Schematic Effects on Probability Problem Solving

S. Sonia Gugga

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2012

ABSTRACT

Schematic Effects on Probability Problem Solving

S. Sonia Gugga

Three studies examined context effects on solving probability problems. Variants of word problems were written with cover stories which differed with respect to social or temporal schemas, while maintaining formal problem structure and solution procedure. In the first of these studies it was shown that problems depicting schemas in which randomness was inappropriate or unexpected for the social situation were solved less often than problems depicting schemas in which randomness was appropriate. Another set of two studies examined temporal and causal schemas, in which the convention is that events are considered in forward direction. Pairs of conditional probability (CP) problems were written depicting events E_1 and E_2, such that E_1 either occurs before E_2 or causes E_2. Problems were defined with respect to the order of events expressed in CPs, so that $P(E_2|\ E_1)$ represents the CP in schema-consistent, intact order by considering the occurrence of E_1 before E_2, while $P(E_1|E_2)$ represents CP in schema-inconsistent, inverted order. Introductory statistics students had greater difficulty encoding CP for events in schema-inconsistent order than CP of events in conventional deterministic order.

The differential effects of schematic context on solving probability problems identify specific conditions and sources of bias in human reasoning under uncertainty. In addition, these biases may be influential when evaluating empirical findings in a manner similar to that demonstrated in this paper experimentally, and may have implications for how social scientists are trained in research methodology.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I: INTRODUCTION

With respect to educational and social policy, renewed emphasis on evidence-based practice and the assessment of student learning has highlighted the importance of nuanced understanding of statistical inference for social scientists, educators, and policymakers (Gal, 2005; Milton, 2006; Slavin, 2004, 2008). There has been debate as to the quality of statistical training for social scientists, in that it has been criticized for its overreliance on narrow methodologies as well as inadequately preparing both producers and consumers of research literature (Henson, Hull, & Williams, 2010; Olani, Harskamp, Hoekstra, & van der Werf, 2010; Pallas, 2001; Slavin, 2004). With respect to consumers of scientific literature, a basic understanding of the process of statistical inference is central to evaluating the validity and scope of social science research (Neath, 2010; Paris & Luo, 2010; Slavin, 2003, 2004, 2008; Sloane, 2008) which may significantly affect how empirical evidence is implicated in major policy decisions, such as the use of students' standardized test scores to evaluate teacher effectiveness.

A major body of work into statistical reasoning has illustrated that individuals often disregard statistically prescribed processes, employing heuristics and biases which often produce invalid inferences (see Barbey & Sloman, 2007; Cohen, 1981; Gigerenzer & Hoffrage, 2007; Kahneman & Tversky, 1979; Krynski & Tenenbaum, 2007; Nisbett, Krantz, Jepson, & Kunda, 1993; Stanovich, Toplak, & West, 2008; and Tversky & Kahneman, 1974, 1980 for extensive treatments of the debate). It would be difficult to be optimistic about the efficacy of policy enacted from faulty inferences, and while many of the aforementioned studies demonstrate the persistence of biases in both the statistically naïve and individuals with extensive statistical training, another body of work exhibits particular contexts which may

improve performance on judgments under uncertainty. Improvement has been affected by training, expression of likelihood in frequency formats, or by expressing real-world application to contexts in which individuals are more likely to reason probabilistically with respect to causal or social schemas (Fox & Levav, 2004; Gigerenzer & Hoffrage, 1995, 2007; Girotto & Gonzalez, 2007; Krynski & Tenenbaum, 2007).

In cognition research, the concept of *schema* is often invoked to describe a hierarchically structured unit of knowledge. Through experience, individuals form schemas consisting of elements and the relationships among them. Schemas serve to reduce a wealth of phenomena into meaningful, manageable units, producing categories that ultimately aid in long-term memory (Bower, Black, & Turner, 1979; Marshall, 1993; Nisbett, et al., 1993; Schank & Abelson, 1977). The novice-expert distinction in problem-solving research is often delineated by experts' recognition of appropriate schemas depending on deep formal problem structure. For novice problem-solvers, it has been shown that the surface features of a problem may provide cues to an inappropriate formal solution schema (Anderson & Thompson, 1989; Gick & Holyoak, 1980; Reusser, 1988; Ross, 1984, 1989).

A mathematical word problem's cover story is its real-world context and semantic content, considered *surface features*, in that they are usually not intended to affect a problem's technical difficulty or formal solution processes. When a cover story relates to social interactions, however, pragmatic reasoning schemas may be invoked in the mind of the problem solver. Generally, "pragmatic reasoning schemas" embody the rules and relationships of real-world social situations, abstracted and classified (Cheng & Holyoak, 1985; Cheng, Holyoak, Nisbett, & Oliver, 1993; Cheng & Nisbett, 1993). The difficulty many problem-solvers face in correctly solving conditional reasoning *if-then* problems posed in the abstract has been shown to

be alleviated by placing the problem within a pragmatic, deontic context, such as *permission*, *obligation*, *cheating*, or *cost-benefit* paradigms (Cheng & Holyoak, 1985; Cosmides, 1989).

Problems in the earliest phases of a course in probability theory generally deal with randomizing devices – balls in urns, shuffled decks of cards, flipped coins, or rolled dice. Their stochasticity is apparent: many people are already familiar with them through games of chance. Social science students new to the field of statistical inference often have difficulty transferring probabilistic principles from games of chance to human behavior, in which the randomness of the problem situation is less apparent (Agnoli & Krantz, 1989; Fong, Krantz, & Nisbett, 1993; Howell & Burnett, 1978; Teigen & Keren, 2007) While *if-then* problems in a pragmatic social context are easier to solve than abstract versions, work in the field has shown that, in probability problem solving, replacing randomizing devices with anthropomorphic context tends to increase the difficulty of the problem rather than alleviate it (Bassok, Wu, & Olseth, 1995; Falk & Lann, 2008). It may be that individuals have more difficulty reasoning probabilistically about people in social contexts rather than about non-human entities due to the implicit causal (rather than stochastic) schemas that people's beliefs or action are caused by certain internal traits or intention (Schwartz & Goldman, 1996; Windschitl & Wells, 1998). With respect to training in the social sciences, however, it is necessary to understand that human behavior can have a stochastic component of variability.

The claim made here is that problem solving in probability is particularly affected by domain-specific biases in probabilistic reasoning. It is worth noting that this paper is not intended to enforce or dispute the *heuristics and biases* view that human judgment under uncertainty is broadly flawed and does not conform to what is considered rational in light of

formal statistical rules. Rather, the studies described in this paper will illustrate particular circumstances in which heuristics and biases may be differentially exhibited during probability problem solving, with respect to the schemas depicted in those problems' cover stories. The studies will be discussed in terms of frameworks constructed to describe probabilistic reasoning processes uniquely.

One of the frameworks describes discrepancies between formal quantitative probabilistic assessments (*Type 2*) and on-the-fly, qualitative probabilistic judgments (*Type 1*), and has led to a theory of dual-systems representations for probability judgments (Evans & Frankish, 2008; Fox & Levav, 2004; Sloman, 1996; Sloman & Rips, 1998; Smith & Collins, 2009; Stanovich, et al., 2008; Tversky & Kahneman, 1983; Windschitl & Krizan, 2005; Windschitl & Wells, 1998). While the two processes have been well-differentiated, it remains to be determined how they may interact in probability problem solving contexts. Individuals have been shown to have a strong bias toward Type 1 processes for evaluating social behavior or characteristics: people generally have difficulty understanding social behavior in strictly probabilistic terms. It may be that the interaction of the two processes depends on the details of the schematic representation of the elements of the problem. The effect of adjustments on schematic elements of items has not yet been studied in a probability problem solving context.

Within these theoretical frameworks, experimentally manipulating pragmatic schemas within isomorphic probability problems (while controlling for formal solution schema) would yield insights into how pragmatic schemas affect the difficulty of solving probability problems. The implications of this theoretical framework was explored using a mix of classroom experiments, conducted in introductory probability and statistics courses in a graduate school

of education, and laboratory experiments conducted online. It was thought that including participants from populations other than that of students in social sciences would inform the generalizability of results and allow for comparisons with previous findings. The research program had several phases, each aimed at understanding particular schematic representations.

The first of the real-world schemas that was addressed used probability problems involving real-world situations in which a "first-come, first-served" schema was replaced by random ordering. Subjects' performance on these problems was compared with performance on problems in which randomness is an inherent aspect of the social situation described in the cover story. With respect to real-world context, it was expected that performance would differ on problems in which randomness was imposed compared to problems in which randomness was endemic. Randomness would be expected in situations in which people are selected by lottery or draw items blindly from a container.

Another pair of studies addressed the direction of temporality and causation, examining whether inverting the temporal direction of a schema would affect the difficulty of a conditional probability problem. It was hypothesized that it would be easier for individuals to reason forwards regarding temporal or causal events given the deterministic nature of causal and, by extension, temporal schemas (Cheng & Nisbett, 1993; Tversky & Kahneman, 1980). In other words, it would be easier for subjects to calculate the conditional probability of an event given the probability of events preceding it versus calculating the probability of an event given the probability of events occurring later. Similarly, it was expected that problems asking for the conditional probability of an effect given the probability of its cause(s) would be easier than problems asking for the probability of a cause given the probability of its effect(s). It was supposed that inverting the direction of determination should introduce an additional level of

difficulty to the problem. It was also expected that the perceived causal strength between events would mediate the effect. In addition, I intend to illustrate how schematic effects on probability problem solving may inform instruction on statistical inference and the evaluation of scientific data.

The experiments discussed yield insights into how probability problems are categorized and solved, what effects these pragmatic schemas have on problem-solving success, and how they may inform peculiarities about probabilistic reasoning in general.

