task was assessed using a repeated measures ANOVA with *causal probability* set as the independent variable (($P(E|C) = .25$ and $P(E|\sim C) = .75$, $P(E|C) = .5$ and $P(E|\sim C) = .5$) which was non-significant. Given this result, the samples generated by the participants in the intervention task were assessed. Across all participants, 320 samples were generated of which 149 were correlationally indeterminate. A repeated measures ANOVA was run setting the *generated sample relationship type* (categorization method described above) as the independent variable and this analysis showed a significant main effect such that when the generated sample indicated a positive relationship, the relationship was judged to be stronger than when it was zero or negative, $F(2, 38) = 7.88$, $p = .001$ (figure 8). The results of paired comparison *t*-tests are presented in table 10. Confidence ratings were also analyzed in the same manner and revealed non-significant effects.

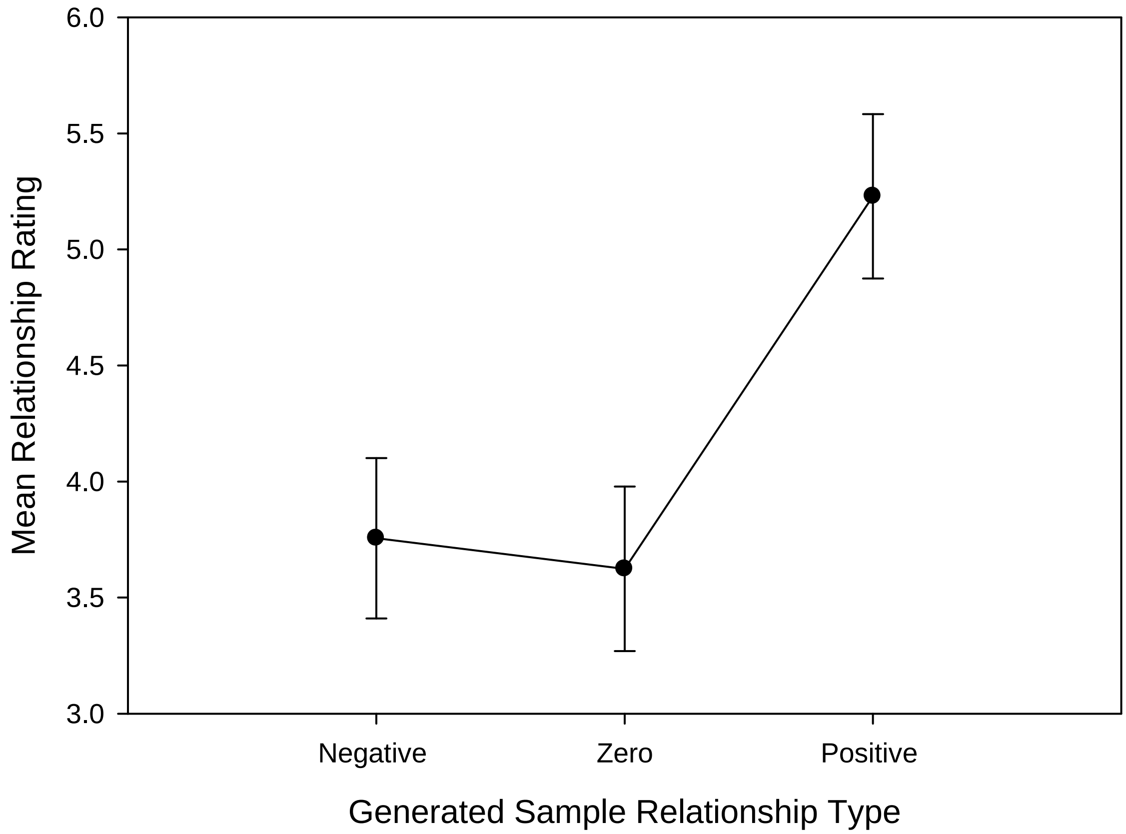


Figure 8. Intervention task results: Mean relationship ratings by *generated sample relationship types (positive, negative, zero)*

<u>Table 10</u>

Results of paired comparison *t*-tests for intervention task mean relationship ratings by *generated sample relationship types (positive, negative, zero)*

Paired Comparison	*N*	*t* value	*p* level
Positive Relationship Type – Negative Relationship Type	20	3.232	= .004*
Positive Relationship Type – No Relationship (Zero) Type	20	3.542	= .002*
Negative Relationship Type – No Relationship (Zero) Type	20	-0.302	= .766

* significant

<u>Final recommendation choice and justification</u>

The final recommendation choices were analyzed using a 2 (*observation task* ΔP *value*) X 2 (*intervention task causal probabilities*) repeated measures ANOVA which revealed no significant effects. Given that no significant effect of *causal probabilities* was found in the analysis of the intervention task ratings, final recommendation choices were also analyzed with respect to *generated sample relationship type* as described above. It should be noted that five participants were excluded from the ANOVA for incomplete data (e.g., participant did not have a mean value for each possible condition). The 2 (*observation task* ΔP *value*) x 3 (*generated sample relationship type*) repeated measures ANOVA revealed a significant main effect of *generated sample relationship type*, $F(2, 28) = 4.22$, $p = .025$ (figure 9). No other effects or interactions were significant. Given the large number of comparisons, a Bonferroni correction was applied (corrected $\alpha = .05/15 = .003$). None of the paired comparison *t*-tests were significant at the Bonferroni corrected alpha level.

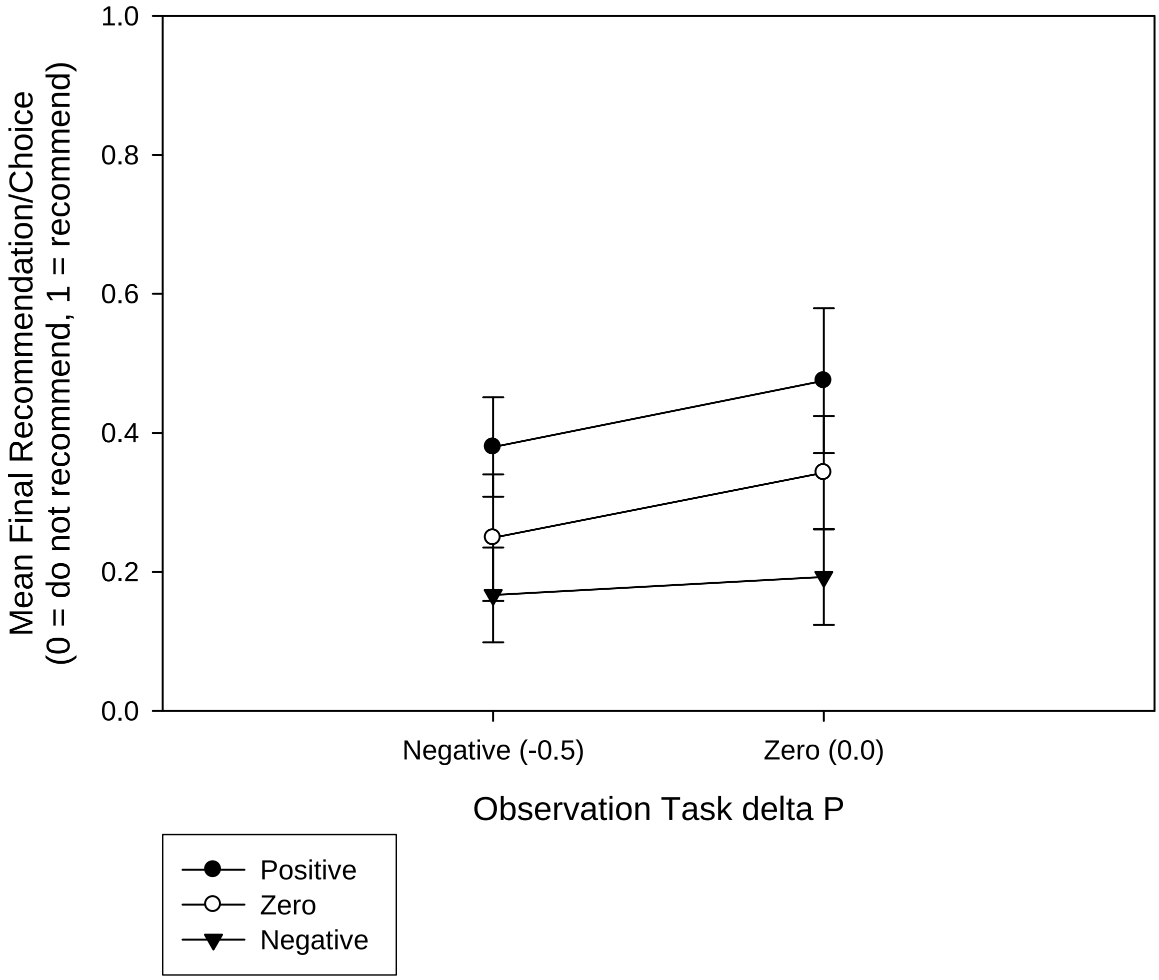


Figure 9. Final recommendation results: Mean recommendation responses by *observation task ΔP value (0.0, -0.5)* and *generated sample relationship type (positive, negative, zero)*

Participants' final judgment justifications were analyzed using a 2 (*observation task ΔP value*) X 2 (*intervention task causal probabilities*) repeated measures ANOVA which revealed a significant effect of *intervention task causal probabilities*, $F(1, 19) = 5.97$, $p = .024$ (figure 10). Paired sample t-tests are presented in table 11. Final recommendation choices were also analyzed with respect to *generated sample relationship type* as described above. Six participants were excluded from the ANOVA for incomplete data. The 2 (*observation task ΔP value*) X 3 (*generated sample relationship type*) repeated measure ANOVA revealed a significant main effect of *observation task ΔP value*, $F(1, 14) = 4.79$, $p = .046$ (figure 11). No other effects or interactions were significant. Given the large number of comparisons, a Bonferroni correction was applied (corrected $\alpha = .05/15 = .003$) see table 12 for results of the paired comparison *t*-tests.

