March 2019

Reference Dependence in Bayesian Reasoning

Alaina N. Talboy
University of South Florida, alaina.talboy@gmail.com

Reference Dependence in Bayesian Reasoning

by

Alaina N. Talboy

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
with a concentration in Cognition, Neuroscience, and Social Psychology
Department of Psychology
College of Arts and Sciences
University of South Florida

Major Professor: Sandra Schneider, Ph.D.
Michael Brannick, Ph.D.
Chad Dube, Ph.D.
Daniel Lende, Ph.D.
Kristen Salomon, Ph.D.

Date of Approval:
March 1, 2019

Keywords: Reference Dependence, Bayesian Reasoning, Problem Structure, Numeracy

Table of Contents

List of Tables

List of Figures

Abstract

The purpose of this dissertation is to examine aspects of the representational and computational influences on Bayesian reasoning as they relate to reference dependence. Across three studies, I explored how dependence on the initial problem structure influences the ability to solve Bayesian reasoning tasks. Congruence between the problem and question of interest, response errors, and individual differences in numerical abilities was assessed. The most consistent and surprising finding in all three experiments was that people were much more likely to utilize the superordinate value as part of their solution rather than the anticipated reference class values. This resulted in a weakened effect of congruence, with relatively low accuracy even in congruent conditions, as well as a different pattern of response errors than what was anticipated. There was consistent and strong evidence of a value selection bias in that incorrect responses almost always conformed to values that were provided in the problem rather than errors related to computation. The one notable exception occurred when no organizing information was available in the problem, other than the instruction to consider a sample of the same size as that in the problem. In that case, participants were most apt to sum all of the subsets of the sample to yield the size of the original sample (N). In all three experiments, higher numerical skills were generally associated with higher accuracy, whether calculations were required or not.

Reference Dependence in Bayesian Reasoning

Diagnostic tests are used in many domains to help distinguish who has or does not have a condition of interest. However, these tests are not perfect, so a positive test result does not always correspond to the presence of a condition. From an individual's standpoint, knowing the likelihood that a positive test result indicates the presence of a condition is an important piece of information that would be helpful to know. This is the *positive predictive value* (PPV) of the test, which compares the subset of those who have the condition and test positive (C+T+) to all of those who test positive (T+).

These types of tests are commonly presented as Bayesian reasoning problems, which are used to evaluate the ability to update prior beliefs based on additional evidence in order to determine a posterior probability. Research over the last 40 years demonstrates that uninitiated reasoners, as well as novice reasoners, tend to have difficulty working through these types of problems to correctly determine the PPV (e.g., Gigerenzer, Gaissmaier, Kurz-Milcke, Schwartz, & Woloshin, 2007; Gigerenzer & Hoffrage, 1995; Hoffrage, Krauss, Martignon, & Gigerenzer, 2015; Johnson & Tubau, 2015; Reyna & Brainerd, 2008; Sirota, Kostovičová, & Vallée-Tourangeau, 2015).

Factors that have been associated with the low accuracy rates observed in Bayesian reasoning tasks are broadly categorized as representational or computational difficulties (Johnson & Tubau, 2015; Talboy & Schneider, 2018a, 2018b). Difficulties with how the problem is cognitively represented by reasoners are typically attributed to how the components of the problem relate to each other, which is not readily apparent in some formulations of the problem.

Although many researchers have attempted various manipulations to encourage reasoners to be aware of and understand the nested structure of the problem, accuracy still falls short and is not consistent across manipulations (Brase, 2014; Garcia-Retamero, Cokely, & Hoffrage, 2015; Sirota, Kostovičová, & Juanchich, 2014; Sirota et al., 2015). This representational issue is compounded by the computational difficulties of extracting and computing the value needed to determine the PPV from the information provided in the problem (Cosmides & Tooby, 1996; Gigerenzer & Hoffrage, 1995; Macchi, 2000). The computational difficulties are especially apparent for those who struggle with numerical concepts compared to those who have stronger numerical skills (Chapman & Liu, 2009; Reyna & Brainerd, 2008; Schwartz, Woloshin, Black, & Welch, 1997; Talboy & Schneider, 2018b).

I propose that many of the representational and computational difficulties associated with Bayesian reasoning tasks are due in part to reference dependence, or the tendency to adopt a given or implied reference point at the start of cognitive deliberations. The contextual structuring provided by the problem description gives uninitiated reasoners a starting point from which to evaluate the values in order to work toward a solution. Across a variety of different types of problem solving, research suggests that inexperienced reasoners will often rely on the problem structure and organization to guide their approach to solution (Chi, Glaser, & Rees, 1981; Talboy & Schneider, 2018a). Relying on that structure can be problematic when the structure presents information in a way that is not consistent with the question being asked. I propose that this reference dependence is a major factor in the solutions that reasoners generate when trying to solve Bayesian reasoning problems.

The purpose of this dissertation is to examine aspects of the representational and computational influences on Bayesian reasoning as they relate to reference dependence. Across three studies, I explored how dependence on the initial problem structure influences the ability to solve Bayesian reasoning tasks. I also evaluated the interplay of adopting a specified reference point when other representational and computational factors are manipulated. Finally, I looked at how individual differences in numerical skill connect to these larger issues.

Bayesian Reasoning Problems

Bayesian reasoning problems are used to assess one's ability to determine the likelihood of having a condition given a positive test result. Reasoners are often asked to determine this *positive predictive value* (PPV) based on three pieces of information: the base rate of the condition within a given population or sample, the true positive rate (indicating those who have the condition and test positive), and the false positive rate (indicating those who do not have the condition but test positive).

One of the first formulations of Bayesian reasoning problems was presented using single-event probabilities for each of the numeric values, which indicates the chance of some event happening without specifying a particular reference class. In this numeric format, the PPV is calculated as a conditional probability using Bayes Theorem. Accuracy was generally very poor (around 15%), with the highest reported accuracy around 30% in the absence of training or other aids (e.g., Casscells, Schoenberger, & Graboys, 1978; Cosmides & Tooby, 1996; Eddy, 1982; Galesic et al., 2009; Gigerenzer & Hoffrage, 1995; Sirota, Juanchich, & Hagmayer, 2014; Sloman, Over, Slovak, & Stibel, 2003).

One of the largest breakthroughs in improving Bayesian reasoning has come through representing the numeric information as natural frequencies instead of single-event probabilities (e.g., Garcia-Retamero & Hoffrage, 2013; Gigerenzer, Gaissmaier, Kurz-Milcke, Schwartz, & Woloshin, 2007; Gigerenzer & Hoffrage, 1995). Natural frequencies provide counts of occurrences, indicating the total number of events in relation to a reference class of a specified size, as demonstrated using the classic mammography problem, shown below:

> 10 out of every 1,000 women at age forty who participate in routine screening have breast cancer. 8 of every 10 women with breast cancer will get a positive mammography. 95 out of every 990 women without breast cancer will also get a positive mammography.
>
> *Q*: Here is a new representative sample of women at age forty who got a positive mammography in routine screening. How many of these women do you expect to actually have breast cancer? ___ out of ___

In each statement, the target subgroup is defined in relation to a specified reference class using integer values. In this format, the PPV is calculated as a joint probability using a simplified form of Bayes Theorem rather than the more difficult conditional probability algorithm (Gigerenzer & Hoffrage, 1995).

When Bayesian reasoning problems are presented in this manner, around 40% of participants typically determine the correct solution in the absence of training (e.g., Brase, 2014; Gigerenzer & Hoffrage, 1995; Micallef, Dragicevic, Fekete, & Assessing, 2012; Sirota et al., 2014). Despite this dramatic improvement with natural frequencies, though, well over half of participants across studies still struggle to determine the correct solution to Bayesian reasoning problems. Many of the remaining difficulties associated with this task can be categorized as representational issues and computational issues (Johnson & Tubau, 2015).

Representation

One of the biggest difficulties associated with correctly determining the solution for Bayesian reasoning problems is identifying the correct reference class for PPV. To do this requires understanding the nested structure of the problem as well as the subset-set relationships between the components of the problem (e.g., Barbey & Sloman, 2007; Brase & Hill, 2017; Girotto & Pighin, 2015; Sirota, Juanchich, et al., 2014). To demonstrate, Table 1 shows how the nested values of a traditional Bayesian reasoning problem are organized using a contingency table. The sections presented with no shading indicate which values are typically included in traditional presentations, such as the subset of those who test positive and have the condition (C+T+) and the subset of those who do not have the condition and test positive (C-T+). The grayed boxes indicate values that could be elucidated but are often not, such as the complementary subsets of those who test negative (C+T- and C-T-). The most critical of these for PPV is the total of those who test positive (T+). This value is needed to determine the reference class or denominator for the PPV.

Table 1. Bayesian Reasoning Task Values Organized in a 2x2 Contingency Table

	TEST POSITIVE	**TEST NEGATIVE**	*Marginal Totals*
CONDITION POSITIVE	Condition Positive and Test Positive (C+T+)	Condition Positive and Test Negative (C+T-)	Total Condition Positive (C+)
CONDITION NEGATIVE	Condition Negative and Test Positive (C-T+)	Condition Negative and Test Negative (C-T+)	Total Condition Negative (C-)
Marginal Totals	Total Test Positive (T+)	Total Test Negative (T-)	Superordinate Set (N)

Some researchers have created external aids, such as icon arrays or mosaic plots that visually demonstrate the nested relationship of all four subsets (e.g., Brase, 2014; Garcia-Retamero & Hoffrage, 2013; Sirota, Kostovičová, & Juanchich, 2014). Others have designed elaborate training programs that teach reasoners how to represent the nested structure and the set-subset relationships (Kurzenhäuser & Hoffrage, 2002; Navarrete & Mandel, 2016; Sedlmeier &

Gigerenzer, 2001; Talboy & Schneider, 2017). However, these types of interventions have had limited success (see Brust-Renck, Royer, & Reyna, 2013; Gaissmaier et al., 2012) and are often not available in the applied situations where Bayesian reasoning problems are typically encountered.

In these applied situations, it may be more practical to manipulate the information that is presented within the problem itself instead of relying on an external aid to elucidate the nested components of the problem. As shown above, the verbal presentations do not include all of the entries relevant to the diagnostic test. It has been widely demonstrated that untrained reasoners tend to create mental representations of a problem based, often exclusively, on the information that is provided within the problem (Johnson-Laird, 1994; Kintsch & Greeno, 1985; Sirota, Juanchich, et al., 2014).

In the case of Bayesian reasoning problems presented in the standard format, this suggests that reasoners are likely to encode only the subset information that is provided in the problem text rather than all possible subsets. Because of this, they may not even recognize the nested set structure inherent in the problem. By providing a verbal presentation that includes all of the component parts of the contingency matrix, reasoners may be more likely to develop a more complete mental model of the problem that helps them understand and manipulate values within the nested structure to determine the correct solution.

Experiment 1 was designed to test whether providing complete subset information will facilitate the appropriate representation of the problems. This manipulation also provides a baseline for comparison for manipulations introduced in Experiments 2 and 3. In order to get uninitiated reasoners to appreciate the structure of the problem, I created a problem formulation that includes all four possible (conjoint) subsets to verbally fill all four cells of the nested structure to move closer to what is accomplished by visual aids or training. By expanding the amount of information available in the problem text, reasoners may be in a better position to recognize the structure of the problem and build a more complete mental model, which may in turn increase ability to correctly identify the PPV.

Computation

Although the inclusion of full subset information may improve the mental model reasoners create based on the problem presentation, this does little to address the computational difficulties associated with combining values to determine the correct marginal reference class needed for the PPV question. Responses to the PPV question may be elicited as a pair of integers (i.e., ___ out of ___ people) wherein the numerator is the subset of interest (C+T+) and the denominator is the total reference class (T+) to which the subset is compared. This is referred to as a frequency response format, which requires a single calculation (addition) to determine the denominator. Alternatively, responses may be elicited as a percentage value (i.e., ___%) wherein the two values needed for the frequency response format are combined through division. This is referred to as a percentage response format.

Although some argue that these computations are simple because they involve basic arithmetic operations like adding or dividing (Johnson-Laird, Legrenzi, Girotto, Legrenzi, & Caverni, 1999; Sloman et al., 2003), there is substantial evidence that many reasoners are unable to complete the computations required to correctly solve Bayesian reasoning problems (Johnson & Tubau, 2015; Mayer, 2003; Reyna & Brainerd, 2008; cf. Talboy & Schneider, in progress). Often, reasoners tend to select values directly from the problem as their denominator for the frequency response format instead of calculating the value needed (Cosmides & Tooby, 1996;

Galesic et al., 2009; Gigerenzer & Hoffrage, 1995; Gigerenzer, Hoffrage, & Kleinbölting, 1991; Talboy & Schneider, 2017, 2018a; Wolfe, Fisher, & Reyna, 2013).

There are two potential explanations for this particular component response error. First, reasoners may be dependent on how the problem information is structured, which leads them to utilize the reference class totals that are focused on in the problem, even when they are not appropriate for reaching the solution. Alternatively, reasoners may have a general bias toward selecting values from the problem for their solution or for assuming that their task is to find the needed value from within the problem. Both of these possibilities are also consistent with a bias toward cognitive ease (Kahneman, 2011), which favors a readily available answer over even seemingly innocuous arithmetic steps such as adding two values together. As Ayal and Beyth-Marom (2014) have shown, accurate performance in solving probability-based reasoning problems drops off dramatically as the need for computations increases.

There is more to computation than just the arithmetic required to add or divide two integer values. In problem solving, computation involves the cognitive processes of conceptualizing what is needed for solution, selecting relevant values from the problem, and applying the appropriate values toward solution. Only then does the ability to perform the correct arithmetic step enter into consideration. Computation in this sense involves a preexisting body of knowledge to enable recognition of which mathematical operation is appropriate for the question being asked. It also involves analytic abilities to correctly interpret the meaning of the provided numbers within the context of the problem in addition to figuring out where to include the values in the operation itself before actually completing the arithmetic step. Errors at any point during this process can contribute to incorrect computations, and as a result, incorrect solutions (Talboy & Schneider, 2018a).

Difficulties with computations are especially apparent for those with low numeracy compared to those with higher numeracy (Chapman & Liu, 2009; Reyna & Brainerd, 2008; Schwartz et al., 1997; Talboy & Schneider, 2017, 2018b). Numeracy is the ability to work with and understand numbers in various numeric formats (Peters et al., 2006; Peters, Hibbard, Slovic, & Dieckmann, 2007). Those who experience more difficulty working with numeric information tend to perform worse on reasoning tasks than those who have higher numeracy (Garcia-Retamero & Galesic, 2010; Hill & Brase, 2012; Johnson & Tubau, 2013; Lipkus, Samsa, & Rimer, 2001; Peters et al., 2006; Schwartz, Woloshin, & Welch, 2005).

Experiment 2 was designed to separate out the extent to which reasoners utilize numeric values as they are presented in the problem text versus completing necessary computations to determine the correct solution. I created problem formulations that do not require computation of the denominator to determine the correct response to compare to problems that do require computation. Comparing accuracy on these two problem formulations will help determine whether reasoners are more likely to select an inappropriate value from the problem to apply as the denominator in the solution, or whether they can be encouraged to complete computations to determine the correct denominator. For each problem formulation, I will also look at the role of numeracy in relation to accuracy.

Reference Dependence

A major proposition of this work is that many of the problems associated with the representation and computational components of Bayesian problem solving are tied to the cognitive process of reference dependence. Reference dependence is the tendency to start cognitive deliberations from a given or indicated reference point. Representational and computational difficulties are compounded when the problem starts from one reference point and the question asks reasoners to assess the information from another reference point (Johnson & Tubau, 2013; Pighin, Tentori, Savadori, & Girotto, 2018; Talboy & Schneider, 2018a, 2018b).

In particular, the traditional Bayesian reasoning problem presents information within a condition nesting (C+ and C-), which provides a framework in which reasoners can organize the subset information. However, they are standardly asked to evaluate the PPV of the problem, which asks reasoners to evaluate the subset of interest within a test nesting (T+). Therefore, the most common presentation of Bayesian reasoning problems partitions the nested sets in a way that is not congruent with the posterior likelihood of interest (Girotto & Gonzalez, 2001; Pighin et al., 2018; Talboy & Schneider, 2017, 2018a, 2018b).

To determine the correct solution, reasoners must ignore the reference class totals that are the focus in the problem in order to calculate the marginal reference class total of all those who test positive (T+). The presence of a competing reference class total in the problem may cause a type of processing interference (e.g., Reyna, 2004; Reyna & Brainerd, 2008), which inhibits the reasoner's ability to evaluate the problem from the alternate reference point.