CHAPTER II: LITERATURE REVIEW

A fairly influential body of research in cognition has demonstrated that human judgment under uncertainly falls somewhere on the spectrum between *vulnerable to biases* and *not rational* (Agnoli & Krantz, 1989; Bar-Hillel & Falk, 1982; Cosmides & Tooby, 1996; Falk, 1989; Kahneman & Tversky, 1979; Mackie, 1981; Nisbett, et al., 1993; Rips, 1994; Stanovich, et al., 2008; Tversky & Kahneman, 1974, 1980). Many of these studies have utilized tasks which elicit qualitative probabilistic judgments or attributions, rather than the formal statistical processes which define the prescriptive standard of rationality. Conversely, another vein of research disputes the validity of the broad indictment of probabilistic reasoning by addressing the disconnect between expecting probabilistic judgments on experimental tasks that do not demand formal processes (Windschitl & Krizan, 2005), raising theoretical objections to the interpretation of results (Cohen, 1981), or illustrating conditions in which performance on statistical reasoning tasks may be improved through changes to context (Fong, et al., 1993; Fox & Levav, 2004; Gigerenzer & Hoffrage, 1995, 2007; Krynski & Tenenbaum, 2007; Schurr & Erev, 2007; Windschitl & Wells, 1998). A number of more recent studies have utilized quantitative problem-solving tasks, illustrating differential effects on computational problem solving (Fantino & Stolarz, 2007; Fox & Levav, 2004; Gigerenzer & Hoffrage, 1995; Krynski & Tenenbaum, 2007; Martin & Bassok, 2005; Villejoubert & Mandel, 2002; Wright & Murphy, 1984).

Computational models of mathematical problem solving define the methods by which people solve problems using methods other than direct linear application of formal, prescribed rules. Computational models often describe how problem solving is affected by schematic representations of a word problem's context (Sweller, 1988; Vosniadou, 1989; Windschitl &

Wells, 1998). Broadly speaking, in cognitive research, the wealth of phenomena in the world are organized into schemas, in which units of information are abstracted and defined along with the relationships among them (Holyoak & Thagard, 1989; Sweller, 1988). With respect to problem solving, a schema serves as a structuring framework, defining elements such as problem type with associated solution procedures (Cheng & Holyoak, 1985; Cheng, et al., 1993; Chi, Feltovich, & Glaser, 1981; Chi, Glaser, & Rees, 1982; Gick & Holyoak, 1980; Hinsley, Hayes, & Simon, 1977; Marshall, 1993; Newell & Simon, 1972; Nisbett, 1993; Pólya, 1941, 1954a, 2004; Rips, 1994; Ross, 1989; Sweller, 1988). Marshall (1993) further defined the elements of a problem schema as consisting of feature elements, constraints, planning, and execution. The utility of a schema is primarily to aid long-term memory, as the elements of a schema are stored such that retrieval of one element will make its associated elements more salient. Studies of the distinction between expert and novice approaches to problem solving have yielded insights into how schematic representation of problems change with increased experience. Notably, novices tend to focus on contextual cues with respect to categorizing problems, creating schemas based on cover story. Experts more readily categorize problems based on schemas defined by domain-specific rules or solution procedure. This distinction has been demonstrated with problems in the domain of physics (Chi, et al., 1981; Chi, et al., 1982; Larkin, Heller, & Greeno, 1980; Larkin, McDermott, Simon, & Simon, 1980; Larkin & Simon, 1987), as well as algebra (Gick & Holyoak, 1980; Hinsley, et al., 1977). Experts' categorization of problems by their deep structure facilitates problem solving by the schematic association of problem type and solution procedure, through what is characterized as *schema induction* (Gick & Holyoak, 1980). Schema induction describes the process by which identification of a problem's type based on structural elements cues a

correct solution procedure; identifying problem type based on spurious elements may cue an incorrect solution procedure and produce an incorrect answer.

Computational models of problem solving explain many of the processes that novice problem solvers engage when approaching a novel problem, including the effect of problem context (Martin & Bassok, 2005; Reusser, 1988; Ross, 1989). This has been studied extensively with respect to the four-card *if-then* conditional reasoning task proposed in abstract by Wason (1968), i.e., the *selection task*. In the selection task, individuals are asked to verify a logical implication rule expressed in "if p, then q" form. A number of investigators have shown that placing the problem within a pragmatic context, as summarized in Table 1, significantly increases the proportion of problem solvers who correctly solve the problem. Investigators, however, have differing explanations for these effects, since the selection task is usually expressed in terms of deontic systems (i.e., those defining norms of necessity and obligation)

Table 1
Expression of the Selection Task Rule in Abstract and Deontic Contexts

Context	Rule
Abstract	If there is a square on one side of the card, then there is a red scribble on the other side (Wason, 1968).
Permission	If a customer is drinking an alcoholic beverage, then he or she must be over twenty-one (Cheng, Holyoak, Nisbett, & Oliver, 1993).
Obligation	If one works for the armed forces, then one must vote in the elections (Cheng, et al., 1993).
Private Exchange	If you give me your ostrich eggshell, then I'll give you duiker meat (Cosmides, 1989).
Cost-benefit	If you spend over £100, then you will receive a free gift (Rips, 1994).

(Cosmides, 1989; Rips, 1994). Cheng and Holyoak (1985) cite the effect of pragmatic schemas, noting that there are context-dependent schematic representations of *permission* and *obligation* paradigms which are engaged in those contexts. Cosmides, rather than credit the concept of pragmatic schema, attributes the effect to the evolutionary salience of obeying or violating social contracts for individuals' enhanced performance on these problems.

A number of the pragmatic schemas applied to the selection task have been studied within the context of event schemas, or scripts, by which individual events are categorized into generalized structures. Shank and Abelson (1977) illustrated that within a restaurant script there are several expected events, including being seated at a table, ordering from a menu, being served food, eating, and paying the bill. Scripts are shared natural categories similar to those for objects; 73% of individuals sampled independently cited each of the events previously mentioned when asked to list the events that occur during a restaurant visit. These categories also play a role in what is recalled about an experience. Individuals are more likely to recall script elements rather than details which are not classified by script elements (Bower, Black, & Turner, 1979). Calling upon an element of a script in word problems facilitates recall of related elements of that script, similar to the process of schema induction described by Gick and Holyoak for problem type and problem solution (1980). If the context of a word problem recalls an established script, the salience of that script may in part affect performance on the problem.

In probability problem solving, however, pragmatic context has not been shown to alleviate item difficulty. Randomness is often a necessary factor in probability problems, and individuals readily understand randomness with respect to games of chance. Problems involving cards, dice, or balls in urns do not require special consideration with respect to stochasticity (Fong, et al., 1993; Howell & Burnett, 1978; Teigen & Keren, 2007). When these

same problems are expressed in a human or social context, however, problem solving is often impeded. For example, with respect to the much studied three-card (i.e., "Monty's dilemma") problem, correct performance on the problem in its original context was impeded when it was expressed with pairs of people replacing the sides of a card, particularly when those people were individualized and named (Falk & Lann, 2008; Fox & Levav, 2004). There have been a number of studies illustrating that individuals have difficulty applying notions of randomness in social judgments (Agnoli & Krantz, 1989; Lehman, Lempert, & Nisbett, 1993; Nisbett, Fong, Lehman, & Cheng, 1987; Windschitl & Wells, 1998). Given instruction in the application of the law of large numbers, study participants have been shown to more readily apply the principle to judgments regarding athletic skill or job performance rather than friendliness or honesty (Nisbett, et al., 1993). Individuals generally have difficulty differentiating between cause and chance when reasoning about people probabilistically, often making what has been termed a *covariance assumption*: rather than attributing human behavior to chance, people tend to assume that behaviors or opinions are determined exclusively by internal factors (Schwartz & Goldman, 1996). Reasoning probabilistically about people can be complicated by considering intent, which is not the case when reasoning about cards or dice (Howell & Burnett, 1978; Nisbett, et al., 1987).

The implicit confounding of the attribution of both cause and chance factors on an outcome may offer insights as to why problem solving in probability is not regularly facilitated by social-pragmatic context. Probability is particularly affected by domain-specific biases that do not affect performance on *if-then* implication tasks. Research in this field has documented discrepancies between formal quantitative probabilistic assessments and on-the-fly, qualitative probabilistic judgments, leading to a generalized theory of dual-systems representations for

probability judgments. These two processes have been contrasted in terms of *extensional* versus *intuitive* (Tversky & Kahneman, 1983); *rational* versus *experiential* (Epstein, Pacini, Denes-Raj, & Heier, 1996); or *rule-based* versus *associative* (Sloman, 1996), among others. For the purposes of this paper, extensional, rational, and rule-based processes will be referred to as *Type 2* and intuitive, experiential, and associative processes will be referred to as *Type 1*, as summarized by Evans and Frankish (2008, see also Stanovich, Toplak, & West, 2008). Type 2 processes engage the formal procedures prescribed by probability theory. Type 1 processes engage the heuristics that have been defined by how they do not adhere to formal probability theory, such as base-rate neglect, the conjunction fallacy, and availability bias (Tversky & Kahneman, 1974, 1980, 1983). For individuals with the appropriate training, Type 2 processes are dominant in situations requiring quantitative judgments and non-social context. For these same individuals, Type 1 processes are usually engaged when reasoning about social situations, indicating the persistence of these heuristics in certain contexts.

While the two processes have been well-differentiated, it remains to be determined how they may interact in probability problem solving contexts. Namely, how may the heuristics employed in Type 1 probability judgments affect formal calculations requiring Type 2 processes? Some research has indicated that, with respect to social judgments, Type 1 reasoning biases can be superseded by the introduction of a quantifiable metric of comparison, which seems to cue formal Type 2 processes (Nisbett, et al., 1993). Stanovich et al. (2008) term this process *Type 2 override*, the process of suppressing Type 1 processes by separating the abstract formal schema from the individual problem schema. The formal schema becomes a simulation representation of the problem space which can then be manipulated using formal solution processes. Often the

recognition that a formal schema is applicable is facilitated by a randomness cue, which serves to introduce a stochastic component onto a situation in which it is not otherwise salient. For example, the well-established restaurant script has been experimentally altered with the detail that an individual orders from an unreadable restaurant menu by dropping a pencil on the menu and requesting the item on which it lands (Fong, et al., 1993). Type 2 override in probability problems with social context may be facilitated by a randomness cue or by demand characteristics of the item, which indicate that a formal quantitative assessment of mathematical probability is required over and above a quick judgment (Nisbett, et al., 1993).