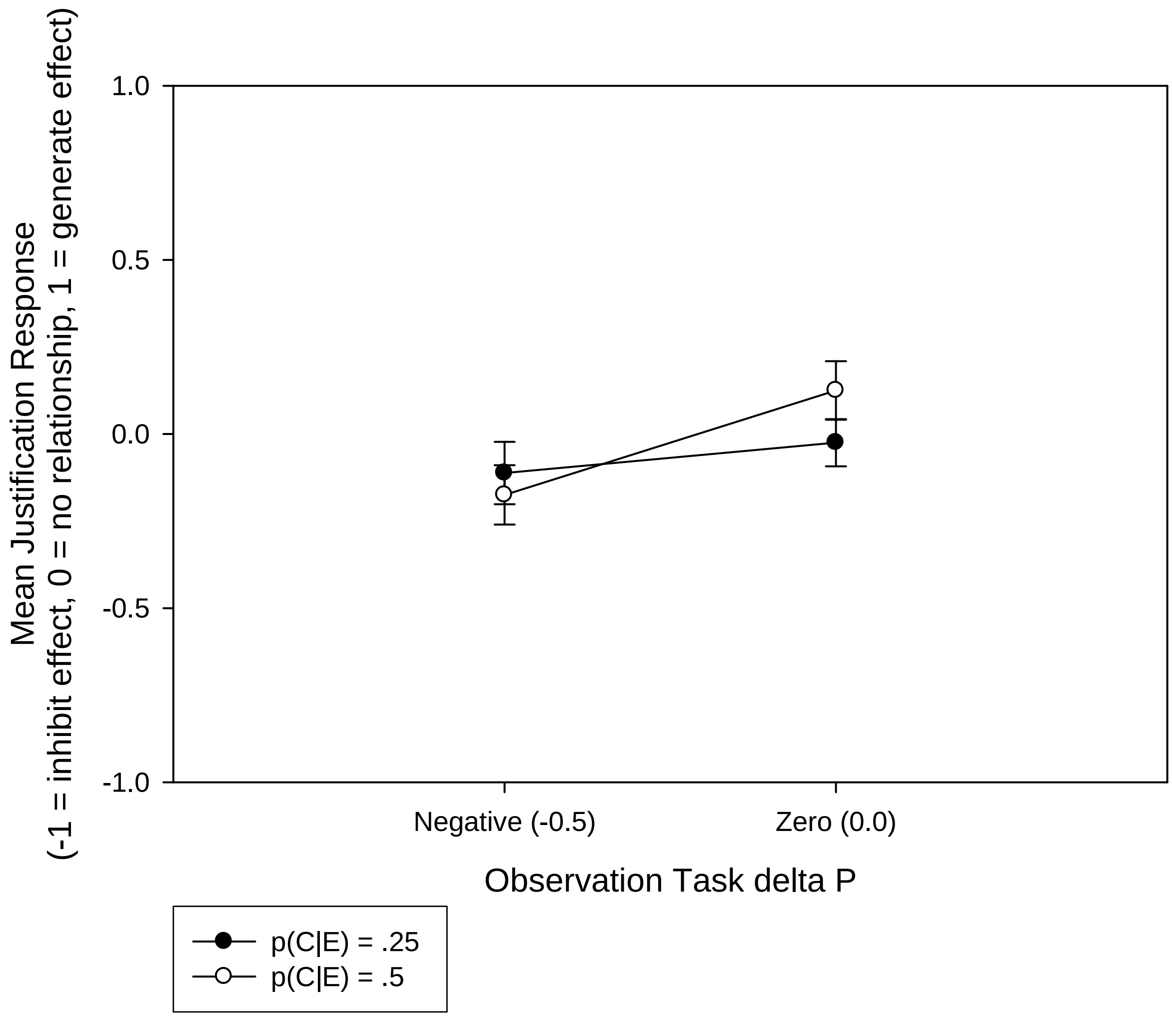


Figure 10. Final justification results: Mean justification responses by *observation task* *ΔP value (0.0, - 0.5)* and *intervention task causal probabilities (0.25, 0 .5)*

Table 11

Results of paired comparison *t*-tests for mean final justification responses by *observation task* ΔP *value (0.0 or 0.5)* and *intervention task causal probabilities (0.75 or 0.5)*

Paired Comparison (ΔP value, p(C|E) value)	*N*	*t* value	*p* level
(0.0, 0.5) – (0.0, 0.25)	20	2.47	= .023*
(0.0, 0.5) – (-0.5, 0.5)	20	1.71	= .104
(0.0, 0.5) – (-0.5, 0.25)	20	2.03	= .056
(-0.5, 0.5) – (0.0, 0.25)	20	-1.26	= .225
(-0.5, 0.5) – (-0.5, 0.25)	20	-0.44	= .666
(0.0, 0.25) – (-0.5, 0.25)	20	0.85	= .406

* significant

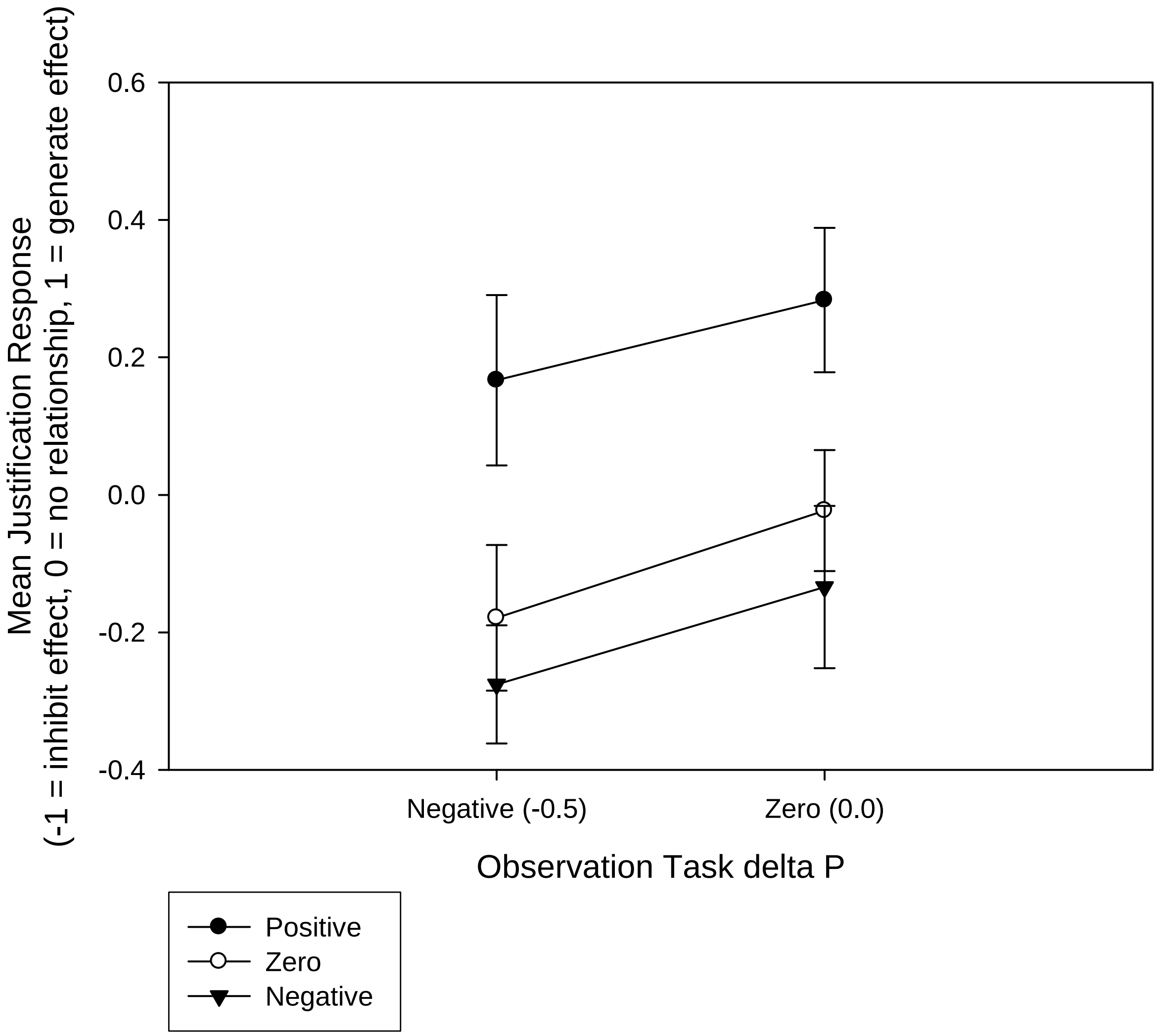


Figure 11. Final justification results: Mean justification responses by *observation task* Δ*P value (0.0, - 0.5)* and *generated sample relationship type (positive, negative, zero)*