Reference dependence is one of the most ubiquitous findings throughout the judgment and decision making literature. A wealth of research indicates that decisions are highly dependent on the reference frame used to present choices (e.g., Dinner, Johnson, Goldstein, & Liu, 2011; Hájek, 2007; Lopes & Oden, 1999; Tversky & Kahneman, 1991), and that many decision heuristics, such as default and anchoring effects, as well as framing effects can be explained by reference dependence. Although the majority of research documenting reference dependence comes from the choice literature, the importance of context in shaping behavior has also been noted in several other domains, including logical reasoning (Johnson-Laird, 2010), problem solving (Kotovsky & Simon, 1990), extensional reasoning (Fox & Levav, 2004)—and now in Bayesian reasoning as well (Talboy & Schneider, 2018a, 2018b).

In particular, reasoners will often use the reference class that is the focus of the problem description as a part of their solution instead of calculating the correct reference class needed for solution. As a result, responses often correspond to the sensitivity of the test (C+T+/C+) instead of the PPV (C+T+/T+; Cosmides & Tooby, 1996; Galesic et al., 2009; Gigerenzer & Hoffrage, 1995; Gigerenzer, Hoffrage, & Kleinbölting, 1991; Talboy & Schneider, 2017, 2018a; Wolfe, Fisher, & Reyna, 2013). In Talboy and Schneider (2018a), I found that about one-third of participants consistently indicated the sensitivity as their response, but also that another one-third of participants used the condition reference class as the subset of interest in relation to the superordinate set—another alternative reference class—as their response for the denominator on the frequency format question.

This critical issue of using the incorrect reference class as part of the solution is often attributed to representational difficulties, and it may also reflect a bias to avoid computations. However, I further argue that both the representational and computational difficulty of identifying the correct reference class is compounded by the fact that reasoners must start evaluating the problem matrix from a different reference point than the one needed for solution.

I refer to this traditional presentation of Bayesian reasoning problems as the "condition-focus" presentation because the problem text explicitly uses the presence or absence of the condition as the reference class in which the subsets are organized and evaluated (Talboy & Schneider, 2018b, 2018a). These problems can be reorganized to focus on the alternate set of reference classes, which indicate the total number of people who test positive (T+) or negative (T-), which includes the reference class that is needed to arrive at the PPV value. I refer to this form as a “test-focus” presentation (Talboy & Schneider, 2018b, 2018a). By changing the organization and focus of the problem, I am manipulating the reference structure that reasoners are likely to rely on to answer the PPV question. These two problem forms are shown in Table 2 using the mammography example.

Table 2. Example Presentations with Partial Subset Information for the Mammography Problem

Condition-Focus (CF)	**Test-Focus (TF)**
In this sample of 10,000 women, 100 have breast cancer.	In this sample of 10,000 women, 1,070 received a positive result on their mammogram.
Of the 100 women who have breast cancer:	Of the 1,070 women who received a positive result on their mammogram:
80 received a positive result on their mammogram.	80 have breast cancer.
Of the 9,900 women who do not have breast cancer:	Of the 8,930 women who received a negative result on their mammogram:
990 received a positive result on their mammogram.	20 have breast cancer.
Imagine another random sample of 10,000 women who had a mammogram.	Imagine another random sample of 10,000 women who had a mammogram.

The traditional condition-focus problem presentation presents information in terms of who has or does not have the condition of interest (C+ and C-, respectively), with additional information about the subsets of testing positive or negative within the two primary reference classes. This form of the problem is considered to be incongruent with the PPV question because each focuses on a different reference class than the T+ class needed for solution. Research has long called for a manipulation that constructively changes the way in which reasoners think about and conceptualize these types of problems to improve intuitive judgment and draw attention to the correct reference class (e.g., Bar-Hillel, 1980; Fischhoff & Bar-Hillel, 1984; Girotto & Gonzalez, 2001; Hoffrage, Hafenbrädl, & Bouquet, 2015).

Talboy and Schneider (2018a, 2018b) attempted to draw attention to the correct reference class by changing how information was organized within the problem. Instead of organizing the subsets of test results into the condition reference classes, I organized the subsets of the condition's presence into the reference classes of those who test positive (T+) and those who test negative (T-). This "test-focus" presentation is congruent with the PPV question as both utilize the same reference class information structure. A congruent test-focus presentation paired with the PPV question (congruent) resulted in over 90% of participants routinely determining the correct solution without any aids compared to less than 30% determining the correct solution on the incongruent condition-focus problems paired with PPV questions (Talboy & Schneider, 2018a, 2018b). These findings demonstrate that when reasoners evaluate a problem that is organized to focus on the same information as the diagnostic question of interest, they are able to consistently identify the correct posterior likelihood of having the condition given a positive test.

The benefit of using a congruent problem-question pairing is attributed to using reference dependence to ease the difficulties associated with both representation and computation of solutions (Talboy & Schneider, 2018a). With a congruent pairing, the organization of the problem information maps directly to the question of interest. Therefore, reasoners do not need to mentally re-structure the problem to get to the requested solution as they would with an incongruent problem-question pairing. This also eliminates computational difficulties in the frequency format because the correct reference class total (i.e., needed denominator) is provided as a value that can be selected directly from the problem rather than a value that must be calculated from two component pieces that are split between the alternate set of reference classes. Therefore, the requirements for solution on the congruent pairing were drastically reduced compared to the more traditional incongruent pairing (e.g., Kotovsky & Simon, 1990; Simon, 1978).

Across all three proposed experiments, I evaluated how reference dependence influences accuracy by using both congruent and incongruent problem-question pairings. For each pairing, I assessed how representational and computational difficulties interplay with the effect of reference class congruence on problem solving, as well as explored the relationship between these effects with respect to error response patterns and numerical skill.

Current Research

In the three studies proposed here, I evaluated the extent to which reference dependence, as well as related representational and computational difficulties, affect accuracy in Bayesian reasoning. Although each experiment is described separately, all conditions were randomly assigned simultaneously to enable cross-study comparisons. Across all three proposed experiments, there is a shared set of general hypotheses, as well as aspects of the stimuli and materials that was common to all. A map of the conditions across the three experiments is provided in Appendix A.

General Hypotheses

Bayesian reasoning problems were presented in either congruent or incongruent problem-question pairings. Congruent test-focus and PPV pairings utilize the same reference class information in both the problem and diagnostic question of interest, which makes the test-focus information problem formulation congruent with the PPV question (C+T+/T+). Incongruent condition-focus and PPV pairings use different reference classes in the problem versus the question, which creates a mismatch between how information is presented and what is required to determine the solution. Therefore, because condition-focus problem formulation focuses on the condition reference classes, it is incongruent with the PPV question that focuses on the test reference class. Based on the power of reference dependence in the conceptualization of the problem, as confirmed in our previous research, I proposed the primary general hypothesis:

H1) *Congruence hypothesis*: Congruent problem-question pairings will result in <u>higher problem solving accuracy</u> on average than incongruent pairings

because the same reference structure provided in the problem is required for solution.

In addition to evaluating reference dependence through the use of congruent and incongruent problem-question pairings, I also evaluated individual's response patterns to determine which types of errors reasoners are making. Error patterns observed in previous research suggests an identifiable pattern of responses that corresponds to values provided in the problem text rather than computational mistakes. This led to the second general hypothesis:

H2) *Value selection bias hypothesis*: Uninitiated reasoners will be more likely to select values as they are explicitly presented in the problem for their solution rather than completing computations to determine the correct response. Therefore, the majority of errors in response patterns will <u>conform to identifiable values from the problem</u>.

Finally, I evaluated accuracy and response patterns on Bayesian reasoning tasks with regard to numerical skill. I was specifically interested in those who have the most difficulty on these types of problems, which requires separating out those with the lowest numerical skillset from those who demonstrate higher numeracy. This led to the third general hypothesis:

H3) *Numeracy hypothesis*: In general, those with higher numeracy will display <u>higher accuracy</u> than those with low numeracy, especially for problems that require computations compared to problems without computations.

Shared Stimuli and Measures

Eight Bayesian reasoning problems were used to test participants' abilities to understand and calculate the positive predictive value (PPV). Table 3 includes the domain, topic, and frequency versions of the base rate, true positive rate, and false positive rate for all eight problems. I used two problems each from four unique content domains to ensure generalizability of results across various areas in which diagnostic tests could be encountered.

Table 3. Condition-focus Essential Components of Each Inference Problem

		Base Rate		**True Positive Rate**	**False Positive Rate**	**PPV**	**SEN**
Domain	Topic	C+	N	(C+T+) | (C+)	(C-T+) | (C-)	%	%
Medical	Mammogram	100	10,000	80 | 100	990 | 9,900	8	80
Medical	Diabetes	50	10,000	48 | 50	4,975 | 9,950	1	96
Legal	Polygraph	50	1,000	47 | 50	47 | 950	37	83
Legal	Recidivism	156	1,000	130 | 156	220 | 844	50	94
Sports	Baseball	185	250	130 | 185	15 | 65	90	70
Sports	Tennis	2,800	10,000	2,000 | 2,800	1,100 | 7,200	65	71
College	Employment	140	200	70 | 140	10 | 60	88	50
College	Exam Prep	350	500	275 | 350	25 | 150	92	79

Note. Base Rate = the number of condition occurrences (C+) within the specified sample size. (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.

All reasoning problems start with a general preamble about the condition of interest, as well as the test used for detecting the condition. In each problem, the preamble emphasizes that the test is not always correct and that specific information regarding correct and incorrect results is provided in the remaining sections of the problem. The rest of the problem information is manipulated to conform to either a test-focus or a condition-focus problem presentation, which is then matched with a congruent or incongruent diagnostic question. An example of each problem form is provided in Appendix B.

For each problem, participants determined the PPV first as a frequency response and then as a percentage response immediately after. The frequency response format question asks, “Of the women from this new sample who test positive, how many do you expect to have breast cancer?” For this format, answers were given as an open-response requiring the correct identification of two relevant integer values (___ out of ___ people) provided in the correct order. The accuracy score for the frequency format was computed by totaling the number of correct integer-pair answers across the eight problems. Because minimal or no calculations were needed to give the correct frequency responses, only exact values were coded as correct. Many of the hypotheses focus on the frequency response format as it provides a distinction between values that can be selected directly from the problem (i.e., the numerator) and values that require computations (i.e, the denominator).

For the percentage answer format, responses were provided using a slider scale ranging from 0% to 100% to the question "If a woman from this new sample tests positive, what is the probability that she will have breast cancer?" The accuracy score for the percentage format was assessed by totaling how many correct estimates are provided on the eight reasoning problems. Based on previous studies (Galesic et al., 2009; Hoffrage & Gigerenzer, 1998; Talboy & Schneider, 2017, 2018b, 2018a), responses on the percentage format are coded as correct if the answer falls within ±5% of the correct response. This range allows for possible rounding errors that could be made during mental calculations as participants are only provided pencil and paper in lieu of calculators.

Participants also completed a numeracy scale to assess their level of numerical skill. Researchers are often unable to make an overarching assessment of numeracy because current prevalent measures tend to focus on myopic components such as subjective impressions of personal numerical abilities (Fagerlin et al., 2007), specific computational skills (Cokely et al., 2012), or understanding of numerical expressions of risk in the form of probabilities or percentages (Lipkus et al., 2001; Weller et al., 2012). Additionally, results are often clouded by low reliability in the measure as some of the questions are clearly related to individual components of numeracy, whereas other map onto specific statistical literacy skills.

Although an ideal scale has yet to be developed, the Abbreviated Numeracy Scale (ANS; Weller et al., 2012) has been used in prior studies to distinguish those who have low numeracy from those who have higher numeracy (Talboy & Schneider, 2018a, 2018b). Therefore, in the current studies, numeracy was measured using the ANS.

Power Analysis

The three proposed experiments were run contemporaneously, and participants were randomly assigned to one of 10 possible between-subjects conditions. Power analysis conducted prior to data collection indicates that these experiments require a minimum of 40 participants per cell to find smaller main effects (η^2 = .06) with a power = .80 and α = .05, and medium-large simple effects (η^2 = .16; Cohen, 1992). The determination of expected effect size was based on the effects found in previous research for congruence and numeracy (Talboy & Schneider, 2018a, 2018b). In total, 594 participants participated in these studies. However, 5 participants were removed prior to analysis due to computer issues or incomplete response sets. In total, 589 participants (66% female) completed all study requirements and were included for analyses.

Overview of Proposed Experiments

In Experiment 1, I evaluated how the amount of information provided in the problem description affects accuracy on Bayesian reasoning tasks. I did this by manipulating the total number of subsets described explicitly in the problem. By expanding the amount of information available in the problem text, reasoners may build a more elaborate mental model that better reflects the structure of the problem (Johnson-Laird, 1994; Kintsch & Greeno, 1985; Sirota, Juanchich, et al., 2014), which may in turn increase accuracy. However, this might instead make discrimination of needed values more difficult. This experiment also established a baseline measure for accuracy in problems with complete subset information, which is used for comparative purposes in Experiments 2.

In Experiment 2, I evaluated the role of computation in Bayesian reasoning. In this study, measuring accuracy and individual differences in response patterns on incongruent problem-question pairings (i.e., condition-focused problem with PPV question; CF-PPV) allows the opportunity to evaluate issues related to both incongruent and congruent pairings. First, I assessed the extent to which incongruent reference class totals introduce processing interference (e.g., Reyna, 2004) that inhibits uninitiated reasoners' ability to determine the correct reference class. I assessed individual response patterns to gauge the general bias toward selecting values directly from the problem for the solution rather than completing computations.

For the congruent problem-question pairings (i.e., test-focused problem with PPV question; TF-PPV), I separated out the extent to which an added computation reduces the overall effect of congruence on accuracy. Again, I assessed individual response patterns to evaluate the extent to which reasoners are biased toward selecting values from the problem rather than completing computations. For both incongruent and congruent problem-question pairings, numerical skill was evaluated with respect to the selection and computation of values needed for the solution.

In Experiment 3, I introduced a new manipulation of reference dependence to determine the extent to which reasoners rely on the reference structure provided in the verbal description of the problem to determine the solution. In both the incongruent and congruent problem-question pairings, I looked at accuracy and individual differences in response patterns, as well as the relationship between numeracy and accuracy.

Each experiment is described separately to allow for assessment of specific hypotheses within each experiment. Running them contemporaneously, though, allowed for comparisons across experiments that further clarify the extent to which these three issues are related, and affect accuracy in Bayesian reasoning. This research was approved by the University of South Florida's Institutional Review Board, shown in Appendix C with a copy of the informed consent.

Experiment 1 – Amount of Information

In each Bayesian reasoning problem, there are four unique subsets of interest (C+T+, C+T-, C-T+, C-T-). Generally speaking, the research conducted on Bayesian reasoning tends to utilize problem forms that only present the two subsets that are needed to calculate the PPV (Brase, 2014; Gigerenzer & Hoffrage, 1995; Micallef et al., 2012; Sirota, Juanchich, et al., 2014; Talboy & Schneider, 2018a). Within these condition-focused problems, the subset of those who test positive is provided within the context of having the condition (C+T+) or not having the condition (C-T+). The pair of complementary subsets are not typically included.

Findings are mixed regarding whether reasoners are aware that potentially important subset information is absent from problem presentations. In some cases, people appear to be insensitive to information that is intentionally left out of problems (Fischhoff, Slovic, & Lichtenstein, 1978; Hammerton, 1973; McDowell & Jacobs, 2017; McDowell, Rebitschek, Gigerenzer, & Wegwarth, 2016). In others, reasoners (when asked) may make simple assumptions about information that they believe should be present but is not (e.g., Hamm, Miller, & Drillings, 1988). Uninitiated reasoners tend to rely on surface features provided in the problem to guide how they determine the solution (Chi, Feltovich, & Glaser, 1981; Chi, Glaser, et al., 1981; Owen & Sweller, 1989; Swanson & Beebe-Frankenberger, 2004; Winner, Engel, & Gardner, 1980). Therefore, responses may conform to surface features made available in the problem whether this includes full or partial subset information.

Adding full subset information may allow reasoners to create a more complete mental model of the problem structure by clarifying all of the component parts and implying the interrelationships that make up the contingency matrix. With partial presentations, half of the subsets are not explicitly delineated. Research concerning mental models and conditional reasoning suggests that this type of implicit information may be encoded into the mental representation of the problem but in a way that is not immediately accessible for problem solving (Johnson-Laird, 1994) or it may be completely ignored (Fischhoff et al., 1978; Hammerton, 1973; McDowell & Jacobs, 2017; McDowell et al., 2016). As a result, reasoners tend to rely more on explicit information because those values are more readily accessible (Johnson-Laird, 1994). Given this, accuracy is expected to increase on problems with full subset compared to partial subset information (Girotto & Gonzalez, 2001; Johnson-Laird, 1994; Legrenzi & Girotto, 1995; Markovits & Barrouillet, 2002).

When the problem presents all of the component pieces, all of the information is explicitly available and thus readily accessible for creating a mental structure that can be manipulated to determine the solution. By having all of the components available, reasoners do not need to mentally manage as many pieces of information because they are available in the problem description, making it less likely for reasoning errors to occur (e.g., Markovits & Barrouillet, 2002). This led to the first hypothesis for amount of information:

H4a) *Mental models hypothesis*: Complete problem information will result in <u>higher problem solving accuracy</u> than partial problem information because it will facilitate the reasoner's development of a comprehensive mental model of the problem structure.