Models of problem solving have defined problem solving processes as sets of required skills or knowledge types as well as sequential sets of phases. Some of these are more relevant as to when Type 2 override may be engaged. Mayer (1992) included *semantic knowledge*, knowledge of the world, and *schematic knowledge*, knowledge of problem type, in a set of knowledge types necessary for problem solving. These definitions parallel the Type 1-Type 2 distinction previously described. Reusser (1996, see also Zahner & Corter, 2010) proposed a five-stage model of mathematical problem solving, with stage three being the phase during which the situation model of the problem is translated into formal mathematical representations. In the example given above, it is during this phase of problem solving that the randomness cue of dropping the pencil on the menu included in the situational model of the problem would be formalized in its mathematical solution. With respect to dual-systems representations theory, Type 1 processes would typically be activated when considering real-world pragmatic schemata, Type 2 processes activated when considering formal solution schema, and both processes interact when negotiating an individual problem schema. When the surface content of a problem reflects an established pragmatic schema, the ease with which specific elements of a

problem map onto the formal solution schema can be affected by the degree to which the problem context adheres to or deviates from the relevant real-world schema (Bassok, et al., 1995; Falk & Lann, 2008; Martin & Bassok, 2005). If so, then probability problem solving can be viewed as a three-way coordination and mapping process, associating real-world knowledge through pragmatic schemas to individual problem schemas to formal schemas. To date, these processes have been studied with respect to general application of the law of large numbers or sampling methodology in a qualitative context. The effects of these adjustments have not yet been studied in a probability problem solving context.

CHAPTER III: STUDY 1

Statement of the Problem

Study 1 examines if probability problem solving is affected by whether randomness is appropriate to the pragmatic schema of an item's cover story. It was expected that Type 1 and Type 2 processes would play a larger or smaller role in problem solving depending on the context of the problem's cover story. Previous studies (Bassok, et al., 1995; Falk & Lann, 2008; Howell & Burnett, 1978; Schwartz & Goldman, 1996; Windschitl & Wells, 1998) have demonstrated that people have greater difficulty reasoning statistically about social situations or human behavior than about simple randomizing devices or abstract entities. Much of the work in this area used tasks which do not require mathematical or formal quantitative solutions, while a few more recent studies have illustrated context effects using quantitative problem-solving tasks as well. Study 1 was designed to examine whether Type 1 reasoning interacts with Type 2 solution processes on a formally isometric permutation problem expressed through different cover stories.

Cover stories for a probability problem were written depicting schemas in which individuals expect an element of randomness as well as those in which the social script dictates that a non-random ordering should be applied. Cover stories in the *randomness-appropriate-schema* (RA) condition depicted randomizing devices such as cards or dice, as well as social situations in which random selection is consistent with the script, such as lotteries.

Cover stories in the *randomness-inappropriate-schema* (RI) condition depict situations in which some identifiable non-arbitrary criterion of ordering is an essential element of the script. Specifically, the *first-come, first-served* (1C1S) line-waiting paradigm is an important aspect of

many service-related schemas, including the much studied restaurant script (Shank & Abelson, 1977). The 1C1S script element defines a sequence of events so that, among people waiting for some service, the first to arrive will be the first served, and so forth, in the order of arrival. In experimental cover stories depicting this type of scenario, the expected 1C1S ordering is replaced with random ordering, which is inconsistent with the expected script. By imposing the social schema onto a probability problem schema in which random assignment is required, individuals' performance on the problem was expected to be altered. With respect to Cheng and Holyoak's pragmatic schema theory (1985), the problem would be rendered more difficult, as adding stochasticity to the problem schema would increase the conceptual distance between the problem context and the pragmatic schema. Conversely, should the addition of stochasticity highlight the violation of the 1C1S rule, as Cosmides's theory of social contract (1989) would predict, performance on the problem may be enhanced. According to this view, individuals' native sensitivity to cheating increases the salience of randomness in the inappropriate situation and may induce Type 2 override.

Additional Conditions

Variants of each primary randomness condition were designed to examine whether specific elements of a problem schema would affect performance on the item. It was expected that within randomness-appropriate (RA) items, cover stories depicting randomizing devices would be solved more frequently than items depicting social schemas. This condition sought to support the idea illustrated by Bassok and colleagues (1996) in which, using a common probability item involving permutations, a probability item in which computers were randomly assigned to secretaries was easier to solve than items in which secretaries were randomly

assigned to computers, even though each cover story described a situation in which randomness was expected. Changing the subject of the probability item from objects to people did not radically alter the script of the problem cover story but affected its difficulty. To investigate this effect in Study 1, two types of RA problems were written with cover stories depicting either (a) randomizing devices or (b) randomizing people in social settings. The items using randomizing devices depicted games of chance such as cards or drawing numbered balls. The RA items with individuals in place of devices depicted social situations similar to a lottery, such as a *Secret Santa* gift exchange, in which people select gifts at random from a bag.

The additional condition applied to RI items examined whether a randomness cue might induce Type 2 override and alleviate the difficulty of a probability problem. As mentioned above, in a series of studies investigating the salience of random components of variance in "everyday" reasoning, Fong et al, (1986) imposed randomness on a situation in which it is unexpected. When ordering a meal in a restaurant, people order food according to some criterion, whether it be calories, food allergy, or personal taste. In their variation of the task, however, Fong et al. describe a cover story in which a man is in Japan on business and must order a meal from a menu in a language which he cannot read. The businessman selects a meal randomly by ordering the item on which his pencil falls. The man enjoys his meal and repeats the procedure at the restaurant another evening, yet is disappointed in his second meal. Fong, et al.'s subjects who read this version of the story more often attributed the businessman's disappointment to stochastic factors, relative to subjects who read a story in which the standard restaurant script was followed. It was thought that the randomness cue highlighted the element of stochasticity, facilitating participants' attributing variance in meal enjoyment to chance factors. In addition, it has been shown that individuals are resistant to attributing variation to

"pure" randomness alone, preferring to attribute unexpected outcomes to unknown causal determinants (Krynski & Tenenbaum, 2007; Luhmann & Ahn, 2005). In addition, individuals prefer to bet on known odds of success of 50/50, rather than unknown, possibly better odds (Heath & Tversky, 1991). In Study 1, the secondary condition within randomness inappropriate (RI) items tested this effect with cover stories written either (c) with a cue or (d) without a cue explaining why random assignment was imposed. Cover stories in each secondary RI condition were identical except that one set of items included an explanation for why random selection has replaced 1C1S order. The explanation provides a causal determinant in the problem's cover story and functions as the randomness cue. It was hypothesized that the explanation would decrease the conceptual distance between the problem schema and social schema, facilitating solution of the problem.

Method

Participants

Volunteers for Study 1 participated via the *Amazon Mechanical Turk* (AMT) platform, which is an online labor market in which individuals are recruited for surveys as well as tasks which are difficult for computers and cannot be automated, such as image recognition and filtering for adult content (Buhrmester, Kwang, & Gosling, 2011; Mason & Suri, 2012). All AMT users must be over age 18 according to the terms of that website, although we excluded data for six individuals who reported their age less than 18 years. Participants were further limited to those subjects with a proportion of accepted submissions of AMT tasks above 95% to exclude users 'fishing' for remuneration without reasonably expected effort. This resulted in 394

participants (Table 2). The AMT website is in English, so all participants were expected to have been proficient in English to a degree, regardless of native language.

Materials

Cover stories were written to present a single formal probability problem in different contexts. The formal problem involved finding the likelihood of a single possible permutation of events. Two cover stories were written for each of the four conditions, resulting in eight parallel, formally isometric probability items (Table 3).

Procedure

Study 1 was administered online as an AMT Human Intelligence Task (HIT). A HIT is a short task for which workers get paid a small amount. The bulk of the tasks on AMT cannot be automated, that is, they entail activities on which humans outperform computers.

Amazon Mechanical Turk workers browse or search for HITs by keyword. Keywords defined for this task included *probability thinking*, *word problems*, *math problem solving*, and *opinions*. Clicking on the HIT title, "Solve a probability problem" took the AMT participant to a screen describing the task. The instructions informed potential participants that they would be completing an introductory probability problem and suggested to have a pen, paper, and calculator at hand. Users clicked a button labeled “Accept HIT”, upon which task appeared onscreen in a frame (Figure 1).

The AMT algorithm allowed each of the eight probability problems to be presented randomly to participants. The HIT containing these items appeared online between March and

Table 2

Study 1: Descriptive Statistics on the Total Sample and by Condition

			RANDOMNESS APPROPRIATE		**RANDOMNESS INAPPROPRIATE**	
		Total	**Randomizing Devices**	**Social Schema Consistent**	**Social Schema Inconsistent, with Explanation**	**Social Schema Inconsistent, without Explanation**
	N (% Total)	*394*	*99 (25.1%)*	*98 (24.9%)*	*99 (25.1%)*	*98 (24.9%)*
Age (years)	*M (SD)*	*28.4 (9.1)*	*27.6 (8.7)*	*29.7 (9.3)*	*28.0 (9.6)*	*28.4 (8.7)*
Gender	*N (%)*					
Male		*270 (68.5%)*	*73 (73.7%)*	*67 (67.7%)*	*64 (65.3%)*	*66 (67.3%)*
Female		*115 (29.2%)*	*23 (23.2%)*	*31 (31.3%)*	*31 (31.6%)*	*30 (30.6%)*
Not reported		*9 (2.3%)*	*3 (3.0%)*	*1 (1.0%)*	*3 (3.1%)*	*2 (2.0%)*
Region/Language	*N (%)*					
English-speaking countries		*275 (69.8%)*	*62 (62.6%)*	*75 (75.8%)*	*73 (74.5%)*	*65 (66.3%)*
European, Spanish-speaking countries		*46 (11.7%)*	*16 (16.2%)*	*8 (8.1%)*	*9 (9.2%)*	*13 (13.3%)*
Asian countries		*73 (18.5%)*	*21 (21.1%)*	*16 (16.3%)*	*16 (16.3%)*	*20 (20.4%)*
Time on task (minutes)	*M (SD)*	*3.47 (2.6)*	*3.61 (2.6)*	*3.61 (2.2)*	*3.44 (2.4)*	*3.24 (2.3)*

Note: Statistical tests indicated no statistical differences among condition on these variables.