Table 12

Results of paired comparison *t*-tests for mean justification responses by *observation task* Δ*P value (0.0 or -0.5)* and *generated sample relationship type (positive, negative, zero)*

Paired Comparison (ΔP value, relationship type)	*N*	*t* value	*p* level
(0.0, negative) – (0.0, zero)	18	-1.11	= .283
(0.0, negative) – (0.0, positive)	20	-2.78	= .012
(0.0, negative) – (-0.5, negative)	20	1.27	= .219
(0.0, negative) – (-0.5, zero)	19	0.52	= .611
(0.0, negative) – (-0.5, positive)	18	-1.64	= .119
(0.0, zero) – (0.0, positive)	18	-2.58	= .019
(0.0, zero) – (-0.5, negative)	18	2.25	= .038
(0.0, zero) – (-0.5, zero)	17	1.21	= .245
(0.0, zero) – (-0.5, positive)	16	-1.31	= .210
(0.0, positive) – (-0.5, negative)	20	3.50	= .002*
(0.0, positive) – (-0.5, zero)	19	2.96	= .008
(0.0, positive) – (-0.5, positive)	18	0.64	= .534
(-0.5, negative) – (-0.5, zero)	19	-0.52	= .612
(-0.5, negative) – (-0.5, positive)	18	-2.92	= .010
(-0.5, zero) – (-0.5, positive)	17	-1.72	= .105

* significant

Effect of scenario/cover story

The potential effect of scenario/cover story was assessed using a repeated measures ANOVA with scenario (levels of which were fertilizer, food allergy, experimental drug, and diet plan) set as the independent variable and observation task rating set as the dependent variable. The analysis revealed a significant effect, $F(3, 57) = 5.801$, $p = .002$. Subsequent paired samples *t*-tests showed that participants significantly rated relationships higher in the chemical scenario than the food ($t(19) = 2.33$, $p = .031$) and drug ($t(19) = 4.26$, $p < .001$) scenarios. Also, participants rated relationships higher in the diet scenario than in the drug scenario ($t(19) = -2.715$, $p = .014$). It is suspected that this effect is a result of the response scale not being optimal for judging negative/preventative relationships.

Experiment 3: Indeterminate and uncorrelated relationships

Observation task

Judgment accuracy in the observation task was assessed using a repeated measures ANOVA with *observation task* Δ*P value* set as the independent variable (ΔP = 0.0 or indeterminate) and relationship rating as the dependent measure. The analysis revealed a main effect of ΔP such that positively correlated samples were rated as higher than uncorrelated samples, $F(1, 19) = 117.23$, $p < .001$ (figure 12). The repeated measures ANOVA with confidence rating set as the dependent measure was not significant, thus indicating that the difference is ratings is not a reflection of confidence levels.

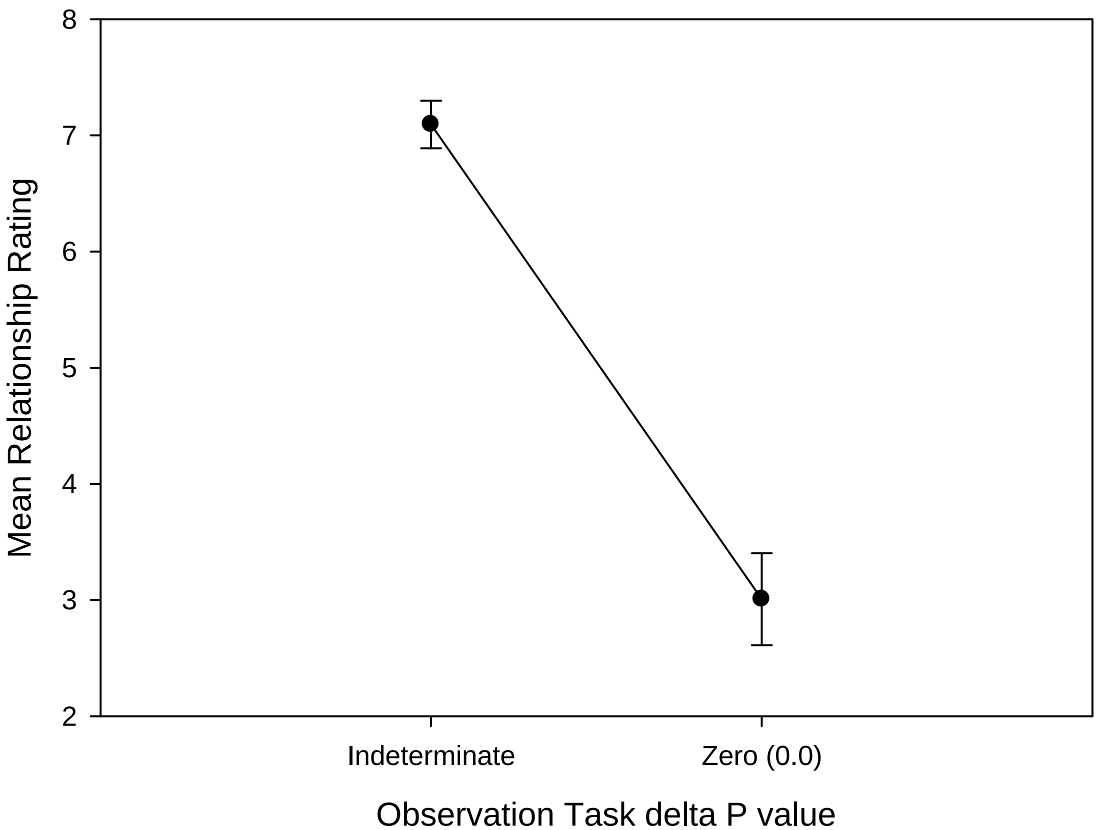


Figure 12. Observation task results: Mean relationship ratings by *observation task ΔP value* (indeterminate and 0.0)

Intervention task

Judgment accuracy in the intervention task was assessed using a repeated measures ANOVA with *causal probability* set as the independent variable ((P(E|C) = .75 and P(E|~C) = .25, P(E|C) = .5 and P(E|~C) = .5) which was non-significant. Given this result, the samples generated by the participants in the intervention task were assessed. Across all participants, 320 samples were generated of which 200 were correlationally indeterminate. A repeated measures ANOVA was run setting the *generated sample relationship type* (categorization method described above) as the independent variable. This analysis showed a significant main effect such that when the generated sample indicated a positive relationship, the relationship was judged to be stronger than when it was zero or negative, $F(2, 38) = 35.20$, $p < .001$ (figure 13). Results of the paired comparison *t*-tests are presented in table 13. Confidence ratings were also analyzed in the same manner and revealed non-significant effects.