However, there is a competing hypothesis for what happens when the amount of information is manipulated. Instead of increasing accuracy, the inclusion of full subset information could decrease accuracy because reasoners will need to discriminate among more values than they would with partial subset information. In choice tasks, having several options can be overwhelming because the need to evaluate more information places a larger cognitive burden on reasoners (Greifeneder, Scheibehenne, & Kleber, 2010; Iyengar & Lepper, 2000; Johnson et al., 2012). For instance, reasoners may not be able to actively attend to all of the options provided when full subset information is given compared to when partial subset information is given (Iyengar & Lepper, 2000). Reasoners may also get confused about which pieces of information are necessary for solution when they must discriminate among a larger set of values that are all relevant to developing a deeper understanding of the interrelated problem structure but are not necessarily needed for solution. This led to the competing hypothesis for amount of information:

H4b) *Discrimination hypothesis*: Complete problem information will result in lower problem solving accuracy than partial problem information because reasoners will need to discriminate among a larger set of values.

These two experiment-specific hypotheses will be evaluated in conjunction with the three overarching hypotheses (congruence, value selection bias, and numeracy).

Method

Participants. Experiment 1 included 236 psychology undergraduates who participated in exchange for extra credit toward a course. Each participant was randomly assigned to one of four between-subjects conditions.

Design. Experiment 1 used a 2 x 2 Congruence (congruent, incongruent) x Information (partial subsets, full subsets) between-subjects design. The primary dependent variable was accuracy on the PPV question in the frequency response format (___ out of ___ people).

Bayesian reasoning problems were presented as either congruent or incongruent problem-question pairings. Congruent test-focus problems with PPV question pairings utilized the same reference class information to structure the problem and diagnostic question of interest. Incongruent condition-focus problems with PPV question pairings used different reference classes to structure the problem and question, which created a mismatch between how information was presented and what was required to determine the solution.

In typical Bayesian problem presentations, only partial subset information is provided (as shown in Table 2). For the condition-focus information structure used in the current study, partial information included the base rate of a condition within a given population (C+ out of N), along with the subset who will have the condition and test positive (C+T+), and the subset who will not have the condition but also test positive (C-T+). In this format, the two complementary subsets of those who test negative (when the condition is present or absent) were not provided. In the test-focus information structure, the base rate of testing positive within a random sample was given (T+ out of N). Additional subset information was provided about who will test positive and have the condition (C+T+), as well as the subset who will test negative but also have the condition (C+T-). The two complementary subsets of those who do not have the condition (and test positive or negative) were not provided.

When full information was given in the problem, all four subsets of those who have or do not have the condition and test positive or negative were clearly indicated within the appropriate reference class. An example of the congruent and incongruent problem that includes full subset information is included in Table 4.

The dependent variable was the number correct on the frequency response format for the eight Bayesian reasoning problems (range: 0-8 correct responses on both numerator and denominator response components). Participants also provided the PPV as a percentage value using a slider scale ranging from 0-100%. To examine the relationship between numeracy and response accuracy, numerical abilities were measured using the 8-item *Abbreviated Numeracy Scale* (ANS; Weller et al., 2012). The scale results in normally distributed scores and has demonstrated sufficient reliability and validity (Cronbach's α = .71; Weller et al., 2013).

Table 4. Example Presentations with Full Subset Information for the Mammography Problem

Incongruent Condition-Focus Problem – Full Subset Information
To determine whether a woman is at risk of breast cancer, doctors conduct mammogram screenings. Sometimes women test positive even when they should test negative or test negative when they should test positive. Here is some information for a random sample of 10,000 women who had a mammogram: In this sample of 10,000 women, 100 have breast cancer. Of the 100 women who have breast cancer: 80 received a positive result on their mammogram. 20 received a negative result on their mammogram. Of the 9,900 women who do not have breast cancer: 990 received a positive result on their mammogram. 8910 received a negative result on their mammogram. Imagine another random sample of 10,000 women who had a mammogram.
Congruent Test-Focus Problem – Full Subset Information
To determine whether a woman is at risk of breast cancer, doctors conduct mammogram screenings. Sometimes women test positive even when they should test negative or test negative when they should test positive. Here is some information for a random sample of 10,000 women who had a mammogram: In this sample of 10,000 women, 1,070 received a positive result on their mammogram. Of the 1,070 women who received a positive result on their mammogram: 80 have breast cancer. 990 do not have breast cancer. Of the 8,930 women who received a negative result on their mammogram: 20 have breast cancer. 8,910 do not have breast cancer. Imagine another random sample of 10,000 women who had a mammogram.

Procedure. The experiment was hosted on Qualtrics.com™. All data were collected during one-hour supervised sessions in a university computer lab equipped with 11 desktop computers, allowing for multiple participants in each session. General instructions were read to each group of participants, with additional instructions provided on each computer to guide participants as they independently completed the experiment. Participants first completed the Abbreviated Numeracy Scale. After numeracy items were completed, an experimenter provided participants with a pencil and a blank paper form numbered 1 through 8, and entered the reference code from the form into the computer (for tracking purposes). Participants then advanced to another instructions page that indicated the paper form should be used for any notes participants felt they needed to complete in the next section.

Each participant completed the eight problems as randomly ordered in Qualtrics.com™ (algorithm from Matsumoto & Nishimura, 1998). Each problem was presented by itself with the frequency response format question presented first. Then they moved forward to a second screen to answer the PPV question using the percentage response format. After completing the eight problems, participants were instructed to see the experimenter for their course credit. At that time, they were given an information sheet that provided a summary about Bayesian reasoning and references to selected articles.

Results

Analysis for Experiment 1 was completed in three parts. First, I evaluated the relationship between numeracy and accuracy on Bayesian reasoning tasks. Then, I analyzed the extent to which accuracy was affected by congruence and amount of information while controlling for differences in numerical skills. Then, response patterns were evaluated to determine the types of errors participants routinely made across the eight Bayesian reasoning problems.

Numeracy. Numeracy was measured using the 8-item Abbreviated Numeracy Scale (Weller et al., 2010). As anticipated, there was a strong positive relationship between numeracy and the accuracy of responses to the frequency response format of solutions provided in the Bayesian reasoning problems, $r(234) = .44$, $p < .001$. Consistent with previous findings (e.g., Garcia-Retamero & Galesic, 2010; Peters et al., 2006; Talboy & Schneider, 2017), stronger numerical skills generally corresponded to higher accuracy rates on the Bayesian reasoning tasks.

Congruence and amount of information. A 2 x 2 Congruence (congruent, incongruent) x Information (partial subsets, full subsets) analysis of covariance was used to evaluate the effects of the two primary between-subjects variables on accuracy while controlling for differences in numeracy. An initial check of covariance assumptions was completed to ensure that numeracy was not related to the independent variables of Congruence, $F < 1$, or Information, $F < 1$.

As expected, those who read congruent problem-question pairings ($M_{adj} = 3.59$) were more accurate than those who read incongruent problem-question pairings ($M_{adj} = 2.59$) even after controlling for differences in numerical skill, $F(1, 232) = 7.37$, $p = .007$, $\eta_p^2 = .04$. Although this was a much smaller effect than predicted based on previous findings, this provides at least weak support for the congruence hypothesis, which argues that the starting point provided by the problem presentation influences the reasoner's determination of the solution.

However, there was no discernable difference in accuracy between those who read problems presented with partial information ($M_{adj} = 3.02$) compared to full information ($M_{adj} =$

3.16), $F < 1$. Additionally, Figure 1 demonstrates that the interaction of congruence and amount of information was not significant, $F(1, 232) = 1.50$, $p = .22$. Against expectations, I did not find support for the mental models hypothesis or the discrimination hypothesis. Providing full information was expected either to increase accuracy by helping reasoners extrapolate a more complete mental model of the problem or to decrease accuracy by forcing reasoners to discriminate among additional problem values. Within the congruent and incongruent pairings, though, accuracy did not appear to significantly change as a function of whether partial versus full information was provided.

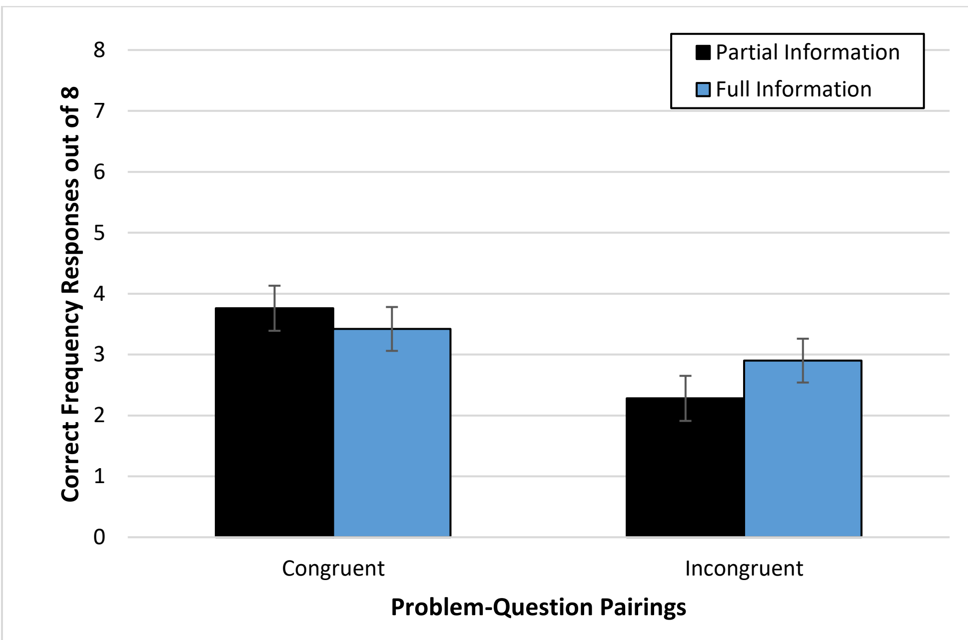


Figure 1. Combined effects of congruence and amount of information on accuracy.

Response patterns. I also analyzed denominator response patterns for each participant across the eight Bayesian reasoning problems to determine if any particular problem-relevant but incorrect value was routinely indicated (e.g., T+, C+, N). Consistent with the congruence hypothesis, previous research has shown that reasoners have a value selection bias in which they typically utilize the reference values provided in the problem presentation as part of their response, even when calculations are required (Talboy & Schneider, 2017, 2018a). Therefore, the majority of responses were expected to conform to the T+ reference class in the congruent pairings and the C+ reference class in the incongruent pairings, with a small portion of reasoners using other reference values provided in the problem such as N (i.e., the total sample).

If participants provided four or more responses that conformed to the same value, this was coded as their predominant response strategy for the denominator of the frequency response format. Participants who did not consistently provide the same type of response across at least four problems, but generally selected values from the problem text for the same answer component were coded as "Other Selected." Those who did not consistently provide the same type of responses and did not appear to be routinely selecting values from the problem were coded as "Other."

Based on the congruence hypothesis, the percentage of participants who consistently identify the correct response (i.e., at least four out of eight responses) was expected to be higher on congruent problem-question pairings than incongruent pairings when either partial or full subset information is provided (e.g., Talboy & Schneider, 2018a). However, as shown in Figure 2, the proportion of those who consistently identified the correct denominator value did not differ between congruent and incongruent problem-question pairings.[1] In each group, approximately half of the participants were able to determine the correct value for the PPV denominator on at least four of the eight problems. This finding will be evaluated in more depth in the next section.

[1] Although the proportion of participants who consistently identified the correct denominator appeared to be similar, the metric used for strategy evaluation did not take differences in rates of identification into account. Of those who consistently identified the correct denominator (at least 4 out of 8), the average correct was higher on both response components for congruent pairings (M = 6.91, SD = 1.12 for denominators and M = 6.90, SD = 1.46 for numerators) than incongruent pairings (M = 6.15, SD = 1.29 for denominators and M = 5.65, SD = 2.60 for numerators).

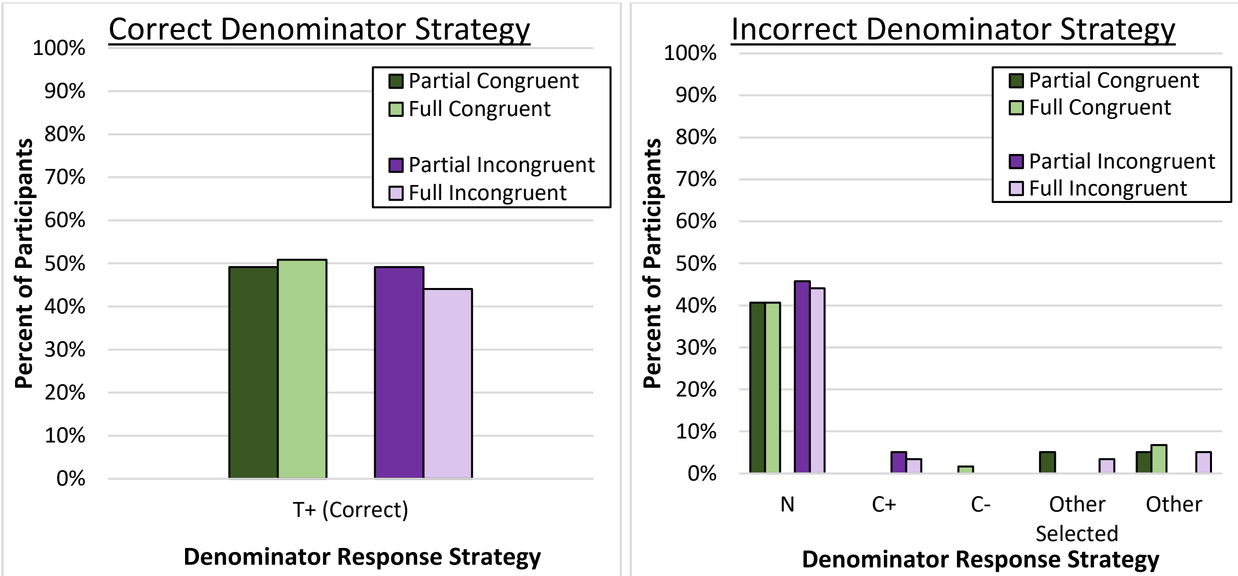


Figure 2. Proportion of participants who consistently used the correct denominator strategy (left) or the incorrect denominator strategy (right) on the frequency response format. C± denotes the total number of people who have or do not have the condition. T± denotes the total number of people who test positive or negative. Total N denotes the total in the superordinate set (i.e., sample size).

In addition to correct responses, I was particularly interested in determining which values were provided when responses conformed to incorrect values. I expected to find a substantial portion of participants consistently using the focal reference class (either T+ or C+ depending on congruence) as their denominator on the frequency response format (e.g., Talboy & Schneider, 2018a). However, a different pattern of response errors was observed (right, Figure 2).

The predominant incorrect strategy was consistent with the overall sample size (N) rather than the conflicting reference class. Few, if any, participants in the congruent or incongruent pairings consistently utilized the C+ reference class or any other provided values. Only a handful of reasoners provided responses that were not present in the problem. This pattern, which is virtually identical for both the full and partial information conditions, provides support for the value selection hypothesis in that reasoners were utilizing a prominent value from the problem for their denominator response. However, the finding is inconsistent with previous research that shows reasoners are typically drawn to the focal reference class more so than any other problem value.

Discussion

In Experiment 1, I found evidence in favor of the three primary hypotheses of numeracy, congruence, and value selection bias. With regard to numeracy, there was a strong positive relationship with accuracy on the Bayesian reasoning tasks. As predicted by the general numeracy hypothesis, those with higher numerical skill were more accurate than those with lower numerical abilities.

I expected a relatively large effect of congruence in which reasoners would be more accurate when starting from a congruent problem presentation than from one that focused on the incongruent reference classes. Although accuracy was generally higher for those who read congruent pairings than those who read incongruent pairings, this effect was much smaller in the current experiment compared to previous studies (e.g., Talboy & Schneider, 2018a, 2018b). Additionally, the inclusion of partial versus full subset information did not appear to affect accuracy for either the congruent or incongruent pairings.

The substantially reduced effect of congruence could be the result of what I thought were minor changes to the problem presentation. Compared to previous problem formulations, the problem form used in the current studies asked reasoners to imagine another sample of the same size prior to reading the PPV question. Introducing this statement could have inadvertently suggested to participants that the superordinate value indicating the size of that sample (N) should be viewed as a focal reference point, making this value much more salient than anticipated. Additionally, the PPV question was simplified to reduce redundancy of indicators pointing to the correct reference class, which could have downplayed how important the T+

value was for solution. These changes may have eliminated any benefit that would have occurred when full subset information was provided compared to when only partial subset information was available.