April 2011. Participants entered an answer in a text box below the item and had the option of commenting on their solution. Below these text boxes, a list of relevant probability formulas (Figure A1) was presented, followed by some demographic items. The "Submit" button followed the demographic items.

Another procedure in addition to the "95% acceptance" rule was implemented to regulate quality of the data by limiting only one response per AMT worker. AMT allowed each HIT to be accepted or rejected before participants were paid. At any point while the HIT was active, investigators were able to download a file with participants' response data and anonymous

Table 3

Study 1: Test Items by Condition

RANDOMNESS APPROPRIATE
Randomizing devices
Eight cards numbered one through eight are shuffled. They are then dealt one at a time. What is the probability they are dealt in increasing numerical order?
A special pool table at the pool hall has only eight balls numbered one through eight. To play, a customer inserts $2.00 and the balls are released into a tray in the side of the table. What is the probability that they appear in increasing numerical order?
Social schema - consistent
Day-use lockers at the gym are assigned randomly by selecting keys from a bowl. The bowl contains keys numbered one though eight. What is the probability that the eight locker keys get assigned in increasing numerical order?
Eight 'Secret Santa' gifts of different value are in a bag. What is the probability that the gifts are randomly selected in order from most to least expensive?

identifiers, select which submissions to accept, and upload the file back to AMT. The HIT instructions explicitly stated that only one HIT per worker ID would be accepted, yet several participants submitted multiple HITs. In these cases only the first HIT submitted was accepted and the rejected tasks were returned to the item pool. An additional constraint limited this occurrence: Participants' worker IDs were compiled in a list and a rule applied so those workers' with IDs in the list were prohibited from further participation. This method helped minimize the frequency of rejected HITs. Each participant was paid 50 cents for an accepted HIT.

Table 3 (continued)

Study 1: Test Items by Condition

RANDOMNESS INAPPROPRIATE
Social schema - inconsistent: 1C1S replaced with random order, with explanation
Eight people are waiting in line to be served at the post office. There is a fire alarm and the post office is evacuated. It was a false alarm, but now the eight people cannot reassemble the line and must be served in random order. What is the probability that they are served in the original order of the line?
Eight parties are waiting for tables at a restaurant. The restaurant computer has crashed, losing the waiting list. Parties must now be seated in random order. What is the probability that the parties are seated in the original order in which they arrived?
Social schema - inconsistent: 1C1S replaced with random order, without explanation
Eight people have arrived at the post office, but are served in random order. What is the probability that they are served in the order of their arrival?
Eight parties have arrived and are waiting for tables at a restaurant, but are seated in random order. What is the probability that the parties are seated in the order in which they arrived?

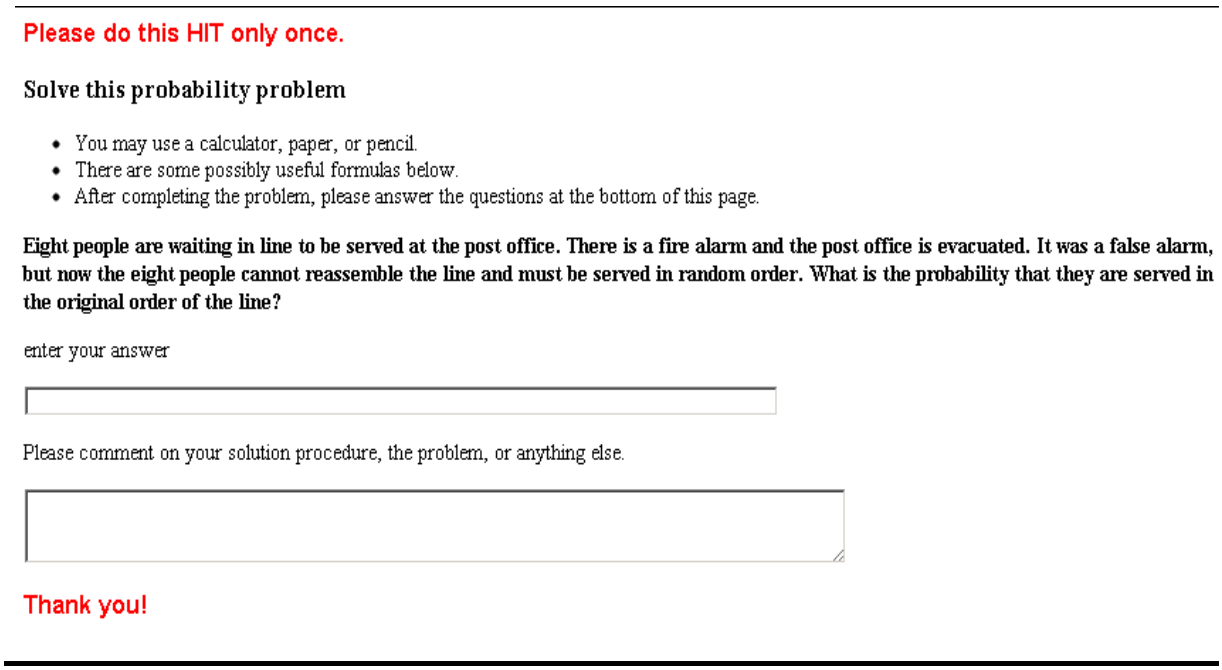
Note: '1C1S' = First-come, first-served

Analyses

Preliminary analyses indicated that there were no differences in proportion correct between pairs of items within each of the four conditions, so observations for both items within each condition were combined into a single group. Due to the diversity in the sample, region was included as a control in analyses to account for possibility of cultural or linguistic effects (see Table 2). Participants had been asked in which country or countries they were educated; these answers were then coded to group participants into three regions: English-speaking countries, Asia, and the rest of the world. The groups are defined in Table A1 in Appendix A. Other covariates considered included gender, age, field of study, and a variable in which AMT had automatically collected data on the time spent on each HIT in seconds. Preliminary analyses on these control variables showed mean time on task differed significantly among

Figure 1

Study 1: Presentation of Each Probability Item on the Amazon Mechanical Turk Website



region groups. Other variables did not correlate with answering an item correctly and were not included in further analyses.

A Generalized Linear Modeling (GZLM) procedure, SPSS GENLIN, was used to model these data according to a binary distribution with logit (i.e., log-odds) link. This procedure is statistically more robust than procedures such as logistic regression in that GZLM does not require that the response data are distributed normally for valid results. Parameter estimates can be transformed and interpreted as odds ratios: the odds of solving an item correctly, given a condition, over the odds of solving the item given the reference condition, holding all other independent variables at fixed values. Measures of model fit were estimated using likelihood ratio chi-square statistics (χ^2_{LR}). Definitions of the primary research question, along with three secondary research questions, may be found in Table 4.

Table 4

Study 1: Research Questions Expressed Through Contrast-Coded Variables for Statistical Analyses

Primary condition	Secondary condition	ST1	ST2	ST3	ST4
RANDOMNESS APPROPRIATE	Randomizing Devices	C	-	-	C
	Social Schema Consistent	C	C	-	R
RANDOMNESS INAPPROPRIATE	Social Schema Inconsistent, with Explanation	R	R	C	-
	Social Schema Inconsistent, without Explanation	R	R	R	-

Note: 'R' = reference group; 'C' = comparison group; '-' = not included in analysis

Primary research question:

ST1 Is performance on randomness appropriate items better than performance on randomness inappropriate items?

Secondary research questions:

ST2 Is performance on items in which the cover story depicts a social schema consistent with random assignment better than on items in which a 'first-come, first-serve' (1C1S) schema is replaced with random assignement?

ST3 Within the randomness inappropriate condition, does an explanation for why 1C1S is replaced with random assignment improve performance?

ST4 Within the randomness appropriate condition, are probability items depicting randomizing devices easier to solve than items depicting social situations in which random assignment is expected?

Results

Overall, 59.4% of participants correctly solved the problem. Solution rates by condition are listed in Table 5. Results by detailed region are presented in Table A2 in Appendix A.

Four contrast-coded variables were used to test the primary research question and three secondary questions. These questions and the groups compared with respect to each question are listed in Table 4. Each research question (RQ) was expressed through a single degree-of-freedom contrast-coded variable. The primary RQ, as well as each secondary RQ, was tested separately in GZLM analyses both with and without covariate region.

Results from GZLM analyses are summarized in Tables 6 through 9, including results from tests of model fit and parameter estimates for each predictor and covariate. The primary research question (ST1) compared performance on randomness-appropriate (RA) items versus

Table 5
Study 1: Proportion Correct for Total and by Condition

Primary condition	Secondary condition	*n (%)* correct	all *N*
RANDOMNESS APPROPRIATE	Randomizing Devices	*66 (66.7)*	*99*
	Social Schema Consistent	*63 (63.6)*	*98*
RANDOMNESS INAPPROPRIATE	Social Schema Inconsistent, with Explanation	*53 (54.1)*	*99*
	Social Schema Inconsistent, without Explanation	*52 (53.1)*	*98*
	TOTAL *N*	***234 (59.4)***	***394***

randomness-inappropriate (RI) items (see Table 6). In an ST1-only model, the model fit the data significantly better than an intercept-only model, X^2_{LR} (1, N = 394) = 5.70, $p < .017$. The odds ratio for RA items versus RI items was 1.64 ($p < .017$).