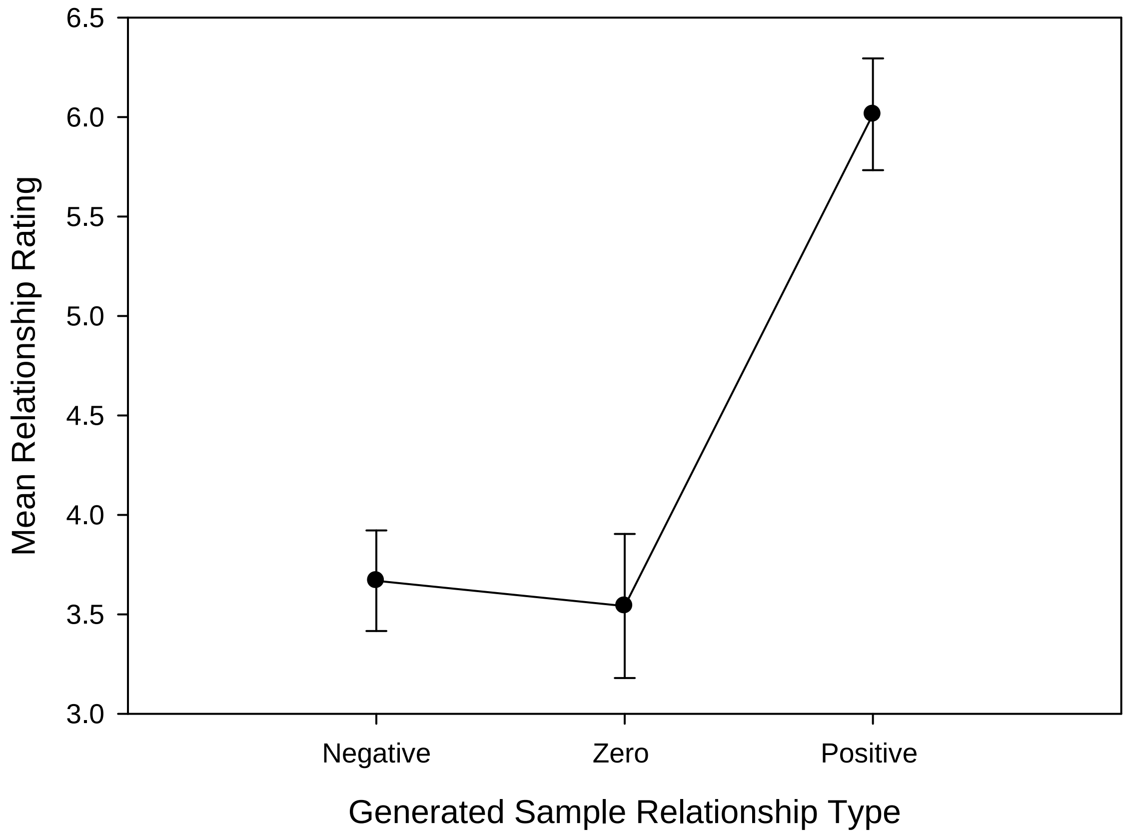


Figure 13. Intervention task results: Mean relationship ratings by *generated sample relationship types (positive, negative, zero)*

Table 13

Results of paired comparison *t*-tests for intervention task mean relationship ratings by *generated sample relationship types (positive, negative, zero)*

Paired Comparison	*N*	*t* value	*p* level
Positive Relationship Type – Negative Relationship Type	20	7.644	< .001*
Positive Relationship Type – No Relationship (Zero) Type	20	7.718	< .001*
Negative Relationship Type – No Relationship (Zero) Type	20	-0.302	= .732

* significant

Final recommendation choice and justification

The final recommendation choices were analyzed using a 2 (*observation task* Δ*P value*) X 2 (*intervention task causal probabilities*) repeated measures ANOVA which revealed a significant main effect of *observation task* Δ*P value*, $F(1, 19) = 100.22$, $p < .001$ (figure 14). Paired comparison *t*-tests are presented in table 14. Given that no significant effect of *causal probabilities* was found in the analysis of the intervention task ratings, final recommendation choices were also analyzed with respect to *generated sample relationship type* as described above. It should be noted that two participants were excluded from the ANOVA for incomplete data (e.g., participant did not have a mean value for each possible condition). The 2 (*observation task* Δ*P value*) X 3 (*generated sample relationship type*) repeated measure ANOVA revealed a significant main effect of *observation task* Δ*P value*, $F(1, 17) = 92.84$, $p < .001$, and a significant main effect of *generated sample relationship type*, $F(2, 34) = 12.66$, $p < .001$ (figure 15). Given the large number of comparisons, a Bonferroni correction was applied (corrected $\alpha = .05/15 = .003$) see table 15 for results of the paired comparison *t*-tests.

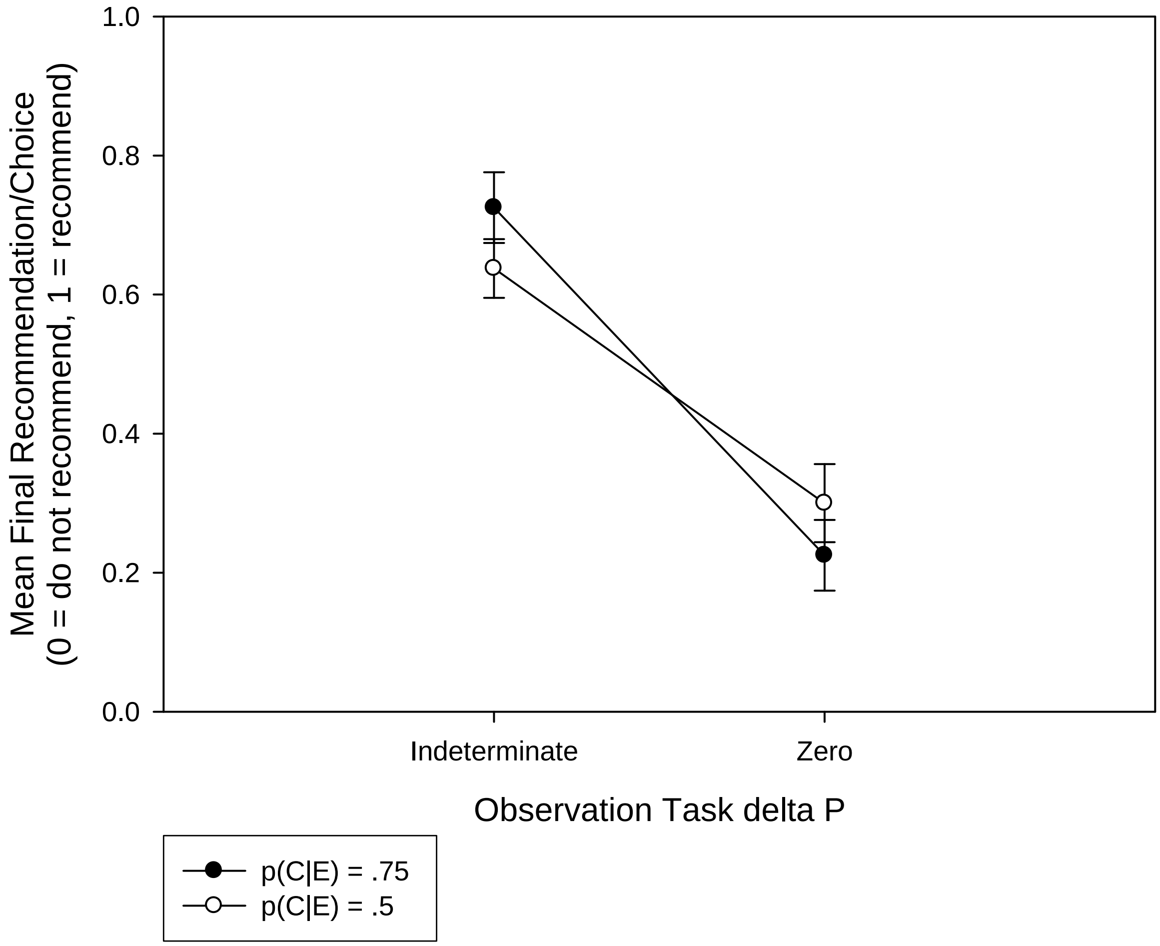


Figure 14. Final recommendation results: Mean recommendation responses by *observation task* Δ*P value (0.0, indeterminate)* and *intervention task causal probabilities (0.75, 0 .5)*

Table 14

Results of paired comparison *t*-tests for mean final recommendation choices by *observation task* Δ*P value (0.0 or indeterminate)* and *intervention task causal probabilities (0.75 or 0.5)*

Paired Comparison (ΔP value, p(C|E) value)	*N*	*t* value	*p* level
(0.0, 0.5) – (0.0, 0.75)	20	0.95	= .356
(indeterminate, 0.5) – (indeterminate, 0.75)	20	1.32	= .201
(0.0, 0.5) – (indeterminate, 0.75)	20	6.03	< .001*
(indeterminate, 0.5) – (0.0, 0.75)	20	6.02	< .001*
(indeterminate, 0.5) – (0.0, 0.5)	20	5.54	< .001*
(0.0, 0.75) – (indeterminate, 0.75)	20	7.65	< .001*