When I evaluated response errors to determine what might be drawing attention away from the correct reference class, hardly any reasoners in either the congruent or incongruent conditions showed evidence of incorrect computations, as the vast majority utilized values that were presented within the problem. Consistent with the value selection prediction, the majority of responses for the denominator conformed to a value presented in the problem. However, reasoners were not selecting the focal reference class as predicted, but instead were predominantly selecting the superordinate value (N). This indicates that reasoners were drawn toward using the overall sample size as their reference value rather than the expected focal reference class. I suspect that the apparent change in focus from the focal reference class to the overall sample size altered the observed response strategies compared to those documented in previous studies (e.g., Talboy & Schneider, 2017, 2018a). Nevertheless, the surprise finding provides confirmation of the importance of salient reference points as a primary determinant of answers to Bayesian reasoning problems.

Experiment 2 – Selection versus Calculation

Although providing problem-question pairings with partial or full subset information focused on a representational issue regarding identification and application of values from the problem to the solution, it did little to elucidate the potential role of computation as an obstacle to Bayesian reasoning. For both congruent and incongruent problem-question pairings, the first value needed for the pair-of-integers frequency response format (i.e., the numerator) could be selected directly from the problem presentation. However, the second value (i.e., the denominator) had to be computed in the incongruent pairing but could be directly selected from the problem in the congruent pairing.

Study 2 was designed to address this inherent confound, and further assess the extent to which the reference dependence hypothesis holds when calculations are required. To do this, full subset information was provided within both congruent and incongruent problem presentations, but without reference class totals in either case. By removing these totals, participants could no longer directly select and apply these values as their responses in either congruent or incongruent pairings. This ensured that both types of pairings required the simple computation of adding two subsets, and neither had the potential for interference from an inappropriate reference class total.

Incongruent Pairings

Regardless of whether partial or full subset information was provided within incongruent pairings, reasoners had to identify and calculate the denominator of the frequency response format to determine the correct solution. (This is also necessary in standard forms of Bayesian reasoning problems.) This value (T+) was partitioned into two pieces of information as subsets of the alternate reference classes (C+T+ and C-T+). Therefore, reasoners had to first identify the two subset values correctly to then compute the total of the relevant test reference class. Although this is similar to what is required in the standard form of Bayesian reasoning problems, it is not clear in the standard form whether the difficulty associated with this step of the reasoning process is primarily related to recognizing which values are needed to sum to T+ or overcoming the pull of the incongruent reference class.

Although results from Experiment 1 suggest that reasoners were drawn to the superordinate set rather than the alternate C+ reference class total, many previous studies indicate that reasoners routinely use the competing C+ reference class value as their preferred denominator (e.g., Cosmides & Tooby, 1996; Talboy & Schneider, 2018a, 2018b). Given the literature (coupled with the fact that the dissertation studies were run contemporaneously), I predicted that when the competing C+ total is omitted from the problem presentation, two different response patterns could emerge, which could help disentangle issues associated with computational requirements from issues more closely tied to reference dependence.

Eliminating the presence of the alternate C+ reference class totals may increase accuracy because I am removing a value that is hypothesized to interfere with reasoning about the correct nested set (e.g., Reyna, 2004; Reyna & Brainerd, 2008). This could potentially reduce the tendency to adopt the inappropriate total and thereby help reasoners recognize the need to engage in the computational process of calculating values to reach the solution. Though if, as in Experiment 1, reasoners are drawn to the superordinate set, the hypothesized increase in accuracy may not be as strong as originally hypothesized.

Conversely, removal of the alternate C+ reference class totals may have little effect on accuracy or may result in even lower accuracy rates than in standard problem forms. This would be expected if reasoners have a strong bias toward selecting values directly from the problem text, rather than computing responses. In particular, the superordinate set is the only remaining higher order value in the problem text that can be used as an organizational cue. This value appears to draw reasoners away from the alternative C+ reference class, as demonstrated in Experiment 1. As a result, removing other competing values may shift reasoners to reliance on the overall sample value or, if removal of the reference class creates additional uncertainty about how to proceed, it may push even more reasoners toward using the overall sample as their preferred denominator compared to the proportion observed in Experiment 1. Therefore, if value selection bias is supported, there should be a sizable portion of reasoners who utilize the superordinate set as a frequency response component.

This led to the competing hypotheses for incongruent pairings in Experiment 2:

H5a) *Interference reduction hypothesis*: For incongruent pairings, removing the competing reference class totals from the problem text may result in <u>higher problem solving accuracy</u> than when reference class totals are provided. If focusing on the competing reference class total is interfering with the problem solving process, removing the competing value should reduce interference, thus increasing accuracy.

H5b) *Value selection hypothesis*: To the extent that the overriding response tendency is to select values from the problem rather than computing values for solution, any effects of interference reduction may be relatively small with response errors conforming to values presented in the problem, especially the value representing the superordinate set or overall sample.

Congruent Pairings

Unlike incongruent problem-question pairings, previous congruent pairings did not require calculations for the denominator of the frequency response because the relevant reference class totals were provided. In previous studies, this precluded isolation of the most important reason for accuracy increases. Did accuracy improve because the problem-question pairing was congruent or because no calculations were required? To address this confound, Experiment 2 compared performance on test-focus problems that included or omitted reference class totals. When reference class totals are omitted, reasoners need to complete the computational step of adding subsets, which is equivalent to what is typically required in the incongruent pairings.

By adding this computation step, I am able to evaluate the extent to which congruence affects accuracy when computational difficulty is comparable for both congruent and incongruent problem-question pairings. Although I still predict that congruence will result in higher accuracy than incongruence even when calculations are required, I do expect a decrease in accuracy on the congruent pairings when there is an added step involving computation.

This decrease in accuracy is predicted to be the result of two different mechanisms. First, even for those who attempt the computation, accuracy may decrease because the arithmetic step may be completed incorrectly, either resulting in "quasi-Bayesian responses" (i.e., responses that use the correct component values but are not combined correctly; Macchi, 2000) or incorrect values. With the first, there should be an increase in responses that is close to correct but computationally inaccurate. With the second, accuracy could decrease due to reasoners' bias in favor of selecting values over computing values (Talboy & Schneider, 2018a). Therefore, there should be an increase in other values from the problem being utilized as the denominator.

H6) *Selection and computation hypothesis*: Removal of reference class totals in congruent pairings is expected to result in <u>lower problem solving accuracy</u> both because of (a) the required computation step adds complexity that is not present when the needed reference class totals are provided in the problem and (b) the bias to select values provided in the problem.

Method

Participants. For Experiment 2, there were 118 psychology undergraduates who were randomly assigned to one of two between-subjects conditions that provided either congruent or incongruent problems without reference class totals. Their performance was compared to data from the two conditions in Experiment 1 that presented full subset information with reference class totals. In all, data from 236 participants were analyzed in Experiment 2. (All conditions were randomly assigned at the same time.)

Design. This experiment employed a 2 x 2 Congruence (congruent, incongruent) x Reference Class Totals (included, omitted) between-subjects design. The primary dependent variable was accuracy on the PPV question in the frequency response format (___ out of ___ people). All four problem-question variations utilized in this experiment included full subset information in either an incongruent condition-focus or a congruent test-focus presentation.

To isolate performance issues, information about the reference class totals was either provided (in the congruent and incongruent full information conditions from Experiment 1) or omitted (newly introduced conditions of Experiment 2). When reference class totals were provided, those who read congruent pairings did not need to complete any calculations to determine the correct denominator as the value could be selected directly from the problem. However, incongruent pairings required an additional computation step to determine the denominator of the frequency response. An example of how these problems appeared when reference class totals were included is shown in Table 4 from Experiment 1.

When the reference class totals were omitted, both the congruent and incongruent pairings required the same level of computation to determine the correct solution. An example of how the problems appeared when reference class totals were omitted is shown in Table 5.

As before, the ANS was used to measure numeracy. The procedure was identical to Experiment 1.

Table 5. *Example Presentations without Reference Class Totals for the Mammography Problem*

Incongruent Condition-Focus Problem without Reference Class Totals
To determine whether a woman is at risk of breast cancer, doctors conduct mammogram screenings. Sometimes women test positive even when they should test negative or test negative when they should test positive. Here is some information for a random sample of 10,000 women who had a mammogram: In this sample of 10,000 women: Of those who have breast cancer: 80 received a positive result on their mammogram. 20 received a negative result on their mammogram. Of those who do not have breast cancer: 990 received a positive result on their mammogram. 8910 received a negative result on their mammogram. Imagine another random sample of 10,000 women who had a mammogram.

Congruent Test-Focus Problem without Reference Class Totals
To determine whether a woman is at risk of breast cancer, doctors conduct mammogram screenings. Sometimes women test positive even when they should test negative or test negative when they should test positive. Here is some information for a random sample of 10,000 women who had a mammogram: In this sample of 10,000 women: Of those who received a positive result on their mammogram: 80 have breast cancer. 990 do not have breast cancer. Of those who received a negative result on their mammogram: 20 have breast cancer. 8910 do not have breast cancer. Imagine another random sample of 10,000 women who had a mammogram.

Results

Numeracy. Numeracy was again assessed in relation to accuracy on the frequency response format using the Abbreviated Numeracy Scale. As found in Experiment 1, results indicated a strong positive relationship between numerical abilities and accuracy on Bayesian reasoning tasks, $r(234) = .47$, $p < .001$. Again, I confirmed that stronger numerical skills are related to higher levels of accuracy on Bayesian reasoning tasks.

Congruence and reference class totals. A 2 x 2 Congruence (congruent, incongruent) x Reference Class Totals (included, omitted) analysis of covariance was used to assess the effects of the two primary between-subjects variables on accuracy while controlling for numeracy. As with the first experiment, an initial check of covariance assumptions was completed and indicated that numeracy was not significantly different within the independent variables of congruence, $F < 1$, and reference class totals, $F < 1$.

Against expectations, the main effect of congruence on accuracy was not significant when controlling for numeracy, $F < 1$. Accuracy was comparable, and relatively low, regardless of whether participants viewed congruent problem-question pairings ($M_{adj} = 3.54$) or incongruent problem-question pairings ($M_{adj} = 3.53$). Counter to previous findings, there was no evidence to suggest that congruence between the problems and questions increased accuracy compared to when incongruent pairings were provided. Additionally, there was no main effect of reference class totals, mean accuracy was similar when reference class totals were provided ($M_{adj} = 3.18$) or omitted ($M_{adj} = 3.89$), $F(1, 231) = 3.6$, $p = .06$, $\eta_p^2 = .02$. This is not entirely surprising as removal of the reference class totals from the congruent pairings was expected to decrease accuracy, whereas removal of reference class totals from the incongruent pairings was expected to increase accuracy.

Pivotal to our hypotheses, however, was the Congruence x Reference class Total interaction effect. As shown in Figure 3, this effect was not significant, $F(1, 231) = 1.79$, $p = .18$. Nevertheless, given specific predictions related to the simple effects of reference class totals inside of congruent and incongruent problem questions pairings, I tested these to see if there was any evidence of predicted effects.

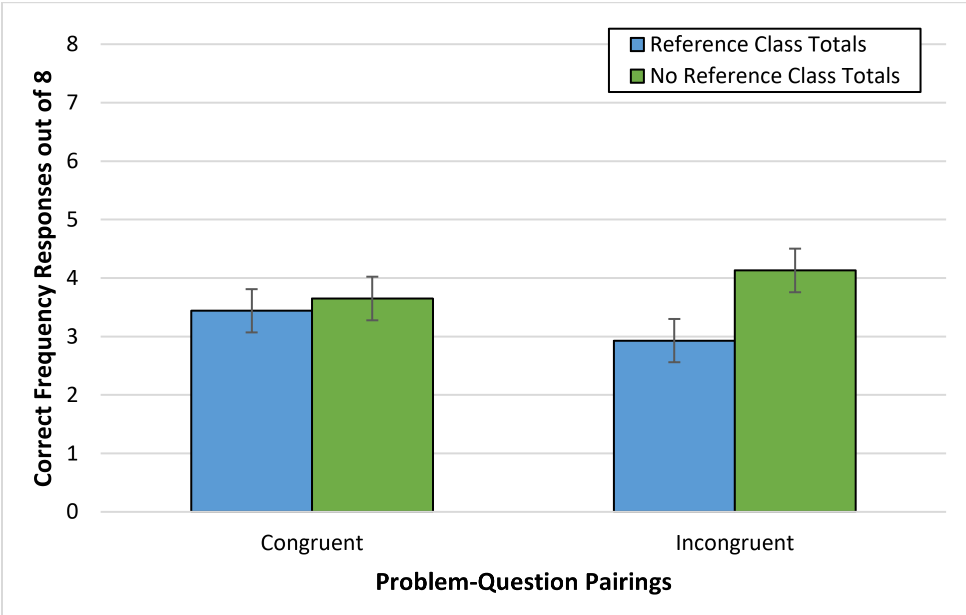


Figure 3. Results for the combined effects of congruence and reference class totals while controlling for numeracy on adjusted means of accuracy. Error bars represent ±1 SE.

For the congruent pairings, I predicted that accuracy would be lower when reference class totals were omitted from versus included due to the added requirement to compute the denominator value. However, there was no evidence that the presence or absence of the correct reference class total significantly affected accuracy in the congruent pairings, $F < 1$. Generally speaking, it appears that reasoners were not thrown off by the need to compute as roughly the same proportion determined the correct denominator regardless of whether the value was provided directly or had to be calculated from the subset information. Nevertheless, as noted in Experiment 1, accuracy rates were unexpectedly low in both conditions. Although accuracy was much lower than that documented in previous research (Talboy & Schneider, 2018a, 2018b), the lack of change in accuracy is surprising and goes against our hypothesis based on the general presumption that, all else equal, reasoners perform worse on problems that require calculation (e.g., McDowell & Jacobs, 2017).

Within the incongruent pairings, I expected that removing the competing reference class totals from the problem might increase accuracy compared to when these values were included. However, the extent to which interference was reduced was also predicted to be relatively small, and possibly compromised by added confusion about how to proceed in answering the problem without a value corresponding to the implied reference class.

The simple effect of reference class totals on accuracy in the incongruent problems supports the interference hypothesis, albeit weakly, $F(1, 231) = 5.21$, $p = .02$, $\eta_p^2 = .02$. This relatively small effect suggests that when reference class totals from the incorrect reference class are present, they may cause at least some interference misleading reasoners about which values are relevant to problem solution. Evaluation of response patterns are considered next to elucidate which values participants were most likely to consistently provide when answering the PPV question, with specific attention to the value selection hypothesis.

Denominator response strategies. I again assessed patterns of denominator responses to determine if errors could be attributed to selection of values given in the problem (and, if so, which ones) or incorrect computations. As in Experiment 1, response patterns for each participant were evaluated to determine if responses errors consistently conformed to identifiable response patterns (on at least four of the eight problems).

In both the congruent and incongruent pairings, regardless of whether the reference class totals were provided, a large majority of incorrect responses aligned with the superordinate set (N) as shown in Figure 4. As predicted by value selection bias hypothesis, reasoners were latching on to the only remaining reference value provided, which was the superordinate value.

As shown in the two conditions discussed in Experiment 1 when reference class totals were provided, very few reasoners consistently used the C+ reference class, even when it was presumably salient in the incongruent condition. This same pattern was also observed when no reference class values were provided. Reasoners from the two new conditions in this experiment also appear to be selecting the overall sample size when they are not sure or cannot (or will not) complete the computations required to determine the correct response.

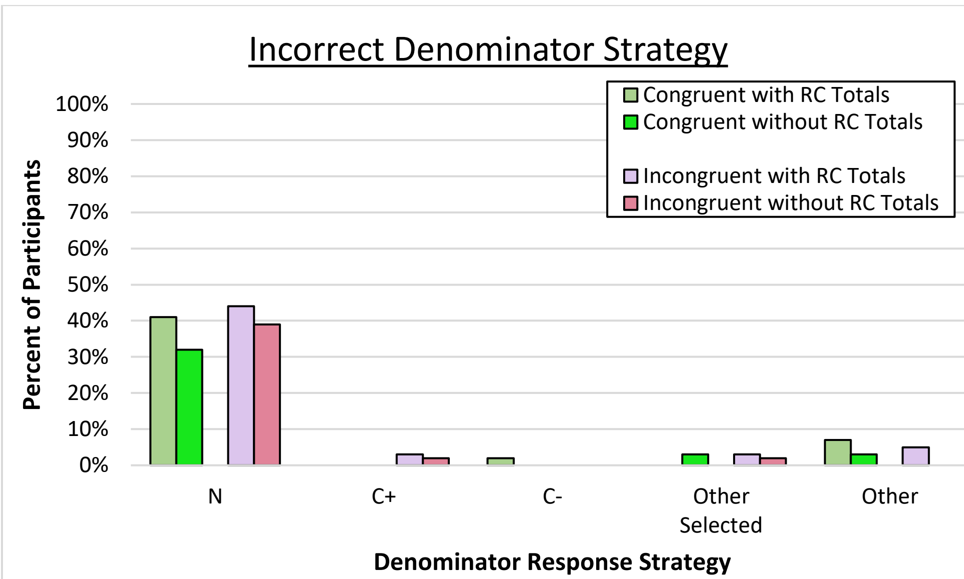


Figure 4. Proportion of participants who consistently used an incorrect denominator strategy on the frequency response format. C± denotes the total number of people who have or do not have the condition. Total N denotes the total in the superordinate set (i.e., sample size).