The model including ST1 and the variables for region fit the data significantly better than the ST1-only model, X^2_{LR} (3, N = 394) = 25.39, $p < .001$. Controlling for region effects, the odds ratio for the RA versus RI condition was 1.67 ($p < .018$). Odds ratios for region, controlling

Table 6

Study 1: Results From Generalized Linear Model Analyses: Primary Research Question (N = 394)

	Odds[a]	*B*	X^2_{LR}	*df*	*p*	Model X^2_{LR}	*df*	*p*
Model: Randomness								
						5.70	*1*	*.017*
ST1: Random appropriate versus Random inappropriate	*1.64*	*0.50*	*5.70*	*1*	*.017*			
Model: Randomness with Covariates								
						25.39	*3*	*.001*
ST1: Random appropriate versus Random inappropriate	*1.67*	*0.51*	*5.56*	*1*	*.018*			
Region[b]			*14.42*	*2*	*.001*			
Asia versus English-speaking	*0.45*	*-0.80*			*.004*			
Europe and Latin America versus English-speaking	*2.02*	*0.70*			*.068*			
Model: Covariate only								
						19.40		*.001*
Region			*19.40*		*.001*			
Asia versus English-speaking	*0.45*	*-0.80*			*.003*			
Europe and Latin America versus English-speaking	*2.66*	*0.98*			*.013*			

[a]Odds are calculated with respect to reference groups defined in Table 4

[b]Reference group = English-speaking countries

for condition, were 0.45 ($p < .004$) for participants from Asia vs. English-speaking countries and 2.02 ($p < .068$) for participants from the rest of the world relative to English-speaking countries.

The effect of randomness as an expected or unexpected characteristic of the schema of a probability item's cover story was statistically significant, controlling for regional differences. Analyses addressing the sensitivity of the effect of ST1 to region are presented in Appendix A.

Tables 7 - 9 summarize results from GZLM analyses on the secondary research questions ST2 – ST4, respectively. None of the tests related to the secondary research questions were significant with or without controlling for region effects.

Table 7

Study 1: Results From Generalized Linear Model Analyses: Additional Research Questions (N = 295)

	Odds[a]	*B*	X^2_{LR}	*df*	*p*	Model X^2_{LR}	*df*	*p*
Model: Social Schema Consistency								
						2.74	*1*	*.098*
ST2: Social schema-consistent versus schema-inconsistent	*1.52*	*0.042*	*2.74*	*1*	*.098*			
Model: Social Schema Consistency with Covariates								
						14.52	*3*	*.002*
ST2: Social schema-consistent versus schema-inconsistent	*1.53*	*0.43*	*2.74*	*1*	*.098*			
Region[b]			*11.78*	*2*	*.003*			
Asia versus English-speaking	*0.39*	*-0.95*			*.003*			
Europe and Latin America versus English-speaking	*1.61*	*0.48*			*.261*			

[a]Odds are calculated with respect to reference groups defined in Table 4

[b]Reference group = English-speaking countries

Table 8

Study 1: Results From Generalized Linear Model Analyses: Additional Research Questions (N = 196)

	Odds[a]	*B*	X^2_{LR}	*df*	*p*	Model X^2_{LR}	*df*	*p*
Model: Explanation								
						0.02	*1*	*.886*
ST3: Schema-inconsistent with explanation versus without explanation	*1.04*	*0.12*	*0.02*	*1*	*.886*			
Model: Explanation with Covariates								
						12.22	*3*	*.007*
ST3: Schema-inconsistent with explanation versus without explanation	*1.00*	*0*	*0*	*1*	*.998*			
Region[b]			*12.20*	*2*	*.002*			
Asia versus English-speaking	*0.27*	*-1.31*			*.001*			
Europe and Latin America versus English-speaking	*1.23*	*0.21*			*.663*			

[a]*Odds are calculated with respect to reference groups defined in Table 4*

[b]*Reference group = English-speaking countries*

Table 9

Study 1: Results From Generalized Linear Model Analyses: Additional Research Questions (N = 198)

	Odds[a]	*B*	X^2_{LR}	*df*	*p*	Model X^2_{LR}	*df*	*p*
Model: Social Context								
						0.20	*1*	*.655*
ST4: Random devices versus Social schema-consistent	*1.14*	*0.13*	*0.20*	*1*	*.665*			
Model: Social Context with Covariates								
						15.79	*3*	*.001*
ST4: Random devices versus Social schema-consistent	*1.04*	*0.04*	*0.02*	*1*	*.902*			
Region[b]			*15.59*	*2*	*.001*			
Asia versus English-speaking	*0.70*	*-0.36*			*.332*			
Europe and Latin America versus English-speaking	*13.53*	*2.61*			*.012*			

[a]*Odds are calculated with respect to reference groups defined in Table 4*

[b]*Reference group = English-speaking countries*

Discussion

The effect of randomness as an expected or unexpected characteristic of the schema of a probability item's cover story was statistically significant. Participants receiving RA items were 1.67 times more likely to solve the problem correctly than participants receiving RI items, accounting for regional differences. Regional comparisons, controlling for condition, indicated that participants from Asia were less than half as likely as participants from English-speaking countries to solve the same type of item, while there was no significant difference between the likelihood of participants from the rest of the world and participants from English-speaking countries to solve the same type of problem.

These results indicate that there is an added level of difficulty when solving a probability item in which the cover story depicts a situation contrived so that the expected order of events is replaced with random ordering. With respect to social contracts (Cosmides, 1989), replacing 1C1S order with random selection should have invoked an innate sensitivity to cheating and facilitated application of randomness. The data in this study do not support this interpretation. Participants performed significantly better on items in which randomness was expected, lending support to Cheng and Holyoak's (1985) pragmatic schema theory. Specifically, the line-waiting, first-come first-served (1C1S) convention may be considered a deontic relation; in those terms, the 1C1S convention may be expressed as "If I arrive before you, then I will be served before you." Removing this element increases the distance between the probability problem schema and the social schema. This distance may further reduce the likelihood of schema induction, the process by which identifying an appropriate formal solution schema is affected by the strength of its association with the problem schema (Gick & Holyoak, 1980).

Analyses on secondary research questions did not illustrate any significant differences. The first of these, ST2 (Table 7), indicated that there was no significant difference between performance on social-schema items with respect to whether randomness was appropriate or not to the schema. The test on ST3 (Table 8) sought to reflect that within the RA condition, it was thought that performance on items depicting randomizing devices would be better than on items depicting social situations. The data in Study 1 did not illustrate this phenomenon. With respect to the RI secondary condition, ST4 (Table 9), the data did not indicate an effect of a randomness cue. Based on previous studies, it was thought that including an explanation for the replacement of 1C1S ordering with random selection would make the stochasticity in the problem more apparent and improve performance on those items. The Study 1 task required calculations and differed from those in which a randomness cue was shown to be effective, which required only qualitative interpretations of the phenomenon (Fong et al., 1986).

For both RA and RI items, it may be that the relative computational simplicity of the formal problem structure did not provide opportunity to demonstrate some hypothesized effects. With respect to the additional conditions within each set of RA and RI probability problems, participants were not asked to provide any interpretation of results. It is possible that further contextual details, beyond whether or not randomness was appropriate, would have affected performance at a more nuanced level than the task demanded, as has been demonstrated by others (Fong, et al., 1993; Howell & Burnett, 1978; Schwartz & Goldman, 1996; Windschitl & Wells, 1998). This may reflect the distinction between Type 1 and Type 2 processes. Requiring calculation has been shown to induce Type 2 override, which may have obscured any meaningful distinction in perceived stochasticity between RA problems depicting objects and those depicting social situations as well as rendered the randomness cue in RI items unnecessary.

CHAPTER IV: STUDY 2

Statement of the Problem

With respect to temporal and causal schemas, it is the convention to reason forwards, considering earlier events before those that happen later, and causes before effects, a preference which has been shown to affect the perception of the strength of a causal relationship (Cheng, et al., 1993; Tversky & Kahneman, 1980). In addition, when given a choice, individuals prefer to wager on the outcomes of events that have yet to happen rather than unknown outcomes of past events (Brun & Teigen, 1990; Fischhoff, 1975, 1976; Rothbart & Snyder, 1970; Wright, 1982). Probability problems addressing conditional probability can be constructed using the same formal problem structure with respect to the likelihood of events in either anterograde or retrograde order. Comparing the difficulty of problems differing in this respect, controlling for other factors, is a particularly apt way to address this issue. It would be expected that both convention and preference for prediction over postdiction may affect problem difficulty with respect to the order of events as depicted in a problem schema. Specifically, a problem asking for the likelihood of a cause given the likelihood of an effect would be more difficult than a problem asking for the likelihood of an effect given a cause. A similar effect should also be found between earlier and later events. In sum, inverting the direction of chronology of events should introduce an additional level of difficulty to a problem despite equivalent computations.

The difficulty introduced by inverting the order of events may be affected by a common error with respect to conditional probability, termed the *fallacy of the transposed conditional* or the *inverse fallacy*, in which $P(A|B)$ is expressed as $P(B|A)$ (Bar-Hillel & Falk, 1982; Batanero, Henry, & Parzysz, 2005; Diaconis & Freedman, 1981; Díaz & Fuente, 2007; Krynski & Tenenbaum,

2007; Mackie, 1981; Neath, 2010; Tversky & Kahneman, 1980; Villejoubert & Mandel, 2002). Among these discussions there has been speculation of conditions in which this fallacy is more or less likely to occur, but little evidence demonstrating the phenomenon systematically within judgment under uncertainty.