* significant

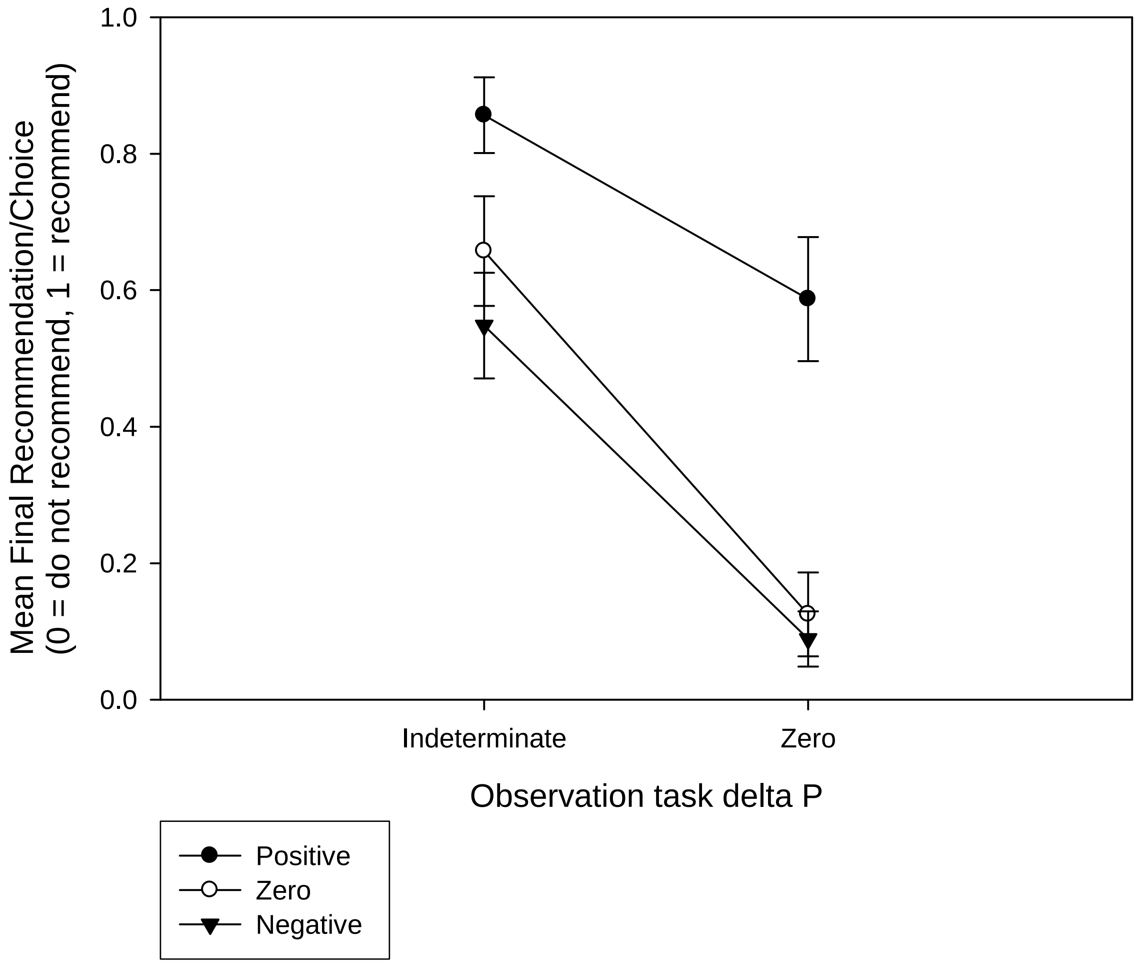


Figure 15. Final recommendation results: mean recommendation responses by *observation task* Δ*P value (0.0, indeterminate)* and *generated sample relationship type (positive, negative, zero)*

Table 15

Results of paired comparison *t*-tests for mean recommendation responses by *observation task* Δ*P value (0.0 or indeterminate)* and *generated sample relationship type (positive, negative, zero)*

Paired Comparison (ΔP value, relationship type)	*N*	*t* value	*p* level
(0.0, negative) – (0.0, zero)	19	-0.42	= .678
(0.0, negative) – (0.0, positive)	20	-4.95	< .001*
(0.0, negative) – (indeterminate, negative)	20	-6.22	< .001*
(0.0, negative) – (indeterminate, zero)	19	-6.83	< .001*
(0.0, negative) – (indeterminate, positive)	20	-11.83	< .001*
(0.0, zero) – (0.0, positive)	19	-4.24	< .001*
(0.0, zero) – (indeterminate, negative)	19	-4.41	< .001*
(0.0, zero) – (indeterminate, zero)	18	-5.78	< .001*
(0.0, zero) – (indeterminate, positive)	19	-9.18	< .001*
(0.0, positive) – (indeterminate, negative)	20	0.08	= .934
(0.0, positive) – (indeterminate, zero)	19	-1.01	= .327
(0.0, positive) – (indeterminate, positive)	20	-3.33	= .004
(indeterminate, negative) – (indeterminate, zero)	19	-0.95	= .354
(indeterminate, negative) – (indeterminate, positive)	20	-3.31	= .004
(indeterminate, zero) – (indeterminate, positive)	19	-1.87	= .078

* significant

Participants' final judgment justifications were analyzed using a 2 (*observation task* Δ*P value*) X 2 (*intervention task causal probabilities*) repeated measures ANOVA which showed a significant main effect of *observation task* Δ*P value*, $F(1, 19) = 40.05$, $p < .001$ (figure 16). Paired comparison *t*-tests are presented in table 16. Final judgment justifications were also analyzed with respect to *generated sample relationship type* as described above. Two participants were excluded from the ANOVA for incomplete data. The 2 (*observation task* Δ*P value*) X 3 (*generated sample relationship type*) repeated measures ANOVA revealed a significant main effect of *observation task* Δ*P value*, $F(1, 17) = 32.10$, $p < .001$; a significant main effect of *generated sample relationship type*, $F(2, 34) = 12.50$, $p < .001$; and a significant interaction, $F(2, 34) = 4.07$, $p = .026$ (figure 17). Given the large number of comparisons, a Bonferroni correction was applied (corrected $\alpha = .05/15 = .003$; see table 17 for results of the paired comparison *t*-tests).

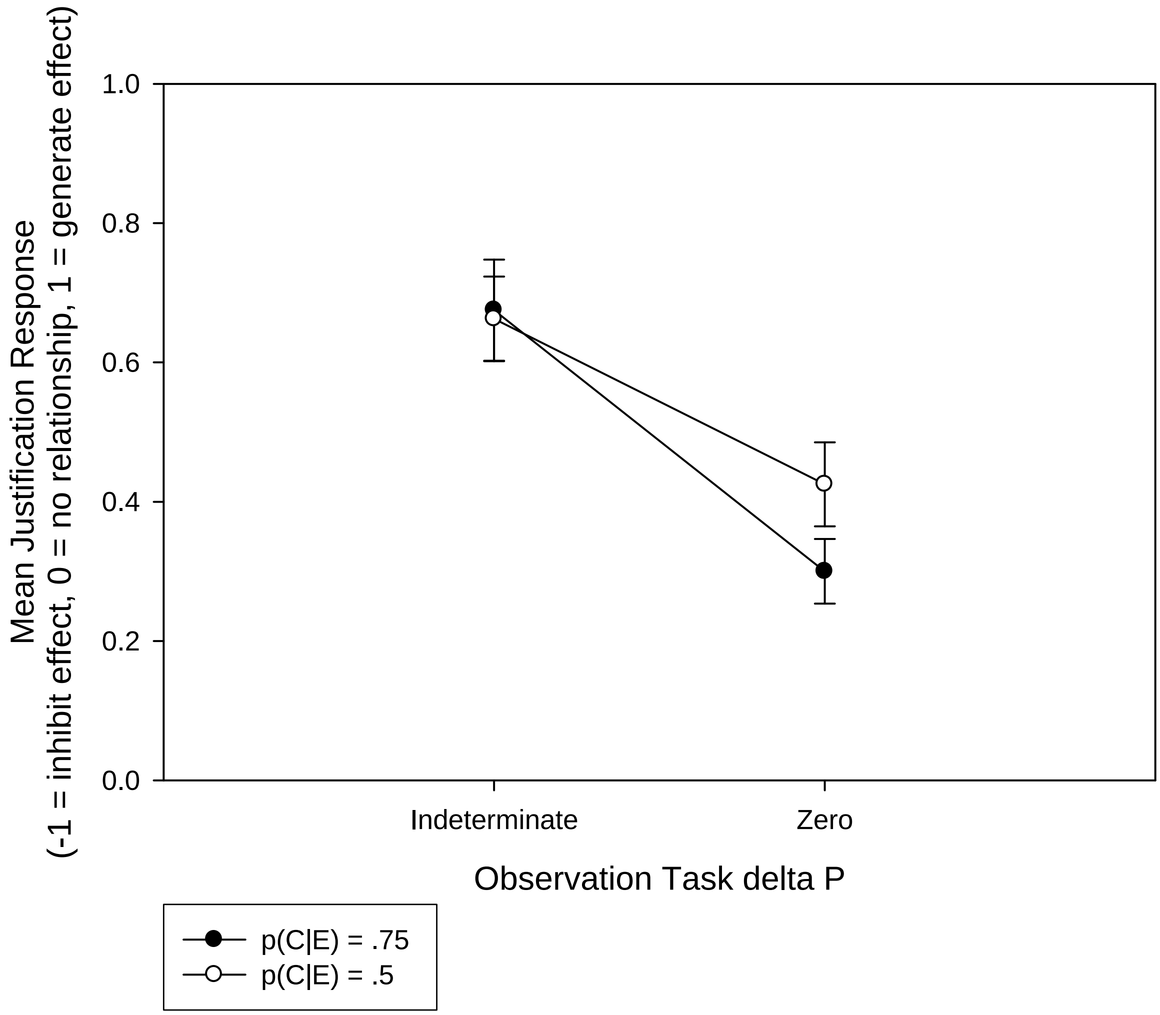


Figure 16. Final justification results: Mean justification responses by *observation task ΔP value (0.0, indeterminate)* and *intervention task causal probabilities (0.75, 0 .5)*