Numeracy and calculations. In this study, there was an additional numeracy prediction that the performance advantage for those high in numeracy would be particularly strong for problems that required calculation to determine the correct response compared to problems without computations. The relationship between numeracy and accuracy was fairly strong on congruent problems that required adding two values together to determine the correct denominator, $r(57) = .45$, $p < .001$, as well as on those that did not require calculation, $r(57) = .32$, $p = .01$. Although in the predicted direction, the difference between these two correlation coefficients was not significant, $z = 0.81$, $p = .21$. Thus, there is insufficient evidence to conclude that numeracy was more or less important when calculations were required to arrive at the correct solution compared to when the value could be selected directly from the problem, and that at least some level of numerical proficiency was required to determine the correct response in either case.

Discussion

In Experiment 2, I was particularly interested in the role of calculation in solving Bayesian reasoning tasks. I found additional support for the general relationship between numeracy and accuracy, as well as the value selection bias. However, the primary hypothesis of congruence was not supported in this study. Reasoners who read congruent problem-question pairings performed about the same on these tasks as those who read incongruent pairings. Although stronger numeracy predicted higher accuracy on Bayesian reasoning tasks, there was insufficient evidence to suggest numeracy was more or less important when calculations were required on the congruent pairings. Further, incorrect responses from both congruent and incongruent pairings virtually always corresponded to values provided in the problem rather than computation errors.

As in Experiment 1, the two newly introduced conditions of Experiment 2 also showed that the majority of reasoners who did not determine the correct response in either the congruent or incongruent pairings consistently utilized the superordinate set (N) rather than any other value provided in the problem presentation. In line with the value selection bias, it appears that removing reference class totals from the incongruent problems pushed reasoners to look for other salient reference values (such as N) to use as part of their solution rather than completing calculations.

Experiment 3 – Reference Dependence

Experiment 3 provides a new assessment of the role of reference dependence in Bayesian reasoning by altering the nesting of subset information. This was done by evaluating performance on problems that organized the four subsets into two related reference classes (C+ and C- or T+ and T-) or a single superordinate set (N), or provided no explicit organizing information (i.e., an unlabeled nesting). By directly manipulating the organization of the subsets into different types of nested structures, I can further assess the extent to which the presence or type of key reference values affects reasoners' abilities to determine the PPV in Bayesian reasoning tasks.

Reference Dependence

The majority of Bayesian reasoning tasks explicitly organize the subsets of interest into reference classes, through both verbal and visual nesting. Talboy and Schneider (2018a, 2018b) found that the majority of participants utilized the focal reference class as part of their solution for PPV questions, particularly on the denominator of the frequency response format. When the problem-question pairings were congruent, this resulted in the correct response as the focal reference class from the problem (T+) was required for solution. However, when the problem-question pairings were incongruent, reasoners typically used the alternate C+ reference class as their denominator of the frequency response (Cosmides & Tooby, 1996; Galesic et al., 2009; Gigerenzer & Hoffrage, 1995; Gigerenzer, Hoffrage, & Kleinbölting, 1991; Talboy & Schneider, 2017, 2018a; Wolfe, Fisher, & Reyna, 2013).

Reference dependence suggests that reasoners will rely on the context in which the subsets are organized to determine the solution to Bayesian reasoning problems. In the problem formulations from Experiments 1 and 2, there were two ways reasoners were encouraged to use the problem structure to determine the solution to PPV questions. First, a visual structuring (indentation) demonstrated how the subsets were nested within each of the indicated reference classes. Second, there was a verbal organization describing which subsets belong to which reference classes, along with information that all of the values were drawn from the superordinate set.

Results from Experiment 2 suggested that in the absence of the value for the nesting reference class, decision makers, especially in incongruent problems, rely on another potential reference class value: the superordinate set size (N). Results from both Experiments 1 and 2 suggest I inadvertently reinforced the superordinate set as another prominent reference point by adding an additional statement about this value into the problems. When reasoners did not determine the correct denominator values in those two studies, the majority of them used the superordinate value (N) for their denominator instead of the focal reference classes from the problem. In Experiment 3, I explicitly test the influence of drawing attention to the potential relevance of this value by introducing a problem formulation in which the visual organization created nesting under the superordinate set rather than the condition or test reference classes. An example of superordinate set nesting is provided in Table 6. By visually nesting all four possible subsets into the superordinate set, the superordinate becomes the explicit reference class within which the subsets are organized or nested.

Table 6. *Example Presentations with Superordinate Set Organization for the Mammography Problem*

Incongruent Condition-Focus Problem with Superordinate Set Organization
To determine whether a woman is at risk of breast cancer, doctors conduct mammogram screenings. Sometimes women test positive even when they should test negative or test negative when they should test positive. Here is some information for a random sample of 10,000 women who had a mammogram: In this sample of 10,000 women: 80 have breast cancer AND received a positive result on their mammogram. 20 have breast cancer AND received a negative result on their mammogram. 990 do not have breast cancer AND received a positive result on their mammogram. 8910 do not have breast cancer AND received a negative result on their mammogram. Imagine another random sample of 10,000 women who had a mammogram.
Congruent Test-Focus Problem with Superordinate Set Organization
To determine whether a woman is at risk of breast cancer, doctors conduct mammogram screenings. Sometimes women test positive even when they should test negative or test negative when they should test positive. Here is some information for a random sample of 10,000 women who had a mammogram: In this sample of 10,000 women: 80 received a positive result on their mammogram AND have breast cancer. 990 received a positive result on their mammogram AND do not have breast cancer. 20 received a negative result on their mammogram AND have breast cancer. 8910 received a negative result on their mammogram AND do not have breast cancer. Imagine another random sample of 10,000 women who had a mammogram.

If reference dependence plays a primary role in how people go about solving these problems, this superordinate reference point should function similarly to the condition-focus reference class in the standard incongruent problem-question pairing from previous research (e.g., Talboy & Schneider, 2018a, 2018b). Rather than determining the correct reference class total of all those who test positive (T+), which is the appropriate denominator for the PPV question, reasoners were expected to utilize the superordinate set as the reference class for their solution because it is the focal reference value presented in the problem. Because this appeared to be such a focal reference point in the previous two studies, I expect that the proportion who use this value will be relatively large in all three problem formulations for Experiment 3, with its use being highest when the subsets are organized into the superordinate value.

In this superordinate nesting, the four subsets were still organized to be congruent or incongruent with the PPV question, though the organization was more implicit than in previous manipulations. In this case, congruence refers only to the ordering of the subsets and the ordering of the features within each subset statement. For the congruent test-focus problem, the test result was always indicated prior to the condition. For the incongruent condition-focus problem, the condition was always indicated prior to the test result. This type of organization provided an internal, embedded structure that reasoners could still utilize as a reference point for their deliberations, although the manipulation was much more subtle.

The implicit congruence of the problem with the PPV question becomes more important when all explicit reference points are eliminated from the problem presentation, as they were in the second novel problem formulation created for this experiment. The second problem formulation removed all explicit structural reference cues such as the reference classes or superordinate set (except, in hindsight, for the statement about imagining another random sample of the same number of people). In this problem formulation shown in Table 7, there were no visual or verbal nesting components that explicitly indicated how the subsets were related to one another. The implicit level of congruence or incongruence was retained, though, as the four subsets were organized to focus on either the test result or the condition status first, with the sentences belonging to each congruent or incongruent reference class occurring together.

Although the superordinate set and reference classes could be computed based on the values provided if the reasoners determined they were necessary for solution, there were no other explicit reference values within the problem presentation that could be used as a guide. (Note. The superordinate value was not present in this version of the statement asking participants to imagine a new random sample, so it could not be used as a guide as it might have been in other conditions.) The implicit verbal ordering of the reference classes within individual descriptions of each subset may not be as helpful as providing an explicit anchor, but it might still be used as an organizational cue for solution.

Table 7. *Example Presentations with No Explicit Organization for the Mammography Problem*

Incongruent Condition-Focus Problem with No Explicit Organization
To determine whether a woman is at risk of breast cancer, doctors conduct mammogram screenings. Sometimes women test positive even when they should test negative or test negative when they should test positive. Here is some information for a random sample of women who had a mammogram: 80 have breast cancer and received a positive result on their mammogram. 20 have breast cancer and received a negative result on their mammogram. 990 do not have breast cancer and received a positive result on their mammogram. 8910 do not have breast cancer and received a negative result on their mammogram. Imagine another random sample of the same number of women who had a mammogram.
Congruent Test-Focus Problem with No Explicit Organization
To determine whether a woman is at risk of breast cancer, doctors conduct mammogram screenings. Sometimes women test positive even when they should test negative or test negative when they should test positive. Here is some information for a random sample of women who had a mammogram: 80 received a positive result on their mammogram and have breast cancer. 990 received a positive result on their mammogram and do not have breast cancer. 20 received a negative result on their mammogram and have breast cancer. 8910 received a negative result on their mammogram and do not have breast cancer. Imagine another random sample of the same number of women who had a mammogram.

By removing the structural organization, our intention was to eliminate the reference class structure that intuitively led reasoners to rely on these values for solution. (For more on the defining role of structural organization in problem solving, see Kotovsky & Simon, 1990; Simon, 1973.) Therefore, accuracy was expected to decrease as the explicit structure on which uninitiated reasoners depend was removed, particularly the provision of a structural reference point. Even with this anticipated decrease in accuracy, though, the congruent pairings were still expected to result in, at least, slightly higher accuracy than incongruent pairings because of the congruent statements could be more readily matched to the order of what was being asked for in the problem.

H7) *Reference dependence hypothesis*: Accuracy will decrease when subsets are organized using the superordinate set compared to explicit congruent reference classes. Accuracy will decrease further when no explicit reference point is provided. This <u>decrease in problem solving accuracy</u> is expected due to the reduction in explicit contextual organization of the nested sets via provided useful/helpful reference classes.

Value Selection Bias

Consistent with what I found in Experiments 1 and 2, incorrect responses on the denominators were expected to coincide with key reference values provided in the problem presentations. The value selection bias hypothesis predicts that participants are more likely to offer given values rather than computing values to arrive at their answer. Taken together with predictions of the reference dependence hypothesis, I expected that reasoners would have a tendency to utilize whichever structural reference point was provided (either reference class or superordinate set value) as the anchor for answering the question of interest.

With regard to use of the superordinate set, results from Experiment 1 and 2 demonstrate that the majority of reasoners used this value for their denominator in conditions where this value was readily available. In problem structures that organized subsets into the reference classes or the superordinate value, I anticipated based on the previous results that a large proportion of reasoners would select N as their denominator.

When these structural reference points were removed, as they were in the third problem formulation, I expected reasoners' responses to be more erratic, with selection errors being more common than errors involving any kind of computation. Nevertheless, I predicted some tendency to rely on the implicit reference structure that was used to organize the subset statements for determining their solution.

Method

Participants. For Experiment 3, there were 235 psychology undergraduates who were randomly assigned to one of four novel between-subjects conditions. Their performance was compared to data from the two newly-introduced conditions in Experiment 2 that provided reference class organization with full subset information but no reference class totals. In total, data from 353 participants were analyzed in Experiment 3.

Design. This experiment employed a 2 x 3 Congruence (congruent, incongruent) x Organization (reference classes, superordinate set, none) between-subjects design. The dependent variable was average accuracy on the frequency response format for the eight Bayesian reasoning problems (range: 0-8 correct responses).

Each problem variation utilized in this experiment included full subset information with no condition or test reference class totals in either an incongruent condition-focus or a congruent test-focus presentation. In this way, all participants were required to complete computations to determine the correct denominator values. Problems were either nested in reference classes or the superordinate set, or were not give any explicit visual nesting. Numerical abilities were measured using the ANS.

The procedure was identical to Experiments 1 and 2.

Results

Numeracy. Numeracy was again assessed using the ANS to determine the relationship between numeracy and accuracy on the frequency response format. Consistent with results of Experiments 1 and 2, there was a strong positive relationship between numeracy and frequency response accuracy, $r(351) = .41$, $p < .001$. This provides further evidence for the general numeracy hypothesis suggesting that at least some level of numeric proficiency is required to perform well on these types of tasks.

Congruence. A 2 x 3 Congruence x Organization (reference class, superordinate, none) analysis of covariance was used to analyze the effects of congruence and problem organization on accuracy while controlling for numeracy. An initial check of the ANCOVA assumptions indicates that numeracy was not significantly different across the congruence conditions, $F < 1$, or organization conditions, $F(1, 343) = 2.69$, $p = .07$.

As found in Experiment 2, there was no main effect of congruence on accuracy when controlling for numeracy, $F < 1$. On average, those who read congruent problem-question pairings ($M_{adj} = 3.87$) performed similarly to those who read incongruent problem-question pairings ($M_{adj} = 4.07$), solving on average about half of the problems experienced. There was no main effect of problem organization, $F(2, 346) = 1.17$, $p = .31$, but there was a Congruence x Organization interaction, $F(2, 346) = 7.55$, $p = .001$, $\eta_p^2 = .04$, which is shown in Figure 5.

Simple effects analysis was completed to evaluate the effect of organization within each level of congruence. For those who read congruent problem-question pairings, I expected that accuracy would decrease as organizational structure was decreased. Within the congruent pairings (left, Figure 5), accuracy was significantly different across the three problem organizations, but was not consistent with the predicted pattern, $F(2, 346) = 6.92$, $p = .001$, $\eta_p^2 = .04$. Accuracy was comparable between the presumably helpful reference class structure and the presumably less helpful superordinate organization, $p = .29$. Furthermore, removing all explicit organization actually increased accuracy compared to the reference class organization, $p = .01$, and compared to problems organized into the superordinate set, $p < .001$.

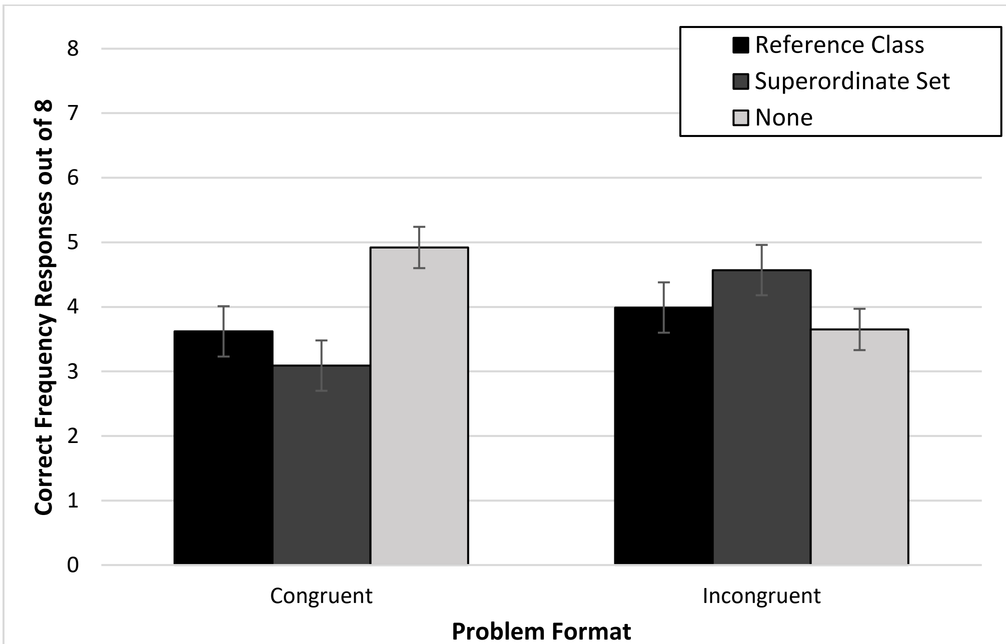


Figure 5. The interaction of congruence and problem organization on accuracy. Error bars indicate ±1 SE.

In retrospect, this lends additional support to the interference hypothesis, which suggests that reasoners were latching on to a value provided in the problem presentation that drew attention away from the correct reference class value. In this case, the superordinate value, which was provided several times in the reference class and superordinate organizations but not provided in the problems with no organization, may have been causing an unintended interference effect wherein participants were (erroneously) drawn to the superordinate value as the denominator for the PPV estimate.