In Study 2, items were constructed to test whether the order of events affects errors in expressing conditional probability, in turn making an item more difficult to solve. Consider two events, *A* and *B*. Given $P(A)$, $P(B)$, and $P(A \text{ and } B)$, it is possible to solve for either $P(A|B)$ or $P(B|A)$ using the same formula and calculations. If in the cover story of the problem, *A* occurs before *B* or causes *B*, the conditional probability $P(B|A)$ reflects a schema-consistent, intact order of events. The conditional probability $P(A|B)$, on the other hand, represents those events in an order inconsistent with a temporal schema. The events are considered with respect to an inverted order: the conditional probability demands consideration of an event's occurrence in light of an event which is described as happening later.

In Study 2 temporally-ordered events were depicted in probability problems with a cover story involving catching an express or local bus (Event *A*) and arriving on time to an appointment (Event *B*). These problems required Bayes's formula for a correct solution. Causal events were depicted in problems with a cover story involving a physical state (cause) and a disease (effect). Specifically, causal items involved the relationship between obesity (Event *A*) and Type 2 diabetes (Event *B*). Causal problems involved calculating a conditional probability from joint and simple probabilities.

Method

Participants

Two cohorts participated in Study 2. One cohort of participants (N = 19) was recruited from classes in probability and statistics from a graduate school in the social sciences. They were paid $5 for their participation. They were predominantly female, n = 13 (68.4%). To contrast these participants with the cohort in Study 1, they were asked whether they were AMT workers. All but four participants reported that they had never heard of AMT (79.0%).

Another cohort of participants (N = 59) consisted of all students enrolled in two sections of an introductory probability and statistical inference course administered online during the Summer 2011 term.

Materials

Test items are listed in Table 10 by condition. Each set of items is formally isomorphic, requiring the same mathematical solution. The pair of causal items contained the same information and required the application of the same formula. Differences in the cover story between items are highlighted in the table here for clarity; test materials were not formatted thusly. *Order-intact* (NT) items ask for a solution calculating the conditional probability of a later event (arriving on time) given an earlier event (catching an express bus) or the probability of an effect (having diabetes) given a cause (being obese). *Order-inverted* (NV) items ask for a solution calculating the conditional probability of an earlier event (catching an express bus)

given a later event (arriving on time) or the probability of a cause (being obese) given an effect (having diabetes).

Table 10

Study 2: Test Items by Condition

Temporal items
Order Intact (forward direction)
You are waiting to meet your friend. He phones saying that he is getting on the next bus. From experience you know that the probability that he arrives on time is 25%. When he arrives on time, it is 90% likely that he took an express bus. When he arrives late, it is 65% likely that he took an express bus. He took an express bus. What is the likelihood that he arrives on time?
Order Inverted (backward direction)
You are waiting to meet your friend. He phones saying that he is getting on the next bus. From experience you know that the probability of getting an express bus is 25%. On an express bus, he is 90% likely to arrive on time. On a local bus, he is 65% likely to be on time. He arrived on time. What is the likelihood that he took an express bus?
Causal items
Order Intact (forward direction)
There is some evidence that obesity causes Type 2 diabetes. About 25% of the United States adult population is obese. 8% of the US adult population have Type 2 diabetes. 5% of the population are obese and have Type 2 diabetes. A randomly selected adult is obese. What is the probability that this person has Type 2 diabetes?
Order Inverted (backward direction)
There is some evidence that obesity causes Type 2 diabetes. About 25% of the United States adult population is obese. 8% of the US adult population have Type 2 diabetes. 5% of the population are obese and have Type 2 diabetes. A randomly selected adult has Type 2 diabetes. What is the probability that this person is obese?

Participants recruited in person received test items in a six-page packet in which they were provided with space to write out solutions. An introductory page was followed by three pages with one probability item per page. Questions were balanced so that each participant received one NT item and one NV item. A "distractor" item from Study 1 was presented between the two Study 2 items. The fifth page in the packet included demographic questions assessing age, race, gender, countries of education, and field of study. The remainder of the page was allocated for comments. The last page of the packet was a sheet of probability formulas (Figure A1).

Participants in the online course were administered test items within a required five-item quiz. Quizzes were required periodically in this course and students had experience completing them in the online distance-learning environment. Each item was followed by a space into which the student was asked to type the solution as well as show work. Examples of student work can be seen in Figures 3 and 4 on pages 44 - 45. The temporal item was third and the causal item fifth in the quiz. As with the in-person cohort, each student received one NT and one NV item.

Procedure

In-person volunteers

Participants were recruited from statistics courses during the Spring and Summer 2011 terms. At the beginning of a class meeting, students were told that an experimenter would arrive ten minutes before the end of the class to recruit volunteers. At the end of class, the experimenter described the general nature of the study, remuneration, and its purpose as dissertation research. Volunteers remained after class, signed consent forms, and were given

test packets randomly. There was no specific time limit, and students took between 15 and 25 minutes on the task. Participants submitted their packets individually when done, were paid, and thanked effusively. Twelve students were willing to participate but unable to stay after class. These students left their email addresses and were later contacted to schedule individual meetings. Six of these students completed the task in the library and were similarly paid and thanked.

Online students

In the online courses, the quizzes with Study 2 items were posted in the course distance learning platform in June 2011. Students in the online courses were given a period of 45 minutes to complete and submit the quiz. Students received instructions to show work and were prohibited from collaborating. These participants were required to enter supporting formulae for their answers; they did not enter an answer alone.

A course teaching assistant (TA) copied each student's quiz into a word document, removing all identifying information, which was provided to investigators. Quizzes were graded for correct solution by both the TA and an investigator.

Analyses

Variables of interest

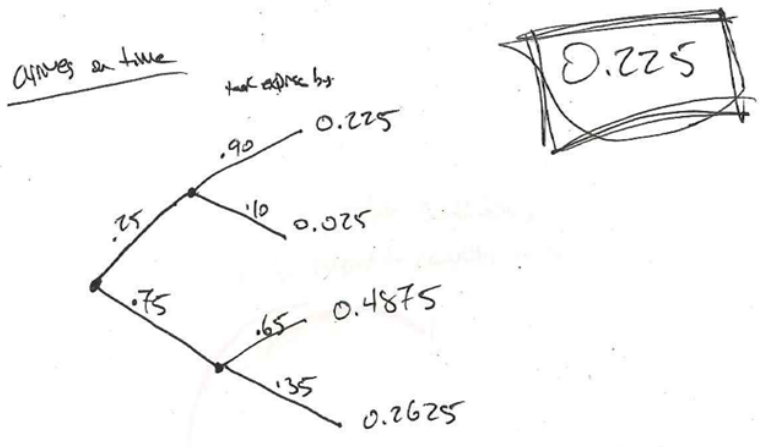
There were two dependent variables (DVs) for each item. First, each solution was scored to indicate whether a student had solved for the correct mathematical answer, allowing for rounding error.

Second, written or typed work was coded to indicate whether a participant had formally symbolized conditional probability (CP) correctly. Formalizing CP correctly is a necessary yet not sufficient condition in solving these test items and the most relevant step with respect to the experimental manipulation. A student needed to express the relevant CP in the correct order either explicitly, e.g., "$P(A|B) = .90$", or use the value of the CP in its appropriate position within a formula. For example, if the student used the formula $P(A|B) = P(A \text{ and } B) / P(B)$, the values used must reflect the correct events. If the problem text asked for $P(A|B)$ but the student used values expressing $P(A \text{ and } B) / P(A)$, which equals $P(B|A)$, the CP was coded as incorrect.

In some written protocols it was possible to infer that a student correctly identified CP from a tree diagram (see Figure 2). In the example in Figure 2, the structure of the tree indicates that the student is expressing the given CPs in the proper order, while numeric values given in the problem text are assigned to the correct branches. In such cases CP was coded as correct.

Figure 2

Study 2: A Tree Diagram Drawn by a Participant in the Temporal Inverted Condition



A few participants did not show work on the items but reported a correct solution. In these cases coding CP was scored as correct. As previously stated, coding CP correctly is a necessary step in solving these items, so a correct answer implied that CP was coded in the correct order. In written protocols showing work, there was only one case in which a student coded CP incorrectly and arrived at a correct answer. Acknowledging that case as rare, it was considered reasonable to use this rule.

The primary independent variable of interest was order of events, either NT or NV.

Interrater reliability

To estimate the interrater reliability (IRR) of the scheme coding both mathematical solution and formal expression of CP, a third of the online quizzes were coded by a second investigator with considerable experience with statistics instruction but not otherwise affiliated with the study. Agreement was measured using Cohen's kappa (k).

Experimental analyses

Proportion correct was calculated for each item for both in-person and online cohorts. Proportion correct on coding CP was also calculated for both cohorts.

The procedure and samples for each of the two Study 2 cohorts were not similar enough to justify combining their data. Binary logit GZLM analyses were performed on data from the online cohort only due to the small size of the in-person sample. These analyses used the same methodology as described for Study 1 but with no covariates.

Error analyses

A set of post-hoc analyses examined the frequency of several well-documented types of errors in probability problem solving. Errors in formalizing CP were classified into one of three common types: *transposed conditional*, in which $P(E_1|E_2)$ was expressed as $P(E_2|E_1)$; *compound substitution*, in which $P(E_1$ and $E_2)$ was mistaken for $P(E_1|E_2)$, and *simple substitution*, in which $P(E_1)$ or $P(E_2)$ was substituted for $P(E_1|E_2)$. Previous research has shown that these are the most common errors in understanding conditional probability (Bar-Hillel & Falk, 1982; Falk, 1986; Krynski & Tenenbaum, 2007; Mackie, 1981; Neath, 2010; Villejoubert & Mandel, 2002), although not all errors observed in the data could be classified into one of these categories. Figures 3 and 4 (pages 44 - 45) illustrate examples of CP expressions coded correctly and by common error.

Results

Interrater reliability

Interrater reliability was measured using Cohen's *kappa*. For temporal (bus) items agreement on solving the problem was $k = .88$, $p < .001$ and for coding CP, $k = .79$, $p < .001$. For causal (diabetes) items, IRR on correct solution was perfect, $k = 1$; IRR on coding CP was $k = .89$, $p < .001$. Agreement was considered good to very good.