Table 16

Results of paired comparison *t*-tests for mean justification responses by *observation task* ΔP *value (0.0, indeterminate)* and *intervention task causal probabilities (0.75, 0.5)*

Paired Comparison (ΔP value, p(C|E) value)	*N*	*t* value	*p* level
(0.0, 0.5) – (0.0, 0.75)	20	-1.60	= .126
(0.0, 0.5) – (indeterminate, 0.5)	20	-3.23	= .004*
(0.0, 0.5) – (indeterminate, 0.75)	20	-2.65	= .016*
(indeterminate, 0.5) – (0.0, 0.75)	20	-4.31	< .001*
(indeterminate, 0.5) – (indeterminate, 0.75)	20	-0.13	= .900
(0.0, 0.75) – (indeterminate, 0.75)	20	-6.10	< .001*

* significant

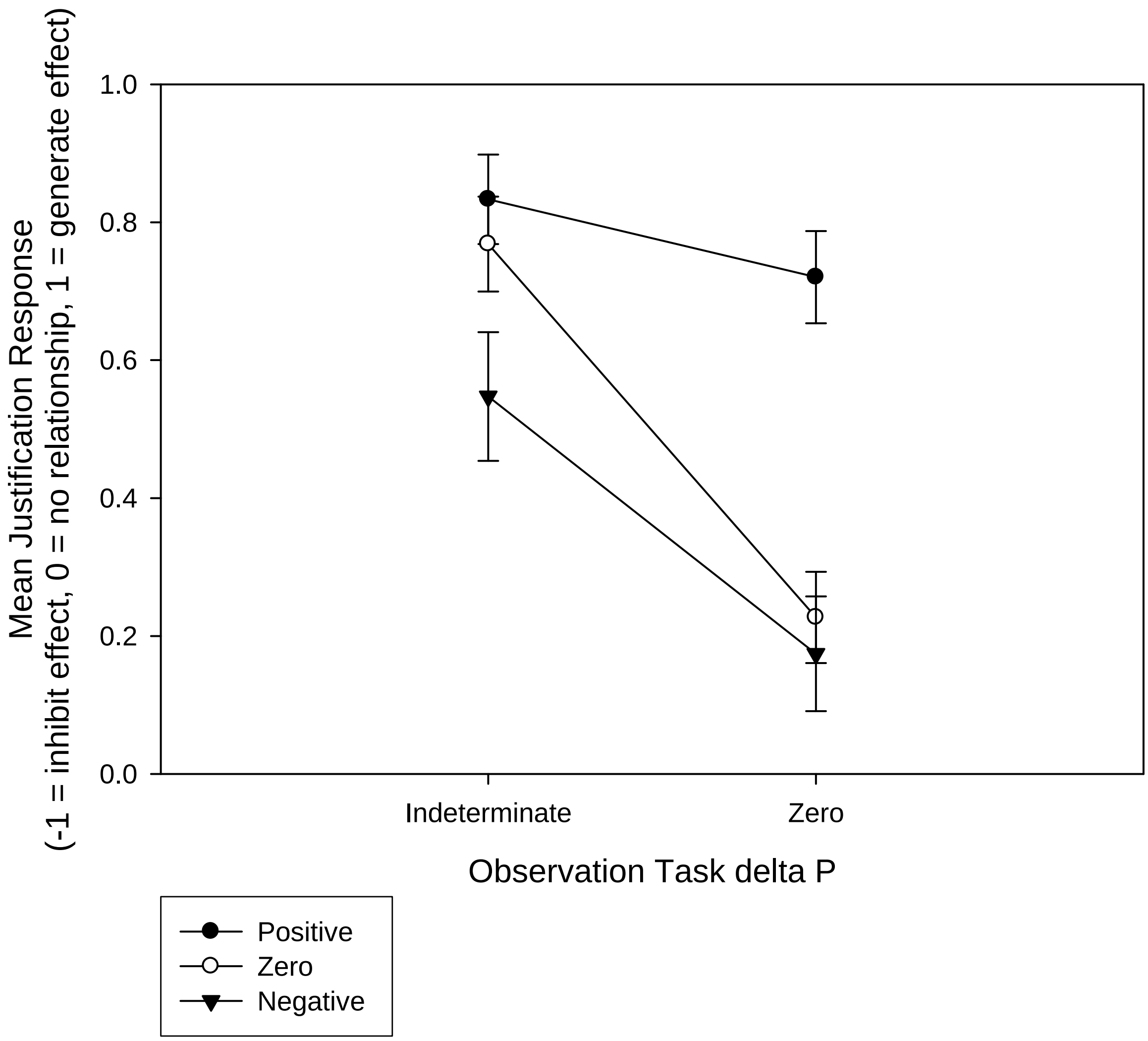


Figure 17. Final justification results: mean justification responses by *observation task ΔP value (0.0, indeterminate)* and *generated sample relationship type (positive, negative, zero)*

Table 17

Results of paired comparison *t*-tests for mean justification responses by *observation task ΔP value (0.0 or indeterminate)* and *generated sample relationship type (positive, negative, zero)*

Paired Comparison (ΔP value, relationship type)	*N*	*t* value	*p* level
(0.0, negative) – (0.0, zero)	19	-0.56	= .580
(0.0, negative) – (0.0, positive)	20	-5.73	< .001*
(0.0, negative) – (indeterminate, negative)	20	-3.54	= .002*
(0.0, negative) – (indeterminate, zero)	19	-3.69	= .002*
(0.0, negative) – (indeterminate, positive)	20	-7.17	< .001*
(0.0, zero) – (0.0, positive)	19	-5.09	< .001*
(0.0, zero) – (indeterminate, negative)	19	-2.54	= .021
(0.0, zero) – (indeterminate, zero)	18	-5.74	< .001*
(0.0, zero) – (indeterminate, positive)	19	-5.90	< .001*
(0.0, positive) – (indeterminate, negative)	20	1.42	= .172
(0.0, positive) – (indeterminate, zero)	19	0.26	= .797
(0.0, positive) – (indeterminate, positive)	20	-0.94	= .361
(indeterminate, negative) – (indeterminate, zero)	19	-0.89	= .385
(indeterminate, negative) – (indeterminate, positive)	20	-2.73	= .013
(indeterminate, zero) – (indeterminate, positive)	19	-1.27	= .222

* significant

Effect of scenario/cover story

The potential effect of scenario/cover story was assessed using a repeated measures ANOVA with scenario (levels of which were fertilizer, food allergy, experimental drug, and diet plan) set as the independent variable and observation task rating set as the dependent variable. The analysis revealed non-significant effects, $F(3, 57) = .619$, $p = .619$.

Discussion

The results of this set of experiments indicate the respective and joint influence of two types of information (observation and intervention) on judgments of causality. The active or inactive role taken by the perceiver in acquiring the information is the primary discrepancy between the two types of information. It was predicted that information acquired by means of an intervention task (or active role) would be more influential on final judgments than observed data (inactive role). The results of these experiments suggest the latter, however. Specifically, final judgments about the effectiveness of a causal candidate in generating the desired outcome were largely reflective of the objective correlation value in the observed data set (observation task). With regard to future military studies, the results of this set of experiments suggest the heavily influential role of observed correlational information in causal judgment which may be applicable to associative learning techniques in a training environment.

Positively correlated and uncorrelated relationships

In the first experiment, participants were presented with positively correlated and uncorrelated samples in the observation task and were accurate in the judgment of these samples. This finding is consistent with the literature. In the intervention task, participants judged generated positive samples to have a moderate relationship and generated negative samples and generated uncorrelated samples to have a weak relationship. The lack of distinction between negative and uncorrelated samples was seen across experiments. Possible explanations are offered below.