For incongruent pairings, it was expected that accuracy would be low in all cases, as there was no problem structure that was designed to draw attention to the correct reference class. Within the incongruent pairings (right, Figure 5), accuracy did not differ when the problem was organized into the reference class structure or superordinate set, or when all explicit organization was removed, $F(2, 346) = 1.71$, $p = .18$. Accuracy did not change as the confusing problem structure in the incongruent pairings was altered to introduce another misleading reference point (i.e., the superordinate value) or when the overarching reference point was removed. This suggests that reasoners were not helped, nor necessarily hindered further, by changes in the explicit organization when starting from an (either explicit or implicit) organization that is incongruent with the question of interest.

Response patterns. To further assess the role of organizational structure, I evaluated denominator responses for each participant to determine who consistently conformed to identifiable response patterns (on at least four of the eight problems). Response patterns were coded using the same process from Experiments 1 and 2 to allow for comparison across studies. Reasoners who did not consistently provide the correct denominator are shown in Figure 6.

As found in Experiments 1 and 2, the vast majority of incorrect responses conformed to the total sample size (N) rather than any other value provided in the problem. Although this had been found in the previous studies, the finding was particularly surprising in the no explicit organization problems, as reasoners would have to calculate the total N value by adding together all four subset values. This suggests that participants had little trouble performing the addition calculation of a given value if they felt that it would provide the answer needed. Thus, the calculation itself is not necessarily the difficulty associated with completing (frequency) Bayesian reasoning tasks. Instead, it seems that reasoners are not sure which values are needed for solution. In this case, they had a guiding statement at the end that directed attention back toward the overall sample size (in very generic terms), which may have been used as a cue that this value was needed for solution. This is the only finding that goes against the value selection bias, and instead suggests that reasoners look for organization cues to determine which values are needed for solution. Value selection may only become the predominant tendency when reasoners are confused about how to reach the solution.

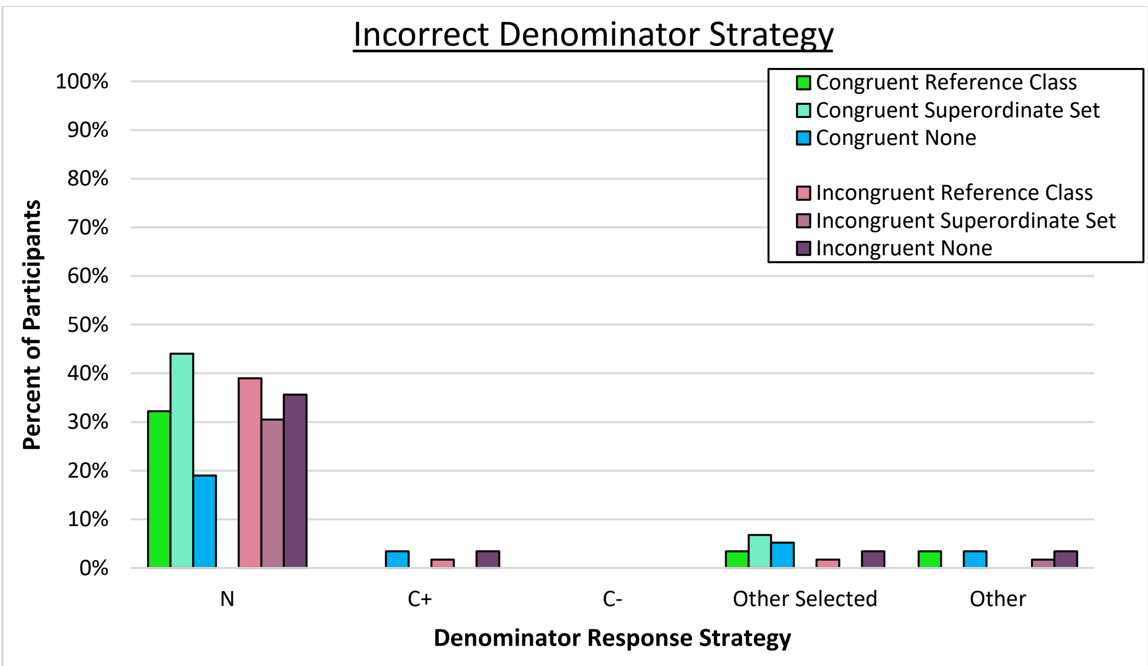


Figure 6. Proportion of participants who consistently used the correct denominator strategy (left) or the incorrect denominator strategy (right) on the frequency response format. C± denotes the total number of people who have or do not have the condition. Total N denotes the total in the superordinate set (i.e., sample size).

Discussion

In Experiment 3, I found clear evidence suggesting the importance of the structural organization of the problem. The congruence hypothesis was supported when all explicit organizational cues were eliminated from the problem presentation, but not on problems that organized subsets into the newly-created superordinate set condition or the reference class conditions (without totals) from Experiment 2. Unexpectedly, when the only reference point value available was the superordinate set, accuracy on incongruent pairings was typically higher than comparable congruent pairings.

When reasoners were not correctly calculating the denominator value, the majority of them utilized the superordinate value (N) much more so than any other value across all problem forms. Based on the response patterns, it appears that the superordinate value (N) was creating a salient reference point in the problem organizations, even when the value was not explicitly provided. Instead of demonstrating a value selection bias in this case, reasoners were calculating the total N from the four subsets. Numerical skill was again a strong predictor of accuracy.

General Discussion

The goal of these three experiments was to evaluate the representational and computational aspects of Bayesian reasoning tasks as they relate to reference dependence. The most consistent and surprising finding in all three experiments was that people were much more likely to utilize the superordinate value as part of their solution rather than the anticipated reference class values. This resulted in a weakened effect of congruence, with relatively low accuracy even in congruent conditions, as well as a different pattern of response errors than what was anticipated. There was consistent and strong evidence of a value selection bias in that incorrect responses almost always conformed to values that were provided in the problem rather than errors related to computation. The one notable exception occurred when no organizing information was available in the problem, other than the instruction to consider a sample of the same size as that in the problem. In that case, participants were most apt to sum all of the subsets of the sample to yield the size of the original sample (N). In all three experiments, higher numerical skills were generally associated with higher accuracy, whether calculations were required or not.

Reference Dependence

The results of these three experiments indicate that the initial presentation of the problem directly informed how reasoners responded to the *positive predictive value* (PPV) question. However, the responses I observed were unexpected, and highlighted the importance of reference dependence in ways that were not predicted.

I proposed that many of the representational and computational difficulties associated with Bayesian reasoning tasks were due to reference dependence, or the tendency to adopt a given or implied reference point at the start of cognitive deliberations. In the Bayesian reasoning problems used here, important reference points in the problem were specified through verbal and visual indicators, and I predicted that reasoners would utilize these cues to determine which values were key reference values needed for solution. Although some used these reference values for solution, there was another aspect of the problem formulations used in the current studies that may have unintentionally pulled attention to a different reference point, the superordinate value (N).

Previous studies have reported similar accuracy rates whether the estimates provided in the Bayesian reasoning problems referred to the initial sample (e.g., Sirota et al., 2014) or to a new representative sample of the same size (e.g., Gigerenzer & Hoffrage, 1995). A large-scale meta-analysis evaluated problem forms that explicitly stated a superordinate value in the problem description (but not necessarily that the problem values should be applied to a new sample of the same size) and found that having this information available did not improve performance (McDowell & Jacobs, 2017). Thus, I did not expect to observe an impact of describing the solution needed in terms of a new sample.

However, in hindsight, I realize that accuracy on those problems was consistently low, as all previous studies used the incongruent problem-question pairings. The lack of a discernible difference from introducing a new sample could be the result of a floor effect. Moreover, because the pattern of errors were not typically reported, I cannot differentiate whether accuracy was generally low because reasoners were using the superordinate value when it was salient, or because they were latching on to the incongruent reference class from the problem.

In the current studies, the superordinate value was emphasized in the problem and the question throughout all of the problem presentations except for the problem form in Experiment 3 that provided no explicit organizational structure. Even there, the reference to imagining a new sample of the same size likely served as a potent cue to find that value. Thus, the majority of incorrect denominator responses in both congruent and incongruent problem-question pairings conformed to the superordinate value rather than any other predicted value. Even in the simplest form of the congruent problem-question pairings tested in Experiment 1 (which was meant to be a replication of Talboy and Schneider, 2018a), where the correct denominator value could be directly selected from the problem, the focus on the superordinate value provided enough pull that accuracy was only around 50% rather than at the expected 80% mark.

By drawing attention to the superordinate value in the question being asked of participants, I apparently made this one of the, if not the most, salient reference point. The dependence on this reference point rivaled the strength of the correct reference point that was included in the congruent problems. In the incongruent problems, it was even stronger than the incorrect reference class within which the subsets were nested. In all experiments, it was almost the only other value suggested for the denominator other than the correct value. An additional study is currently underway to confirm that the instruction to consider a new sample of the same size can account for the superordinate value becoming a focal reference point, leading to these differences in accuracy and changes in the denominator response strategies between the current and past experiments. Regardless, the unexpected reliance on this reference point lends support to the larger hypothesis that reference dependence plays a large role in performance on Bayesian reasoning tasks.

Congruent Problem Structuring

In addition to the verbal and visual cues indicating potential reference points, the overall structure of the problem can either guide reasoners to the correct values needed for solution (in a congruent pairing) or cause confusion about which values should be used in the answer (in an incongruent pairing). The congruent format was expected to increase accuracy because the reference values that were highlighted in the problem structure through verbal and visual cues aligned with the question of interest (Talboy & Schneider, 2018a, 2018b). Alternatively, the incongruent format was expected to cause confusion because the reference values were not consistent with those needed to find the correct solution.

However, the effect of congruence was weak in these studies compared to previous research (e.g., Talboy & Schneider, 2018a, 2018b). Further, the congruence effect was reversed in Experiment 3 when the problems were organized into the superordinate set, resulting in higher accuracy on the incongruent pairing rather than the congruent pairing. The weak effect of congruence (and possibly the reversed effect) is likely the result of unintentionally focusing attention on the superordinate (N) value rather than the reference classes. If so, this also highlights the importance of the problem and question structuring, and in particular, signaling the implied reference point of interest for obtaining the solution.

When all explicit organization is removed and reasoners are left with the four subsets subtly organized in a congruent or incongruent manner, the congruent pairings resulted in higher accuracy than incongruent pairings. In this case, the superordinate value was no longer directly highlighted as a reference point at the beginning and end of the problem, though attention was still drawn to the overall sample size in very generic terms without a numeric value assigned to it. It appears that removing the numeric indicator of the superordinate value may have reduced interference and increased accuracy, but that even a non-numeric focus on this organizational cue may still have limited accurate identification of the correct reference class. This is consistent with the finding that the total sample size was by far the most common error in this condition even though participants had to add all four given values to produce this answer.

Even though the effect of congruence in the above condition was relatively small, the manipulation of congruence when explicit organization was removed was much subtler compared to previous studies. Rather than being a means to explain the alignment between verbal and visual indicators of the reference classes in the problem and those needed for solution, congruence in Experiment 3 referred to the ordering of the problem statements to focus on the test result or condition presence first. Structurally, the explicit reference points were eliminated as all four subsets were presented in a single block rather than broken out into visually and verbally defined reference classes. Therefore, it is not surprising to find this effect is much smaller than in studies in which reference points were structurally signaled.

Amount of information. Reasoners create mental representations of the problem structure based almost exclusively on the information that is provided (Johnson-Laird, 1994; Kintsch & Greeno, 1985; Sirota, Juanchich, et al., 2014). In accordance with the mental models approach, I hypothesized that giving reasoners full subset information would result in higher accuracy than when only partial information is provided. However, I also introduced an alternative discrimination hypothesis, that including additional subset information could result in lower accuracy because reasoners would have greater difficulty discriminating among a larger set of values.

Against expectations, though, inclusion of partial versus full information in the problem presentation did not have any discernible effect on accuracy, regardless of congruence between the problem and question of interest. In each case, performance on the congruent pairing was moderate, with performance on the incongruent pairings relatively low as has been found in any number of previous studies (e.g., Gigerenzer, Gaissmaier, Kurz-Milcke, Schwartz, & Woloshin, 2007; Gigerenzer & Hoffrage, 1995; Hoffrage, Krauss, Martignon, & Gigerenzer, 2015; Johnson & Tubau, 2015; Reyna & Brainerd, 2008; Sirota, Kostovičová, & Vallée-Tourangeau, 2015).

The lack of a difference in accuracy with full information suggests that including the two non-essential subsets did not help reasoners figure out how to find the solution. It also did not hurt their ability. However, the unexpected focus on the superordinate value (N), which then became a salient reference point, may have overshadowed any possible effect of being given full subset information versus partial information. A future study could reassess whether the inclusion of full subset information alters accuracy without the distraction of a second salient reference point.

Removing interference from misleading reference points. Unlike the full versus partial information manipulation, changing or removing focal reference values from the problem had a differential effect on accuracy for reasoners who read incongruent or congruent pairings.

Within the incongruent pairings, focal reference points that organized information in a way that did not draw attention to the correct solution were expected to cause an interference effect that reasoners would need to overcome in order to accurately solve the problem. Evidence for this was seen in the congruence effect observed in Experiment 1. In Experiment 2, when the competing C+ reference class was the nesting structure but the values were removed, accuracy increased, though only by a small amount. So, there seemed to be a slight release from interference, but the error analysis suggested that this was countered by the tendency for reasoners to be drawn to the N value. Further, in Experiment 3 when the explicit structure was eliminated, a large portion of reasoners were consistently calculating the total sample (N) instead of the correct T+ reference value. This suggests that the incongruent problem presentation did not clearly elucidate which reference value was needed for the PPV question, even after competing values were eliminated from the problem description.

Additional evidence for the interference hypothesis was unexpectedly found with structural changes to the congruent problem-question pairings. When all organizational cues and totals were eliminated from the problem presentations in Experiment 3, those who read sentences with a congruent ordering performed substantively better than when subsets were nested within test reference classes or the superordinate set (N). The unstructured condition was the only problem in which the total sample N value was not provided anywhere in the problem. This finding suggests that the presence of the N value must have been interfering with the ability to take advantage of the correct nested structure. This goes back to the general presumption that the reference class values would be the most salient reference point in the problem presentation, and that congruence would help reasoners organize the subset information into a structure that leads directly to the solution. Instead, by focusing on the superordinate (N) value in the question, I likely introduced an interference effect in the congruent pairings that was not anticipated. This suggests the importance of ruling out additional salient reference points within the problem as well as the question being asked.

Value Selection Bias

In addition to the primary findings regarding reference dependence and congruence, I also investigated the general bias toward selecting values from the problem rather than completing calculations when reasoners were unsure of how to determine the solution. In each experiment, virtually all reasoners who did not determine the correct response utilized values directly from the problem to fill in each component of their solution. Hardly any reasoners made responses that were consistent with calculation errors. This provided strong support for the value selection bias (while also confirming the strong draw to the superordinate set value as the denominator).

When reasoners are not familiar with the type of problem that has to be solved, they tend to rely on surface features to guide their solution (Chi, Feltovich, et al., 1981; Chi, Glaser, et al., 1981; Owen & Sweller, 1989; Swanson & Beebe-Frankenberger, 2004; Winner et al., 1980). This tendency may reflect a lack of willingness to engage the mental resources needed to fully flesh out the problem space (e.g., "the lazy controller," Kahneman, 2011), which results in responses consistent with the identifiable problem values rather than calculation mistakes. This bias may also reflect a general belief that relevant values should be readily available in the problem description, and so calculations should not be needed (Talboy & Schneider, 2018a). I consistently observed this value selection bias as the default response strategy when reasoners did not determine the correct response in all three experiments. This highlights the need to identify and develop methods for overcoming the tendency to select values in order to better assess what steps are actually needed to get to the correct solution.

However, in Experiment 3 when no explicit organization was provided, the majority of reasoners provided responses that were consistent with the superordinate value even though this number was not explicitly provided in the problem description. In this case, it appears that reasoners were (successfully) calculating the total sample size from the four subsets provided. This suggests that reasoners were using cues from the problem presentation, such as the final statement about considering a new sample of the same size, to determine which values they thought were needed in the response.

Calculations

Much of the existing literature suggests that computational difficulties make Bayesian reasoning tasks inherently difficult, particularly for those with low numerical skill compared to those with higher numeracy (Chapman & Liu, 2009; Reyna & Brainerd, 2008; Schwartz, Woloshin, Black, & Welch, 1997; Talboy & Schneider, 2018b). Further, even though some argue that this basic addition operation is simple (Johnson-Laird, Legrenzi, Girotto, Legrenzi, & Caverni, 1999; Sloman et al., 2003), there is substantial evidence that many reasoners are unable to complete this simple step to correctly solve the problem (Johnson & Tubau, 2015; Mayer, 2003; Reyna & Brainerd, 2008; cf. Schneider & Talboy, in progress). However, the difficulty is not necessarily with the calculation itself, but with determining the relationship between different subsets.