Experimental analyses

Proportion correct for both solution and expression of CP for each condition are summarized in Table 11 by cohort. There were no significant differences between performance on items by condition for the in-person cohort. Correlations among the four DV are reported in Appendix B.

Table 11

Study 2: Performance on Items by Cohort and Condition

In-Person (*N* = 19)

Temporal order		**Intact**	**Inverted**	**Total**
	n per group (%)	*9 (47.4)*	*10 (52.6)*	
Solving the problem	*n (%) correct*	*4 (33.3)*	*1 (10.0)*	*4 (21.0)*
Coding conditional probability	*n (%) correct*	*6 (66.7)*	*8 (80.0)*	*14 (73.7)*
Causal order		**Intact**	**Inverted**	**Total**
	n per group (%)	*9 (47.4)*	*10 (52.6)*	
Solving the problem	*n (%) correct*	*4 (44.4)*	*4 (40.0)*	*8 (42.1)*
Coding conditional probability	*n (%) correct*	*4 (44.4)*	*5 (50.0)*	*9 (47.3)*

Online (*N* = 59)

Temporal order		**Intact**	**Inverted**	**Total**
	n per group (%)	*31 (52.5)*	*28 (47.5)*	
Solving the problem	*n (%) correct*	*5 (16.1)*	*7 (25.0)*	*12 (20.3)*
Coding conditional probability	*n (%) correct*	*16 (51.6)*	*24 (85.7)*	*40 (67.8)*
Causal order		**Intact**	**Inverted**	**Total**
	n per group (%)	*31 (52.5)*	*28 (47.5)*	
Solving the problem	*n (%) correct*	*21 (67.7)*	*14 (50.0)*	*35 (59.3)*
Coding conditional probability	*n (%) correct*	*23 (74.2)*	*21 (75.0)*	*44 (74.6)*

Results from binary logit GZLM analyses on the online data are summarized in Table 12. Of models testing the four DVs examined in Study 2, only the model estimating likelihood of formalizing CP for the temporal items fit the data significantly better than the intercept-only model, X^2_{LR} (2, N = 59) = 8.24, p < .004. The odds ratio for the effect of intact order relative to inverted order is 0.17 (p < .008).

Error analyses

Errors by cohort and condition are reported in Table 13. There were no significant differences in the number of errors between conditions for either temporal or causal items for the in-person cohort. For the online cohort, the group asked to solve for a CP in intact order produced more transposed conditional and compound substitution errors in expressing CP than the group asked to solve for CP in inverted order, but statistical tests were not significant. The directionality of the results will be further addressed in the Discussion.

Table 12

Study 2: Results From Generalized Linear Model Analyses – Online Cohort Only

N = 59	Odds[a]	*B*	X^2_{LR}	*df*	*p*
Temporal order					
Solving the problem	*1.73*	*0.55*	*0.72*	*1*	*.401*
Coding conditional probability	*5.62*	*1.73*	*8.24*	*1*	*.004*
Causal order					
Solving the problem	*2.10*	*0.74*	*1.93*	*1*	*.165*
Coding conditional probability	*0.96*	*-0.04*	*0.01*	*1*	*.943*

[a]*Reference group = inverted*

Table 13

Study 2: Distribution of Prototypical Encoding Errors by Cohort and Condition

In-Person (*N* = 19)

Temporal Items	**Intact**	**Inverted**	**Total Errors**
n errors per condition (% error type)	*2 (40.0)*	*3 (60.0)*	*5*
Transposed Conditional	*1 (50.0)*	*1 (33.3)*	*2 (40.0)*
Compound Substitution	*0*	*1 (33.3)*	*1 (20.0)*
Simple Substitution	*0*	*0*	*0*
Unclassified	*1 (50.0)*	*1 (33.3)*	*2 (40.0)*

Causal Items	***Intact***	***Inverted***	***Total Errors***
n errors per condition (% error type)	*4 (44.4)*	*5 (55.6)*	*9*
Transposed Conditional	*1 (25.0)*	*3 (80.0)*	*5 (55.6)*
Compound Substitution	*3 (75.0)*	*1 (20.0)*	*4 (44.4)*
Simple Substitution	*0*	*0*	*0*
Unclassified	*0*	*0*	*0*

Online (*N* = 59)

Temporal Items	**Intact**	**Inverted**	**Total Errors**
n errors per condition (% error type)	*18 (85.7)*	*3 (14.7)*	*21*
Transposed Conditional	*5 (27.8)*	*0*	*5 (23.8)*
Compound Substitution	*6 (33.3)*	*0*	*6 (28.6)*
Simple Substitution	*1 (5.6)*	*1 (33.3)*	*2 (9.5)*
Unclassified	*6 (33.3)*	*2 (66.7)*	*8 (38.1)*

Causal Items	**Intact**	**Inverted**	**Total Errors**
n errors per condition (% error type)	*7 (46.7)*	*8 (53.3)*	*15*
Transposed Conditional	*1 (14.3)*	*2 (25.0)*	*3 (20.0)*
Compound Substitution	*3 (42.9)*	*1 (12.5)*	*4 (26.7)*
Simple Substitution	*0*	*0*	*0*
Unclassified	*3 (42.9)*	*5 (62.5)*	*8 (53.3)*

Figure 3

Study 2: Solutions and Prototypical Errors to Temporal Items from Online Cohort.

Temporal item: intact order

You are waiting to meet your friend. He phones saying that he is getting on the next bus. The probability that he arrives on time is 25%. When he arrives on time, it is 90% likely that he took an express bus. When he arrives late, it is 65% likely that he took an express bus. He took an express bus. What is the likelihood (probability) that he arrives on time?

Correct solution

P(Express) = P(OT and Express) + P(Late and Express)
=(.25)(.90) + (.75)(.65) = .225 + .4875
= .7125 is probability of taking an express bus
So P(OT|Express) = P(OT and Express) / P(Express)
0.315789474
= . 316 is the probability that he is on time given he took an express bus.

Transposed Conditional and Simple Substitution Errors

P(OT) = .25 P(L) = .75
P(OT/EX) = .90 **[*Should be P(EX/OT)*]**
P(L/EX) = .65 **[*Should be P(EX/L*)]**
P(OT) = ?
P(OT n EX) = P(OT)P(OT/EX) = (.25)(.90) = .225
P(L n EX) = P(L)P(L/EX) = (.75)(.65) = .4875
P (OT) = late/not late and also express bus = .4875 + .225 = .7125 **[*Has solved for P(E*)]**

Temporal item: inverse order

You are waiting to meet your friend. He phones saying that he is getting on the next bus. The probability of getting an express bus is 25%. On an express bus, he is 90% likely to arrive on time. On a local bus, he is 65% likely to be on time. He arrives on time. What is the likelihood (probability) that he took an express bus?

Correct solution

Let E = Express Bus Let OT = On time Let L = Local Bus
Using Bayes Rule:
P(E|OT) = P(OT|E)*p(E) / P(OT)
P(OT) = P(OT and E) + P(OT and L)
P(OT) = P(OT|E)p(E) + P(OT|L)*p(L)
P(OT) = (.9)*(.25) + (.65)*(.75) = .7125
So (.9)*(.25) / (.7125) = .3158

Simple Substitution Error

The probability of being on time when taking express: (0.25)(0.9)=0.225, and the probability of being on time when taking local: (0.75)(0.65)=0.4875. When adding all these probabilities, the likelihood of being on time is (0.225)+(0.4875)=0.7125.
Hence, the probability of taking express bus is 71.25% **[*Student solved for P(E*)]**

Compound Substitution Error

Probability of express bus (A) given arrival on time (B).
P(A^B)=P(A and B)/P(B)
P(A^B)=P(.90)/P(.25)
P(A^B)= **[*Student assigned value of P(B/A) to P(A and B)*]**

Note: Explanatory comments are in brackets

Figure 4

Study 2: Solutions and Prototypical Errors to Causal Items from Online Cohort

Causal item: intact order

There is some evidence that obesity causes Type 2 diabetes. About 25% of the United States adult population is obese. 8% of the US adult population have Type 2 diabetes. 5% of the population are obese and have Type 2 diabetes. A randomly selected adult is obese. What the probability that this person has Type 2 diabetes?

Correct solution:

Define:
P(obese)=P(O)=0.25,
P(type 2 diabetes)=P(2D)=0.08,
Also it is given that:
P(O and 2D)=0.05.
Then, P(2D, given O)=P(O and 2D)/P(O)=0.05/0.25=0.2

Transposed Conditional Error

0.5/0.8=0.625 **[***P(O and D)/P(D) = P(O|D), rather than P(D|O)***]**

Compound Substitution Error

0.25 * 0.05 = 0.0125 **[***P(O)P(D|O) = P(O and D)***]**

Causal item: inverse order

There is some evidence that obesity causes Type 2 diabetes. About 25% of the United States adult population is obese. 8% of the US adult population have Type 2 diabetes. 5% of the population are obese and have Type 2 diabetes. A randomly selected adult has Type 2 diabetes. What the probability that this person is obese?

Correct solution:

P(O) = .25
P(Type2) = .08
P(O^Type2) = .05 = P(O|Type2)*P(Type2)
P(O|Type2) = .05/.08 = .625

Transposed Conditional Error

let P(OB) = the probability that someone is obese = 25%
P(T2)= the probability that someone has type 2 = 8%
P(OB and T2)= 5%
P(T2 | OB) = P (T2 and OB)/ P (OB)
P(T2|OB) = 5% /25%
= .2 **[***Should have solved for P(OB|T2)***]**

Compound Substitution Error

5% **[***5% is given as P(Obese and T2 Diabetes)***]**

'2D', 'T2', and 'D' = 'having diabetes'. 'O' and 'OB' = 'obese'.
Note: Explanatory comments are in brackets

Discussion

Order impacted the likelihood of formalizing conditional probability for the temporal items but not causal items. No significant differences were found in the likelihood of solving test items with cover stories depicting events of intact or inverted order for both causal and temporal items.