At the end of a trial, participants were asked to give a final recommendation regarding the causal candidate's effectiveness in generating the desired outcome. When the observation task sample and the intervention task generated sample were positive, the mean final recommendations were greatest and close to one indicating that participants most frequently judged these causal candidates to be effective. Alternatively, when the observation task sample and the intervention task generated sample were zero or uncorrelated, then the mean final recommendations were lowest and close to zero indicating that participants most frequently judged these candidates to be ineffective. In sum, when the observation and intervention task samples were congruent, or in agreement, participants' final recommendations were very accurate. However, when the samples were incongruent, participants' final recommendations were less accurate. Specifically, when the observation task sample was positive and the intervention task generated sample was negative or zero, participants' mean final recommendations were lower than when congruent. When the observation task sample was zero and the intervention task generated sample was negative or positive, participants' mean final recommendations were greater than when congruent. To some extent, this finding was expected such that incongruent sample types would muddy the waters with respect to final judgments. However, there was a significant main effect of observation task sample and a significant interaction thus suggesting that the observation task sample was more influential on final judgments than intervention task samples. This finding was not expected. This pattern was also seen in the mean final justifications. One potential explanation for this result may be a primacy effect. Participants always saw the observed sample prior to generating a sample. It is possible that participants held steadfast to their initial judgment of the causal candidate's effectiveness.

Negatively correlated and uncorrelated relationships

In the second experiment, participants were presented with negatively correlated and uncorrelated samples in the observation task. Participants' judgments were inaccurate in this task such that participants did not judge negative samples differently than zero correlation samples. It is possible that this is an indication that the judgment scale was inappropriate for use with negatively correlated samples. The scale ranged from 0 to 10 asking that the participant rate the strength of the relationship, but not the direction. Possibly, participants were confused by the direction of the relationship and inability to rate the relationship negatively. Likewise, participants rated negative and zero correlation generated samples in the intervention task similarly and lower than generated positive samples.

The analysis of the mean final recommendations revealed a significant main effect of intervention task generated sample relationship type which indicates that generated negative samples were more frequently not recommended to be effective than generated uncorrelated samples and generated positive samples (e.g., negative > uncorrelated > positive). This suggests that participants were sensitive to the negative relationships but unable to indicate it as such using the observation and intervention task rating scale. This is further supported by the pattern of results shown by the mean justifications. Specifically, participants indicated that causal candidates for which they observed negatively correlated samples (both from the observation and intervention tasks) inhibited the desired outcome. This effect is weaker than for the first experiment and positively correlated samples which is consistent with the existing literature regarding preventative relationships and the asymmetry between judgments of generative relationship and preventative relationships. Logically, it would seem that preventative relationships would simply be the “flip” of generative relationships but responses and inferences to such would suggest otherwise. Perhaps, people struggled with reasoning about negation (Wason & Johnson-Laird, 1972) such that the absence of an effect is more complex to attribute to the presence of a causal candidate rather than the absence of an alternate causal candidate. The present study was limited to single causal candidates and effects for purposes of simplicity but the complex real world environment is not structured as such. To fully understand the generative and preventative relationship asymmetry, future experiments need to incorporate competing causal candidates to test the above stated explanation.

Indeterminate and uncorrelated relationships

In the final experiment, participants were presented with indeterminate and uncorrelated samples. The indeterminate samples were structured such that the probability of the sample being drawn from a positively correlated population was greater than that from a negatively or uncorrelated population (the sample was indicative of a positive relationship). As predicted and consistent with previous work investigating inferences drawn from indeterminate samples, participants rated the indeterminate samples as having a strong relationship and uncorrelated samples has having a very weak or no relationship. Results from the intervention task were similar to the results from the other two experiments such participants rated generated positive samples as having a stronger relationship than generated negative or zero relationship samples. It should be noted that participants in this experiment generated more indeterminate samples in the intervention task than in the other two experiments. Potentially, participants’ intervention actions were influenced by the presentation of indeterminate samples in the observation task.

The results of the mean final recommendations and justifications are similar to that found in the other two experiments. Mean final recommendations are most extreme when the observation and intervention task sample characteristics are congruent. When incongruent, participant more frequently chose to recommend causal candidates for which they had observed an indeterminate sample than an uncorrelated sample. Interestingly and unlike the other incongruent conditions, participants more frequently recommended samples for which they had observed an uncorrelated sample and generated a positive sample. This would suggest that participants were more heavily influenced by the positive generated sample than the observed uncorrelated sample which is inconsistent with the results of the other two experiments. While the pattern (as shown in figure 15) appears similar to that in the other two experiments, the effect of generated sample relationship type is stronger in this particular condition. A similar pattern is seen in the mean justification results.

Limitations and future studies

Given the nature of the intervention task, it is impossible to control the samples generated by participants. While in some regards this is a limitation (incomplete data were excluded from analysis), the data set generated by the participant provides valuable insight into what information the participant deems important and informative and alternatively, that which is deemed unnecessary or uninformative. In this set of studies, participants primarily generated samples that were correlationally indeterminate, such that the causal candidate was applied on each observation (i.e., did not vary), however, a few of the samples generated were indeterminate such that the causal candidate was not applied on any observation (indeterminate-absent). Previous studies of inference from correlationally indeterminate samples that included indeterminate-absent samples found that participants consistently were unsure of what inference could be made from these samples (e.g., Kelley, 2007). Thus, it is interesting that few participants generated these samples in addition to drawing inferences from them. In a future study, it would be interesting to employ a yoking procedure in that participants' derive or generate intervention task samples (as in the current set of experiments) and rate the relationship between the causal candidate and effect variable. These samples would then be presented in an observation task format and another rating of the relationship would be given. The two ratings of the same sample presented in both an observation and intervention format would then be compared to evaluate whether the same conclusions and inferences are drawn from the derived samples in an alternate presentation format.

It is difficult to interpret the results of the second experiment employing negatively correlated samples. The appropriateness of the provided rating scale is questionable given that, despite the request to assess *strength* rather than *direction*, a rating below zero was not allowed. Some, but not all, participants accurately identified negative relationships as evidenced by the justification responses. In fact, an asymmetry between judgments of positive and negative relationships was shown. More research is needed to understand how people interpret negative/preventative relationships (also suggested by Hattori & Oaksford, 2007). People do not, in fact, conceptualize preventative relationships as the reverse of generative relationships, in the domain of correlation detection and causal judgment (Kelley, Anderson, & Doherty, 2007). Future studies employing alternate causal candidates may be beneficial in better understanding how inferences are drawn regarding preventative relationships.

Conclusions

The current set of experiments presented participants with information regarding the presence and absence of a causal candidate and an effect variable through the use of an observation and an intervention task. Participants made judgments after each task and then were asked to incorporate all of the information presented to them to give a final recommendation and justification regarding the effectiveness of the causal candidate in producing the desired outcome. The results of the set of studies showed that participants' final judgments were influenced by both types of information. However, final judgments were largely reflective of the objective sample correlation in the observation task. These findings are inconsistent with previous work which suggests that the active role of the participant in an intervention task is essential to learning whether a variable causes change in another beyond learning the covariation between the variables. The results are more consistent with theories of normative correlation detection than of dual process (heuristic and analytic) causal judgment. The results also have implications for response scale formatting with regard to the type of relationship being presented which will be implemented in future studies of causal judgment and correlation detection conducted by researchers at USAARL.

The results of this study lend to the facilitation of a more effective Soldier in that they build on the understanding of human processing of cause and effect relationships. By understanding how humans form beliefs about the relationships in their environment and which cues are most salient and influential in increasing the likelihood of accurate belief formation, more efficient information displays can be designed to apply brevity and intuitiveness to training and operational environments. Of particular interest for future studies are conditions under which short-cut heuristics are likely to be used by Soldiers (e.g., sleep deprivation) thus increasing the probability of an error in judgment. By understanding cue salience, the presented cues can be manipulated to exploit these short-cuts and ultimately decrease the likelihood of an error.

References

Anderson, J. R., and Sheu, C. F. 1995. Causal inferences as perceptual judgments. Memory & Cognition. 23: 510-524.

Baker, A. G., Murphy, R. A., and Vallee-Tourangeau, F. 1996. Associative and normative accounts of causal induction: Reacting to verses understanding a cause. In D. R. Shanks, K. J. Holyoak, & D. L. Medin (Eds.), The psychology of learning and motivation: Vol. 34. Causal learning: pp.1-46. London: Academic.

Busemeyer, J. R. 1990. Intuitive statistical estimation. In Anderson, N. H. (Ed.), Contributions to Information Integration Theory: pp. 187-215. Hillsdale, N. J.: Lawrence Erlbaum Associates, Inc.

Chapman, L. J. 1967. Illusory correlation in observational report. Journal of Verbal Learning and Verbal Behavior. 6,:151-155.

Cheng, P.W. 1997. From covariation to causation: A causal power theory. Psychological Review. 104: 367-405.