In problem solving, computation involves determining what is needed for solution, selecting the relevant values from the problem description, and then applying the correct values toward the solution (Schneider & Talboy, under review; Talboy & Schneider, 2018a). In this process, reasoners are expected to have an existing body of knowledge that helps them recognize which mathematical operation is appropriate for the question of interest. This also requires analytic abilities to correctly interpret the meaning of values provided in the problem. Although I did not attempt to evaluate each step required in computation, I did want to isolate the issues directly related to calculating the correct denominator (i.e., adding two values together).

Separating out the issue of adding the two relevant subsets together from the larger problem structuring issue was accomplished in Experiment 2. Responses of reasoners who read congruent problem-question pairings in which the denominator could be selected directly from the problem were compared to responses of those who read congruent problems in which the denominator had to be calculated by adding two subsets together. To our surprise, there was no evidence to suggest that the added calculation step increased the difficulty of solving the congruent problem-question pairing. Further, in Experiment 3, when reasoners were not provided any organizing value, a large portion of reasoners who read congruent or incongruent pairings actually calculated the superordinate value more often than any other value. This suggests that reasoners can and do complete simple calculations when they are given statement cues suggesting these are important values (such as the statement at the end of the problem telling them to consider another sample of the same size).

The lack of a discernible difference in accuracy in the congruent pairings, regardless of whether calculations were required or not, suggests that a simple addition step is not what is inhibiting accuracy on these nested set problems. Instead, when reasoners are starting from a problem that flows directly to the solution, the added calculation step may be viewed as a nominal change that reasoners readily take on to correctly answer the question, but only when starting from a problem structure that leads directly to the correct (or at least an obvious) solution (see also, Schneider & Talboy, in progress). If the difficulty reasoners have is not the actual step of adding values together, this suggests that the breakdown occurs in one of the other prerequisites for computation (Schneider & Talboy, under review; Talboy & Schneider, 2018a) or in the general process of understanding how nested subsets function in relation to one another.

Conclusion

The purpose of the current research was to evaluate the effect of reference dependence within Bayesian reasoning, and the extent to which reference dependence accounts for many of the representational and computational difficulties associated with Bayesian reasoning tasks. What may appear to be simple changes in how reasoning problems are presented can actually fundamentally alter the way reasoners interpret and utilize problem information to determine solutions. When reasoners are not sure what they need to determine the correct response, they tend to utilize the values provided directly in the problem presentation. Which value they choose will critically depend on the reference points they identify based on the problem structure and the way the question is asked. When calculation is needed, the calculation step of summing values does not in itself appear to be a hindrance to accuracy when reasoners are starting from a problem presentation that maps directly to the question of interest. This again suggests the issue is knowing which values are relevant and how to organize those values cognitively in order to arrive at the correct solution.

References

Ayal, S., & Beyth-Marom, R. (2014). The effects of mental steps and compatibility on Bayesian reasoning. *Judgment and Decision Making*, *9*(3), 226–242.

Bar-Hillel, M. (1980). The base-rate fallacy in probability judgments. *Acta Psychologica*, *44*, 211–233.

Barbey, A. K., & Sloman, S. A. (2007). Base-rate respect: From ecological rationality to dual processes. *Behavioral and Brain Sciences*, *30*(03), 241-254; discussion 255-297. http://doi.org/10.1017/S0140525X07001653

Brase, G. L. (2014). The power of representation and interpretation: Doubling statistical reasoning performance with icons and frequentist interpretations of ambiguous numbers. *Journal of Cognitive Psychology*, *26*(1), 81–97. http://doi.org/10.1080/20445911.2013.861840

Brase, G. L., & Hill, W. T. (2017). Adding up to good Bayesian reasoning: Problem format manipulations and individual skill differences. *Journal of Experimental Psychology: General*, *146*(4), 577–591. http://doi.org/10.1037/xge0000280

Bruine de Bruin, W., McNair, S. J., Taylor, A. L., Summers, B., & Strough, J. (2014). “Thinking about Numbers Is Not My Idea of Fun”: Need for Cognition Mediates Age Differences in Numeracy Performance. *Medical Decision Making*, *35*, 17–22. http://doi.org/10.1177/0272989X14542485

Brust-Renck, P. G., Royer, C. E., & Reyna, V. F. (2013). Communicating numerical risk: Human factors that aid understanding in health care. *Review of Human Factors and Ergonomics*, *8*(1), 235–276. http://doi.org/10.1177/1557234X13492980

Casscells, W., Schoenberger, A., & Graboys, T. B. (1978). Interpretation by physicians of clinical laboratory results. *The New England Journal of Medicine*, *299*(18), 999–1001.

Chapman, G. B., & Liu, J. (2009). Numeracy, frequency, and Bayesian reasoning. *Judgment and Decision Making*, *4*(1), 34–40.

Chi, M. T. H., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science*, *5*, 121–152. http://doi.org/10.1207/s15516709cog0502_2

Chi, M. T. H., Glaser, R., & Rees, E. (1981). Expertise in problem solving. *Technical Report No. 5*.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155–159. http://doi.org/10.1037/0033-2909.112.1.155

Cosmides, L., & Tooby, J. (1996). Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgment under uncertainty. *Cognition*, *58*(1), 1–73.

Dinner, I., Johnson, E. J., Goldstein, D. G., & Liu, K. (2011). Partitioning default effects: Why people choose not to choose. *Journal of Experimental Psychology: Applied*, *17*(4), 332–341. http://doi.org/10.1037/a0024354

Eddy, D. M. (1982). Probabilistic reasoning in clinical medicine: Problems and opportunities. In A. Kahneman, D., Slovic, P., & Tverskey (Ed.), *Judgment Under Uncertainty: Heuristics and Biases.* (pp. 249–267). Cambridge University Press.

Fischhoff, B., & Bar-Hillel, M. (1984). Focusing techniques: A shortcut to improving probability judgments? *Organizational Behavior and Human Performance*, *34*(2), 175–194.

http://doi.org/10.1016/0030-5073(84)90002-3

Fischhoff, B., Slovic, P., & Lichtenstein, S. (1978). Fault trees: Sensitivity of estimated failure probabilities to problem representation. *Journal of Experimental Psychology: Human Perception and Performance*, *4*(2), 330–344. http://doi.org/10.1037/0096-1523.4.2.330

Fox, C. R., & Levav, J. (2004). Partition-edit-count: Naive extensional reasoning in judgment of conditional probability. *Journal of Experimental Psychology: General*, *133*(4), 626–642. http://doi.org/10.1037/0096-3445.133.4.626

Gaissmaier, W., Wegwarth, O., Skopec, D., Muller, A., Broschinski, S., & Politi, M. C. (2012). Numbers Can Be Worth a Thousand Pictures: Individual Differences in Understanding Graphical and Numerical Representations of Health-Related Information. *Health Psychology*, *31*(3), 286–296. http://doi.org/10.1037/a0024850.supp

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*(3), 368–371. http://doi.org/10.1177/0272989X08329463

Garcia-Retamero, R., Cokely, E. T., & Hoffrage, U. (2015). Visual aids improve diagnostic inferences and metacognitive judgment calibration. *Frontiers in Psychology*, *6*(July), 932. http://doi.org/10.3389/fpsyg.2015.00932

Garcia-Retamero, R., & Galesic, M. (2010). Who profits from visual aids: Overcoming challenges in people's understanding of risks. *Social Science & Medicine*, *70*(7), 1019–1025. http://doi.org/10.1016/j.socscimed.2009.11.031

Garcia-Retamero, R., & Hoffrage, U. (2013). Visual representation of statistical information improves diagnostic inferences in doctors and their patients. *Social Science & Medicine*, *83*, 27–33. http://doi.org/10.1016/j.socscimed.2013.01.034

Gigerenzer, G., Gaissmaier, W., Kurz-Milcke, E., Schwartz, L. M., & Woloshin, S. (2007). Helping doctors and patients make sense of health statistics. *Psychological Science in the Public Interest*, *8*(2), 53–96.

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review*, *102*(4), 684–704.

Gigerenzer, G., Hoffrage, U., & Kleinbölting, H. (1991). Probabilistic mental models: a Brunswikian theory of confidence. *Psychological Review*, *98*(4), 506–528. http://doi.org/10.1037/0033-295X.98.4.506

Girotto, V., & Gonzalez, M. (2001). Solving probabilistic and statistical problems: A matter of information structure and question form. *Cognition*, *78*(3), 247–276. http://doi.org/10.1016/S0010-0277(00)00133-5

Girotto, V., & Pighin, S. (2015). Basic understanding of posterior probability. *Frontiers in Psychology*, *6*(May), 1–3. http://doi.org/10.3389/fpsyg.2015.00680

Greifeneder, R., Scheibehenne, B., & Kleber, N. (2010). Less may be more when choosing is difficult: Choice complexity and too much choice. *Acta Psychologica*, *133*(1), 45–50. http://doi.org/10.1016/j.actpsy.2009.08.005

Hájek, A. (2007). The reference class problem is your problem too. *Synthese*, *156*(3), 563–585. http://doi.org/10.1007/s11229-006-9138-5

Hamm, R. M., Miller, M. A., & Drillings, M. (1988). Interpretation of conditional probabilities in probabilistic inference word problems. *Unpublished Manuscript*.

Hammerton, M. (1973). A case of radical probability estimation. *Journal of Experimental Psychology*, *101*(2), 252–254. http://doi.org/10.1037/h0035224

Hill, W. T., & Brase, G. L. (2012). When and for whom do frequencies facilitate performance?

On the role of numerical literacy. *The Quarterly Journal of Experimental Psychology*, *0218*(January 2015), 1–26. http://doi.org/10.1080/17470218.2012.687004

Hoffrage, U., & Gigerenzer, G. (1998). Using natural frequencies to improve diagnostic inferences. *Academic Medicine*, *73*(5), 538–540.

Hoffrage, U., Hafenbrädl, S., & Bouquet, C. (2015). Natural frequencies facilitate diagnostic inferences of managers. *Frontiers in Psychology*, *6*(June), 1–11. http://doi.org/10.3389/fpsyg.2015.00642

Hoffrage, U., Krauss, S., Martignon, L., & Gigerenzer, G. (2015). Natural frequencies improve Bayesian reasoning in simple and complex inference tasks. *Frontiers in Psychology*, *6*(OCT). http://doi.org/10.3389/fpsyg.2015.01473

Iyengar, S. S., & Lepper, M. R. (2000). When choice is demotivating: Can one desire too much of a good thing? *Journal of Personality and Social Psychology*, *79*(6), 995–1006. http://doi.org/10.1037/0022-3514.79.6.995

Johnson-Laird, P. N. (1994). Mental models and probabilistic thinking. *Cognition*, *50*, 189–209. http://doi.org/10.1016/0010-0277(94)90028-0

Johnson-Laird, P. N. (2010). Mental models and human reasoning. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(43), 18243–50. http://doi.org/10.1073/pnas.1012933107

Johnson-Laird, P. N., Legrenzi, P., Girotto, V., Legrenzi, M. S., & Caverni, J. P. (1999). Naive probability: A mental model theory of extensional reasoning. *Psychological Review*, *106*(1), 62–88. http://doi.org/10.1037/0033-295X.106.1.62

Johnson, E. D., & Tubau, E. (2013). Words, numbers, & numeracy: Diminishing individual differences in Bayesian reasoning. *Learning and Individual Differences*, *28*, 34–40.

http://doi.org/10.1016/j.lindif.2013.09.004

Johnson, E. D., & Tubau, E. (2015). Comprehension and computation in Bayesian problem solving. *Frontiers in Psychology*, *6*(July), 938. http://doi.org/10.3389/fpsyg.2015.00938

Johnson, E. J., Shu, S. B., Dellaert, B. G. C., Fox, C., Goldstein, D. G., Häubl, G., … Weber, E. U. (2012). Beyond nudges: Tools of a choice architecture. *Marketing Letters*, *23*(2), 487–504. http://doi.org/10.1007/s11002-012-9186-1

Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.

Kintsch, W., & Greeno, J. G. (1985). Understanding and solving word arithmetic problems. *Psychological Review*, *92*(1), 109–129. http://doi.org/10.1037/0033-295X.92.1.109

Kotovsky, K., & Simon, H. A. (1990). What makes some problems really hard: Explorations in the problem space of difficulty. *Cognitive Psychology*, *22*(2), 143–183. http://doi.org/10.1017/CBO9781107415324.004

Kurzenhäuser, S., & Hoffrage, U. (2002). Teaching Bayesian reasoning: an evaluation of a classroom tutorial for medical students. *Medical Teacher*, *24*(5), 516–521. http://doi.org/10.1080/0142159021000012540

Legrenzi, P., & Girotto, V. (1995). Mental models in reasoning and decision making processes. In A. Garnham & J. Oakhill (Eds.), *Mental models in cognitive science: Essays in honour of Phil Johnson-Laird* (pp. 95–118). Erlbaum UK: Psychology Press.

Lipkus, I. M., Samsa, G., & Rimer, B. K. (2001). General performance on a numeracy scale among highly educated samples. *Medical Decision Making*, *21*(1), 37–44. http://doi.org/10.1177/0272989x0102100105

Lopes, L., & Oden, G. (1999). The role of aspiration level in risky choice: A comparison of cumulative prospect theory and SP/A theory. *Journal of Mathematical Psychology*, *43*(2), 286–313. http://doi.org/10.1006/jmps.1999.1259

Macchi, L. (2000). Partitive formulation of information in probabilistic problems: Beyond heuristics and frequency format explanations. *Organizational Behavior and Human Decision Processes*, *82*(2), 217–236. http://doi.org/10.1006/obhd.2000.2895

Markovits, H., & Barrouillet, P. (2002). The development of conditional reasoning: A mental model account. *Developmental Review*, *22*(2002), 5–36. http://doi.org/10.1006

Matsumoto, M., & Nishimura, T. (1998). Mersenne twister: A 623-dimensionally equidistributed uniform pseudo-random number generator. *ACM Transactions on Modeling and Computer Simulation*, *8*(1), 3–30. http://doi.org/10.1145/272991.272995

Mayer, R. E. (2003). Mathematical problem solving. In M. Royer (Ed.), *Mathematical Cognition: A Volume in Current Perspectives on Cognition, Learning, and Instruction* (pp. 69–92). Greenwich, CT: Information Age Publishing.

McDowell, M., & Jacobs, P. (2017). Meta-Analysis of the effect of natural frequencies on Bayesian Reasoning. *Psychological Bulletin*, 1–68. http://doi.org/10.1037/bul0000126

McDowell, M., Rebitschek, F. G., Gigerenzer, G., & Wegwarth, O. (2016). A Simple Tool for Communicating the Benefits and Harms of Health Interventions: A Guide for Creating a Fact Box. *MDM Policy & Practice*, *1*(1). http://doi.org/10.1177/2381468316665365

Micallef, L., Dragicevic, P., & Fekete, J. (2012). Assessing the effect of visualizations on Bayesian reasoning through crowdsourcing to cite this version : Assessing the effect of visualizations on Bayesian reasoning through crowdsourcing. *IEEE Transactions on Visualization and Com- Puter Graphics, Institute of Electrical and Electronics Engineers*.

Navarrete, G., & Mandel, D. R. (Eds.). (2016). *Improving Bayesian Reasoning: What Works and Why?* Frontiers in Psychology. http://doi.org/10.3389/978-2-88919-745-3

Nelson, W., Reyna, V. F., Fagerlin, A., Lipkus, I. M., & Peters, E. (2008). Clinical implications of numeracy: Theory and practice. *Annals of Behavioral Medicine*, *35*(3), 261–274. http://doi.org/10.1007/s12160-008-9037-8

Okamoto, M., Kyutoku, Y., Sawada, M., Clowney, L., Watanabe, E., Dan, I., & Kawamoto, K. (2012). Health numeracy in Japan: measures of basic numeracy account for framing bias in a highly numerate population. *BMC Medical Informatics and Decision Making*, *12*(1), 104. http://doi.org/10.1186/1472-6947-12-104

Owen, E., & Sweller, J. (1989). Should problem solving be used as a learning device in mathematics? *Journal for Reearch in Mathematics Education*, *20*(3), 322–328.

Peters, E., Hibbard, J., Slovic, P., & Dieckmann, N. (2007). Numeracy skill and the communication, comprehension, and use of risk-benefit information. *Health Affairs*, *26*(3), 741–748. http://doi.org/10.1377/hlthaff.26.3.741

Peters, E., Västfjäll, D., Slovic, P., Metz, C. K., Mazzocco, K., & Dickert, S. (2006). Numeracy and decision making. *Psychological Science*, *17*(5), 407–413.