Cheng, P.W., and Novick, L.R. 1992. Covariation in natural causal induction. Psychological Review. 99: 365-382. Reprinted in Goldstein, W.M., & Hogarth, R.M. (Eds, 1997). Research on judgment and decision making: Currents, connections, and controversies. New York, N.Y.: Cambridge University Press.

Crocker, J. 1981. Judgment of covariation by social perceivers. Psychological Bulletin. 90: 272-292.

Doherty, M. E., Anderson, R. B., Angott, A. M., and Klopfer, D. S. 2007. The perception of scatterplots. Perception & Psychophysics. 69: 1261-1272.

Doherty, M. E., Anderson, R. B., Kelley, A. M., and Albert, J. 2006. Probabilistically valid inference of covariation from a single *xy* observation when univariate characteristics are known. Poster presented at the annual meeting of the Society for Judgment and Decision Making, Houston, TX.

Griffiths, T. L., and Tenenbaum, J. B. 2005. Structure and strength in causal induction. Cognitive Psychology. 51: 334-384.

Hattori, M., and Oaksford, M. 2007. Adaptive non-interventional heuristics for covariation detection in causal induction: Model comparison and rational analysis. Cognitive Science. 31: 765-814.

Heuer, Jr., R.J. 1999. Biases in perception of cause and effect. In Psychology of Intelligence Analysis (chap. 11). Retrieved November 27, 2007, from http://www.au.af.mil/au/awc/awcgate/psych-intel/art14.html.

Jenkins, H. M., and Ward, W. C. 1965. Judgment of contingency between responses and outcomes. Psychological Monographs: General and Applied. 79: 1, Whole No. 594.

Kao, S. F., and Wasserman, E. A. 1993. Assessment of an information integration account of contingency judgment with examination of subjective cell importance and method of information processing. Journal of Experimental Psychology: Learning, Memory, and Cognition. 19: 1363-1386.

Kareev, Y. 2005. And Yet the Small-Sample Effect Does Hold: Reply to Juslin and Olsson (2005) and Anderson, Doherty, Berg, and Friedrich (2005). Psychological Review. 112: 280-285.

Kelley, A. M. 2007. Bayesian principles and causal judgment. Dissertation Abstracts International: Section B: The Sciences and Engineering. 68 (5-B): 3418.

Kelley, A. M., Anderson, R. B., and Doherty, M. E. 2007. Perception of correlation in determinate and indeterminate data. Manuscript in preparation.

Lagnado, D. A., and Sloman, S. A. 2004. The advantage of timely intervention. Journal of Experimental Psychology: Learning, Memory, and Cognition. 30: 856-876.

Levin, I. P., Wasserman, E. A., and Kao, S. F. 1993. Multiple methods for examining biased information use in contingency judgments. Organizational Behavior and Human Decision Processes. 55: 228-250.

Mandel, D. R., and Lehman, D. R. 1998. Integration of contingency information in judgments of cause, covariation, and probability. Journal of Experimental Psychology: General. 127: 269-285.

McKenzie, C. R. M., and Mikkelsen, L. A. 2007. A Bayesian view of covariation assessment. Cognitive Psychology. 54: 33-61.

Rescorla, R.A., and Wagner, A.R. 1972. A theory of Pavlovian learning: Variations in the effectiveness of reinforcement and non-reinforcement. In A.H. Black & W.F. Prokasy (Eds.), Classical Conditioning II: Current theory and research: pp. 64-99. New York: Appleton-Century-Crofts.

Shanks, D. R. 2004. Judging covariation and causation. In D. J. Koehler and N. Harvey (Eds.), Blackwell Handbook of Judgment and Decision Making: pp. 220-239. Oxford, England: Blackwell Publishing.

Stanovich, K. E. 1999. Who is rational? Studies of individual differences in reasoning. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Steyvers, M., Tenenbaum, J. B., Wagenmakers, E. J., and Blunmd, B. 2003. Inferring causal networks from observations and interventions. Cognitive Science. 27: 453 – 489.

Ward, W. D., and Jenkins, H. M. 1965. The display of information and the judgment of contingency. Canadian Journal of Psychology. 19: 231-241.

Wason, P. C., and Johnson-Laird, P. N. 1972. Psychology of reasoning: Structure and content. Cambridge, MA: Harvard University Press.

Wasserman, E. A., Dorner, W. W., and Kao, S. F. 1990. Contributions of specific cell information to judgments of interevent contingency. Journal of Experimental Psychology: Learning, Memory, and Cognition. 14: 509-521.

White, P. A. 2000. Causal judgment from contingency information: the interpretation of factors common to all instances. Journal of Experimental Psychology: Learning, Memory, and Cognition. 26: 1083-1102.

White, P. A. 2003. Making causal judgments from the proportion of confirming instances: The *pCI* rule. Journal of Experimental Psychology: Learning, Memory, and Cognition. 29: 710-727.

Appendix

Example of a trial.

Instructions (Screen 1):

Imagine that you are an agricultural scientist. You have identified a chemical compound that may be an effective fertilizer. It is your ultimate goal to understand how the chemical compound is related to plant growth. First, you will look at a small sample of data drawn from test results conducted by another researcher. You will be asked to assess the relationship between the chemical compound and plant growth. Next, you will be able to generate your own data and to test the chemical compound. You will be asked to assess this data. Finally, you will be asked to evaluate the chemical compound (labeled as Chemical X) as a potential fertilizer. Please ask the experimenter any questions you may have now.

Observation Task (Screen 2):

Below is a sample of data from a previous experiment. The Plant # is the arbitrary label given to the plant in the experiment. Chemical Applied indicates whether the plant received the chemical compound. Plant Growth indicates whether the plant grew a significant amount. [In this example, $\Delta P = 0.5$]

Plant #	Chemical Applied?	Plant Growth?
163	Yes	Yes
57	No	No
4	No	Yes
111	No	Yes
582	Yes	Yes
312	Yes	No
84	Yes	No
65	No	No

Please rate the relationship between the chemical compound and plant growth.

0 (No relationship) ______________________ 10 (Very strong relationship)

Relationship Rating: ____

How confident are you in your rating?

0 (Not confident) ____________________ 10 (Very confident)

Confidence Rating: ____

Intervention Task (Screen 3):

Now you will be able to generate some of your own data. For each observation, you can either choose to apply the chemical to the plant or not. Afterwards, you will be told whether the plant grew.

Press X to apply the chemical to your first plant.
Press Y not to apply the chemical to you first plant.

(Screen 4): You chose to (apply/not apply) the chemical. The plant (grew/did not grow).

(Screen 5):
Press X to apply the chemical to your second plant.
Press Y not to apply the chemical to your second plant.

(Screen 6): You chose to (apply/not apply) the chemical. The plant (grew/did not grow).

(Screen 7):
Press X to apply the chemical to your third plant.
Press Y not to apply the chemical to your third plant.

(Screen 8): You chose to (apply/not apply) the chemical. The plant (grew/did not grow).

(Screen 9):
Press X to apply the chemical to your fourth plant.
Press Y not to apply the chemical to your fourth plant.

(Screen 10): You chose to (apply/not apply) the chemical. The plant (grew/did not grow).

(Screen 11):
Press X to apply the chemical to your fifth plant.
Press Y not to apply the chemical to your fifth plant.

(Screen 12): You chose to (apply/not apply) the chemical. The plant (grew/did not grow).

(Screen 13):
Press X to apply the chemical to your sixth plant.
Press Y not to apply the chemical to your sixth plant.

(Screen 14): You chose to (apply/not apply) the chemical. The plant (grew/did not grow).

(Screen 15):
Press X to apply the chemical to your seventh plant.
Press Y not to apply the chemical to your seventh plant.

(Screen 16): You chose to (apply/not apply) the chemical. The plant (grew/did not grow).

(Screen 17):
Press X to apply the chemical to your eighth plant.
Press Y not to apply the chemical to your eighth plant.

(Screen 18): You chose to (apply/not apply) the chemical. The plant (grew/did not grow).

(Screen 19): Given the results of the data you just generated, please rate the relationship between the chemical compound and plant growth:

0 (No relationship) 10 (Very strong relationship)

Relationship Rating: ______________________________________

How confident are you in your rating?

0 (Not confident) 10 (Very confident)

Confidence Rating: ______________________________________

Final Assessment:

You have now seen data about Chemical X and plant growth. You've also generated your own data.
Given all the information you have seen, please answer the following questions:

Would you recommend Chemical X for a fertilizer?
1 yes
2 no

Why did you decide to recommend or not recommend Chemical X?

1 Chemical X stops plants from growing.
2 Chemical X doesn't affect plant growth.
3 Chemical X makes plants grow.

DEPARTMENT OF THE ARMY
U.S. Army Aeromedical Research Laboratory
Fort Rucker, Alabama 36362-0577