Pighin, S., Gonzalez, M., Savadori, L., & Girotto, V. (2014). Improving Public Interpretation of Probabilistic Test Results: Distributive Evaluations. *Medical Decision Making*, *35*, 12–15. http://doi.org/10.1177/0272989X14536268

Pighin, S., Tentori, K., Savadori, L., & Girotto, V. (2018). Fostering the Understanding of Positive Test Results. *Annals of Behavioral Medicine*, (February). http://doi.org/10.1093/abm/kax065

Reyna, V. F. (2004). How people make decisions that involve risk: A dual-processes approach. *Current Directions in Psychological Science*, *13*(2), 60–66. http://doi.org/10.1111/j.0963-7214.2004.00275.x

Reyna, V. F., & Brainerd, C. J. (2008). Numeracy, ratio bias, and denominator neglect in judgments of risk and probability. *Learning and Individual Differences*, *18*(1), 89–107. http://doi.org/10.1016/j.lindif.2007.03.011

Schwartz, L. M., Woloshin, S., Black, W. C., & Welch, H. G. (1997). The role of numeracy in understanding the benefit of screening mammography.: at USF Libraries. *Annals of Internal Medicine*, *127*(11), 966–972.

Schwartz, L. M., Woloshin, S., & Welch, H. G. (2005). Can Patients Interpret Health Information? An Assessment of the Medical Data Interpretation Test. *Medical Decision Making*, *25*(3), 290–300. http://doi.org/10.1177/0272989X05276860

Sedlmeier, P., & Gigerenzer, G. (2001). Teaching Bayesian reasoning in less than two hours. *Journal of Experimental Psychology: General*, *130*(3), 380–400. http://doi.org/10.1037/0096-3445.130.3.380

Simon, H. A. (1973). The structure of ill structured problems. *Artificial Intelligence*, *4*(3–4), 181–201. http://doi.org/10.1016/0004-3702(73)90011-8

Simon, H. A. (1978). Rationality as Process and as Product of Thought. *American Economic Association*, *68*(2), 1–16.

Sirota, M., Juanchich, M., & Hagmayer, Y. (2014). Ecological rationality or nested sets? Individual differences in cognitive processing predict Bayesian reasoning. *Psychonomic Bulletin & Review*, *21*(1), 198–204. http://doi.org/10.3758/s13423-013-0464-6

Sirota, M., Kostovičová, L., & Juanchich, M. (2014). The effect of iconicity of visual displays on statistical reasoning: evidence in favor of the null hypothesis. *Psychonomic Bulletin & Review*, *21*(4), 961–8. http://doi.org/10.3758/s13423-013-0555-4

Sirota, M., Kostovičová, L., & Vallée-Tourangeau, F. (2015). Now you Bayes, now you don't:

Effects of set-problem and frequency-format mental representations on statistical reasoning. *Psychonomic Bulletin & Review*, *22*(5), 1465–1473. http://doi.org/10.3758/s13423-015-0810-y

Sloman, S. A., Over, D., Slovak, L., & Stibel, J. M. (2003). Frequency illusions and other fallacies. *Organizational Behavior and Human Decision Processes*, *91*(2), 296–309. http://doi.org/10.1016/S0749-5978(03)00021-9

Swanson, H. L., & Beebe-Frankenberger, M. (2004). The Relationship Between Working Memory and Mathematical Problem Solving in Children at Risk and Not at Risk for Serious Math Difficulties. *Journal of Educational Psychology*, *96*(3), 471–491. http://doi.org/10.1037/0022-0663.96.3.471

Talboy, A. N., & Schneider, S. L. (2017). Improving accuracy on Bayesian inference problems using a brief tutorial. *Journal of Behavioral Decision Making*, *30*(2), 373–388. http://doi.org/10.1002/bdm.1949

Talboy, A. N., & Schneider, S. L. (2018a). Focusing on what matters: Restructuring the presentation of Bayesian reasoning problems. *Journal of Experimental Psychology: Applied*, *24*(4), 440–458. http://doi.org/10.1037/xap0000187

Talboy, A. N., & Schneider, S. L. (2018b). Increasing understanding of diagnostic test outcomes. *Medical Decision Making*, *38*(5), 573–583. http://doi.org/10.1177/0272989X18758293

Tversky, A., & Kahneman, D. (1991). Loss aversion in riskless choice : A reference-dependent model. *The Quarterly Journal of Economics*, *106*(4), 1039–1061.

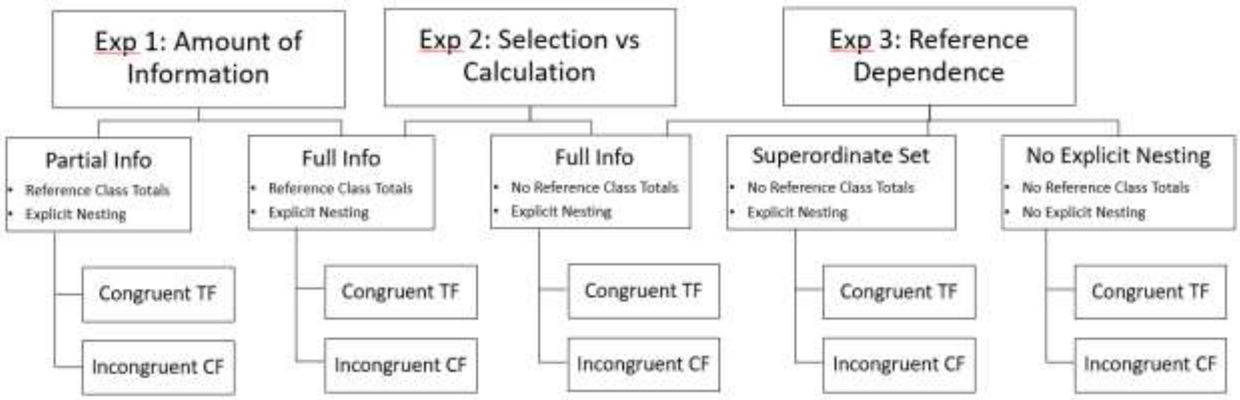
Weller, J. A., Dieckmann, N. F., Tusler, M., Metrtz, C. K., Burns, W. J., & Peters, E. (2013). Development and testing of an abbreviated numeracy scale: A Rasch analysis approach. *Journal of Behavioral Decision Making*, *26*(2), 198–212. http://doi.org/10.1002/bdm.1751

Winner, E., Engel, M., & Gardner, H. (1980). Misunderstanding metaphor: What's the problem? *Journal of Experimental Child Psychology*, *30*(1), 22–32. http://doi.org/10.1016/0022-0965(80)90072-7

Wolfe, C. R., Fisher, C. R., & Reyna, V. F. (2013). Semantine coherence and inconsistency in estimating conditional probabilities. *Journal of Behavioral Decision Making*, *26*(3), 237–246. http://doi.org/10.1002/bdm.1756

Appendix A: Experimental Conditions Map

Map of conditions across the three experiments.



Appendix B: Example Bayesian Reasoning Problems

Example Problem Forms for Experiment 1

Incongruent Condition-Focus Problem – Partial Detail

To determine whether a woman is at risk of breast cancer, doctors conduct mammogram screenings. Sometimes women test positive even when they should test negative or test negative when they should test positive. Here is some information for a random sample of 10,000 women who had a mammogram:

In this sample of 10,000 women, 100 have breast cancer.

Of the 100 women who have breast cancer:
80 received a positive result on their mammogram.

Of the 9,900 women who do not have breast cancer:
990 received a positive result on their mammogram.

Imagine another random sample of 10,000 women who had a mammogram.

Congruent Test-Focus Problem – Partial Detail

To determine whether a woman is at risk of breast cancer, doctors conduct mammogram screenings. Sometimes women test positive even when they should test negative or test negative when they should test positive. Here is some information for a random sample of 10,000 women who had a mammogram:

In this sample of 10,000 women, 1,070 received a positive result on their mammogram.

Of the 1,070 women who received a positive result on their mammogram:
80 have breast cancer.

Of the 8,930 women who received a negative result on their mammogram:
20 have breast cancer.

Imagine another random sample of 10,000 women who had a mammogram.

Incongruent Condition-Focus Problem – Full Detail

To determine whether a woman is at risk of breast cancer, doctors conduct mammogram screenings. Sometimes women test positive even when they should test negative or test negative when they should test positive. Here is some information for a random sample of 10,000 women who had a mammogram:

In this sample of 10,000 women, 100 have breast cancer.

Of the 100 women who have breast cancer:
80 received a positive result on their mammogram.
20 received a negative result on their mammogram.

Of the 9,900 women who do not have breast cancer:
990 received a positive result on their mammogram.
8910 received a negative result on their mammogram.

Imagine another random sample of 10,000 women who had a mammogram.

Congruent Test-Focus Problem – Full Detail

To determine whether a woman is at risk of breast cancer, doctors conduct mammogram screenings. Sometimes women test positive even when they should test negative or test negative when they should test positive. Here is some information for a random sample of 10,000 women who had a mammogram:

In this sample of 10,000 women, 1,070 received a positive result on their mammogram.

Of the 1,070 women who received a positive result on their mammogram:
80 have breast cancer.
990 do not have breast cancer.

Of the 8,930 women who received a negative result on their mammogram:
20 have breast cancer.
8,910 do not have breast cancer.

Imagine another random sample of 10,000 women who had a mammogram.

Example Problem Forms for Experiment 2

Incongruent Condition-Focus Problem without Reference Class Totals

To determine whether a woman is at risk of breast cancer, doctors conduct mammogram screenings. Sometimes women test positive even when they should test negative or test negative when they should test positive. Here is some information for a random sample of 10,000 women who had a mammogram:

In this sample of 10,000 women:

Of those who have breast cancer:
80 received a positive result on their mammogram.
20 received a negative result on their mammogram.

Of those who do not have breast cancer:
990 received a positive result on their mammogram.
8910 received a negative result on their mammogram.

Imagine another random sample of 10,000 women who had a mammogram.

Congruent Test-Focus Problem without Reference Class Totals

To determine whether a woman is at risk of breast cancer, doctors conduct mammogram screenings. Sometimes women test positive even when they should test negative or test negative when they should test positive. Here is some information for a random sample of 10,000 women who had a mammogram:

In this sample of 10,000 women:

Of those who received a positive result on their mammogram:
80 have breast cancer.
990 do not have breast cancer.

Of those who received a negative result on their mammogram:
20 have breast cancer.
8910 do not have breast cancer.

Imagine another random sample of 10,000 women who had a mammogram.

Example Problem Forms for Experiment 3

Incongruent Condition-Focus Problem with Superordinate Set Organization

To determine whether a woman is at risk of breast cancer, doctors conduct mammogram screenings. Sometimes women test positive even when they should test negative or test negative when they should test positive. Here is some information for a random sample of 10,000 women who had a mammogram:

In this sample of 10,000 women:

80 have breast cancer AND received a positive result on their mammogram.
20 have breast cancer AND received a negative result on their mammogram.
990 do not have breast cancer AND received a positive result on their mammogram.
8910 do not have breast cancer AND received a negative result on their mammogram.

Imagine another random sample of 10,000 women who had a mammogram.

Congruent Test-Focus Problem with Superordinate Set Organization

To determine whether a woman is at risk of breast cancer, doctors conduct mammogram screenings. Sometimes women test positive even when they should test negative or test negative when they should test positive. Here is some information for a random sample of 10,000 women who had a mammogram:

In this sample of 10,000 women:

80 received a positive result on their mammogram AND have breast cancer.
990 received a positive result on their mammogram AND do not have breast cancer.
20 received a negative result on their mammogram AND have breast cancer.
8910 received a negative result on their mammogram AND do not have breast cancer.

Imagine another random sample of 10,000 women who had a mammogram.

Incongruent Condition-Focus Problem with No Explicit Organization

To determine whether a woman is at risk of breast cancer, doctors conduct mammogram screenings. Sometimes women test positive even when they should test negative or test negative when they should test positive. Here is some information for a random sample of women who had a mammogram:

80 have breast cancer AND received a positive result on their mammogram.
20 have breast cancer AND received a negative result on their mammogram.
990 do not have breast cancer AND received a positive result on their mammogram.
8910 do not have breast cancer AND received a negative result on their mammogram.

Imagine another random sample of the same number of women who had a mammogram.

Congruent Test-Focus Problem with No Explicit Organization

To determine whether a woman is at risk of breast cancer, doctors conduct mammogram screenings. Sometimes women test positive even when they should test negative or test negative when they should test positive. Here is some information for a random sample of women who had a mammogram:

80 received a positive result on their mammogram AND have breast cancer.
990 received a positive result on their mammogram AND do not have breast cancer.
20 received a negative result on their mammogram AND have breast cancer.
8910 received a negative result on their mammogram AND do not have breast cancer.

Imagine another random sample of the same number of women who had a mammogram.

Appendix C: Institutional Review Board Approval and Informed Consent

USF
UNIVERSITY OF
SOUTH FLORIDA

RESEARCH INTEGRITY AND COMPLIANCE
Institutional Review Boards, FWA No. 00001669
12901 Bruce B. Downs Blvd., MDC035 • Tampa, FL 33612-4799
(813) 974-5638 • FAX (813) 974-7091

2/2/2015

Alaina Talboy, B.A.
Psychology
4202 E. Fowler Avenue, PCD4118G
Tampa, FL 33620

RE:	**Exempt Certification**
IRB#:	Pro00020889
Title:	Evaluating Risky Choices S2015

Dear Mrs. Talboy:

On 2/2/2015, the Institutional Review Board (IRB) determined that your research meets criteria for exemption from the federal regulations as outlined by 45CFR46.101(b):

(2) Research involving the use of educational tests (cognitive, diagnostic, aptitude, achievement), survey procedures, interview procedures or observation of public behavior, unless:
(i) information obtained is recorded in such a manner that human subjects can be identified, directly or through identifiers linked to the subjects; and (ii) any disclosure of the human subjects' responses outside the research could reasonably place the subjects at risk of criminal or civil liability or be damaging to the subjects' financial standing, employability, or reputation.

Approved Item(s):
Evaluating Risky Choices S2015 Study Protocol 012815.docx
ANT SYP Verbal Consent Script 012815.doc

As the principal investigator for this study, it is your responsibility to ensure that this research is conducted as outlined in your application and consistent with the ethical principles outlined in the Belmont Report and with USF IRB policies and procedures.

Please note, as per USF IRB Policy 303, "Once the Exempt determination is made, the application is closed in eIRB. Any proposed or anticipated changes to the study design that was previously declared exempt from IRB review must be submitted to the IRB as a new study prior to initiation of the change."

If alterations are made to the study design that change the review category from Exempt (i.e., adding a focus group, access to identifying information, adding a vulnerable population, or an intervention), these changes require a new application. However, administrative changes, including changes in research personnel, do not warrant an amendment or new application.

Given the determination of exemption, this application is being closed in ARC. This does not limit your ability to conduct your research project. Again, your research may continue as planned; only a change in the study design that would affect the exempt determination requires a new submission to the IRB.

We appreciate your dedication to the ethical conduct of human subject research at the University of South Florida and your continued commitment to human research protections. If you have any questions regarding this matter, please call 813-974-5638.

Sincerely,

John A. Schinka, Ph.D.

John Schinka, Ph.D., Chairperson
USF Institutional Review Board

INFORMED CONSENT

First, I'm going to explain to you what your participation in this study entails and then ask if you want to participate. Keep in mind that you do not have to participate if you don't want to. The name of the PI is Alaina Talboy. The USF IRB number for this research is Pro20889. This study will take approximately 60 minutes, and completion of the study will earn you 2 points. If you decide to withdraw or you are excused by the experimenter, points will be awarded based upon the time you spent in the study. You also have the option of completing a different study listed on SONA to earn 2 points.

Today you will be participating in a study at the University of South Florida that is concerned with how well people understand information about risks. In this study, you will be randomly assigned to one of several conditions. In each condition, you are going to be presented a series of general questions about aspects of your numerical and graphical abilities. You will then be asked a series of questions that concern potential real-life situations involving risk. We want to know how you evaluate the risks in those situations. You may also be asked to provide general demographic information.

Your name will not be associated with any of your study responses. In fact, we will not be asking for written consent, but we will need verbal consent from you. In this way, your privacy and research records will be kept confidential to the extent of the law. Authorized research personnel, employees of the Department of Health and Human Services, the USF Institutional Review Board and its staff, and any other individuals acting on behalf of USF, may inspect the records from this research project. The records will be anonymous.

It is possible, although unlikely, that unauthorized individuals could gain access to your responses. Confidentiality will be maintained to the degree permitted by the technology used. No guarantees can be made regarding the interception of data sent via the Internet. However, your participation in this online survey involves risks similar to a person's everyday use of the Internet. If you complete and submit an anonymous survey and later request your data be withdrawn, this may or may not be possible as the researcher will be unable to extract anonymous data from the database.

The results of this study may be published. Your data will be combined with data from others in the publication. The published results will not include your name or any other information that would personally identify you in any way.

This project presents no risk or harm to you, and there are no anticipated benefits to you. If you have any questions or concerns regarding the research, a written copy of this verbal agreement along with contact information for the principal investigator and the IRB office can be provided to you at your request.

Your participation in this experiment is entirely voluntary, and you may leave at any time should you feel uncomfortable with the procedures. Your decision to participate or not to participate will not affect your status or course grade.

Do you want to participate in the study today?

Yes	No
○	○