The proportion of students solving the items correctly was too low for meaningful analyses comparing solution on the temporal items specifically. Application of Bayes's rule is on the more difficult end of the spectrum of skills required to solve introductory probability problems (Bar-Hillel & Falk, 1982; Díaz & Fuente, 2007) and may have produced a floor effect, preventing the data from varying enough for analysis.

Results for Study 2, where significant, were not as originally expected. It was believed that a test item requiring a solution in which the temporal order of events is inverted would be more difficult than an item with events in intact temporal order. Analyses indicated that the odds ratio for correctly formalizing CP on an intact-order item versus an inverted-order item is 0.087. This result indicates that an individual is 0.1 times as likely to formalize CP correctly for an intact-order item than for an inverted item. In other words, the inverted-order items are 11.5 times as likely than intact-order items to have CP expressed correctly.

Although they were not included in GZLM analyses, examination of written protocols was useful to explain the direction of this result. The conditional probability expressed in the last line of the problem text is only one of three conditional probabilities defined in the temporal items, the other two being given. For items asking for a solution of the CP of intact-order events, the *given* CPs reflect inverted-order events. "Inverted-order" items, on the other hand,

included given CPs reflecting events in intact order. Since when solving any word problem an individual must translate the text of the problem into formal mathematical notation, the direction of given CPs may be of greater relevance to performance on the item than the direction of the CP to be solved (Pólya, 1954a, 2004).

Examining written protocols of the 19 in-person participants also revealed that nine subjects sketched a table or tree; two in the NV condition (in which the CPs given were intact) and seven in the NT condition (given CPs depicted events in inverted order). So, perhaps the increased difficulty of encoding events in inverted order encouraged participants to use graphical devices (see Zahner & Corter, 2010). Further, of the participants using graphical devices to encode the CP events in inverted order, five used trees and two used contingency tables. All but one encoded CP correctly; this participant used a table, which has been shown to obscure rather than facilitate working with CP (Díaz & Fuente, 2007; Gras & Totohasina, 1995).

These results indicated that, to investigate the effect of temporal order on a problem's difficulty, Study 2 hypotheses must be revised and the conditions of the study revisited with items classified according to the direction of given CPs rather than the direction of the CP to be solved. Significant reworking of items was required to examine the effect of order as expressed in our revised hypothesis as well as to more appropriately reflect the abilities of novice statistics students.

CHAPTER V: STUDY 3

Statement of the Problem

The proportion of students correctly identifying the temporal and causal items in Study 2 as asking to solve for a conditional was startlingly low. To control for this, all items were revised to make conditional probability within the problems more salient. Further, examining problem-solving procedures from Study 2 indicated that the order of given conditional probabilities (CP), rather than the direction of CPs to be solved, was more relevant. This is the case in items asking to solve for temporal or causal order intact. We expect that for these revised items it would be more difficult to encode CP when given CPs are expressed with order intact than when expressed with order inverted.

Two further issues arose from Study 2. The difficulty that novice statistics students have in solving problems requiring Bayes's rule introduced challenges outside of the scope of this research and may have obscured order effects. The pair of temporal problems for Study 3 used the same cover story as those in Study 2, and was edited so that participants were still required to encode CP, but solve for the probability of a simple event. Bayes's rule was not required for correct solution. The other issue encountered in Study 2 concerned the relative perceived strength of the causal relationship between the events depicted in the cover story for the causal problems in Study 2, as has been shown to affect problem solving (Krynski & Tenenbaum, 2007; Tversky & Kahneman, 1980; Windschitl & Wells, 1998). Specifically, it was thought that the causal relationship between obesity and Type 2 diabetes may not have enough schematic salience to induce order effects. For Study 3, cover stories were constructed to depict events with greater perceived causal strength to the population of interest, novice students of

probability and statistics. The cover stories for causal items in Study 3 depicted the events of studying for an exam and passing the exam. These items asked participants to calculate a conditional probability given probabilities for joint and compound events.

Method

Participants

Participants were all students enrolled in one of four introductory courses in probability and statistical inference during the Fall 2011 term at a graduate school of social sciences. Two instructors taught these courses; there were 54 students in one instructor's sections and 69 students in the other's. All four sections were taught in person.

Materials

Test items were presented to students as part of an in-class quiz. For one instructor's quizzes, the temporal item was presented first and the causal item third on a three-item quiz. The three items were presented on a single sheet of paper. In the other instructor's sections, the temporal item was presented fourth and the causal item second in a four-item quiz. Each item on this quiz was presented on a separate sheet of paper. Table 14 includes the test items by condition, with differences between pairs of items highlighted for clarity here only. Items presented to students were not formatted as such.

Table 14

Study 3: Test Items by Condition

TEMPORAL ITEMS

Order Intact (forward direction)

You are waiting to meet your friend, who is coming from work. He phones saying he will get on the next bus. From experience you know that if he catches an express bus his chances of being on time are 90%, but if he catches a local bus his chances of being on time are 65%. You also know that 60% of the buses that stop by his work are locals, thus he has a 60% chance of catching a local bus today. What is the probability that he arrives on time?

Order Inverted (backward direction)

You are waiting to meet your friend, who is coming from work. He phones saying he will get on the next bus. From experience you know that when he arrives on time, 90% of the time he has caught an express bus, but when he arrives late, 65% of the time he has caught an express bus. You also know from experience that he is late 60% of the time, thus you figure that he has a 60% chance of being late today. What is the probability that he catches an express bus?

CAUSAL ITEMS

Order Intact (forward direction)

At a journalism school, a professional ethics exam is given to all students at the end of their first year. Extensive research has established that the probability that a student studies specifically for this exam is 70%. The overall proportion of students who pass the exam is 92%. Exactly 66% of the students will study for the exam and pass it. If we know that a student has studied specifically for the exam, what is the probability that the student FAILS?

Order Inverted (backward direction)

At a journalism school, a professional ethics exam is given to all students at the end of their first year. Extensive research has established that the probability that a student studies specifically for this exam is 70%. The overall proportion of students who pass the exam is 92%. Exactly 66% of the students will study for the exam and pass it. If we know that a student has passed the exam, what is the probability that the student DID NOT study specifically for it?

Procedure

Both instructors gave the quizzes during the fourth week of the course, after covering the topic of conditional probability. Quizzes were administered during the last 30 minutes of each one hour, 40 minute session. Students were permitted to use notes and a calculator.

Course TAs scanned completed quizzes into files, removing all identifying information. There was variation by course TA in grading the quizzes so all items were scored and coded by an investigator. A subset of approximately 30% of participants' data was also coded by the same independent rater as in Study 2, using the same interrater reliability procedures.

Analyses

Variables of interest

The dependent variables (DV) of interest are the same as in Study 2, using the same coding criteria. Each item was scored to indicate whether a student had come to the correct mathematical solution and whether the student had encoded conditional probability correctly.

The primary independent variable (IV) of interest is order of given events, *intact-order* (NT) or *inverted-order* (NV). Instructor was also included as a covariate, since items were presented differently on each quiz and the two instructors differed significantly in years of teaching experience.

Interrater reliability

Interrater reliability was calculated using Cohen's kappa (*k*) for both problem solution and encoding CP.

Experimental analyses

Proportion of correct solution and encoding CP was calculated and are reported in Table 15. No statistical tests were applied to these simple proportions.

Binary logit generalized linear model analyses (GZLM) were performed to yield statistical tests of both overall model fit as well as parameter estimates for order effects, controlling for instructor. These analyses used the same methodology as described for Studies 1 and 2.

Error analyses

Errors in formally expressing CP were further coded into one of the common types of mistakes defined for Study 2. The items for Study 3 required different calculations than Study 2, so it was not expected that the same types of errors would necessarily be represented in these data.

Errors by condition were analyzed using GZLM, with one exception discussed below. Analyses were used to compare likelihood of making each type of error, controlling for instructor. Statistical analyses on common error types included only the subset of participants who coded CP incorrectly.

Results

Interrater reliability

Interrater reliability was acceptable for each DV in each set of items. For problem solution in temporal items, *kappa* = 1.00 ($p < .001$); for encoding CP, *kappa* = .56 ($p < .005$). For causal items, agreement in coding problem solution is $k = 1.00$ ($p < .001$); agreement in coding CP correct is $k = .49$ ($p < .012$).

Experimental analyses

Proportion correct by condition is summarized in Table 15. Overall, more participants coded CP correctly for each item type than solved the item correctly. Correlations among the two DVs for both types of item are included in Appendix C.

Table 15

Study 3: Performance on Items by Condition

Temporal items	**Intact**	**Inverted**	**Total**
n per condition (% correct)	*55 (44.7)*	*68 (55.3)*	*123*
Solving the problem	*32 (58.2)*	*35 (52.5)*	*67 (54.5)*
Coding conditional probability	*41 (74.5)*	*39 (57.4)*	*80 (65.0)*

Causal items	**Intact**	**Inverted**	**Total**
n per condition (% correct)	*67 (54.9)*	*55 (45.1)*	*122*
Solving the problem	*17 (25.4)*	*11 (20.0)*	*28 (23.0)*
Coding conditional probability	*43 (64.2)*	*19 (34.5)*	*62 (50.8)*

Results from generalized linear model analyses are summarized in Table 16. There were no significant differences between groups in proportion solving temporal or causal items correctly. For temporal items, the model using order and instructor to predict formalizing CP correctly fit significantly better than an intercept-only model, $X^2_{LR}(2, N = 123) = 9.16, p < .010$. Controlling for instructor effect, temporal order significantly predicted coding CP correctly, $X^2_{LR}(1, N = 123) = 6.02, p < .014$. Parameter estimates indicated that the odds of formalizing CP for order intact relative to order inverted is 2.67 ($p < .017$). The effect of instructor was not significant controlling for temporal order.

Table 16

Study 3: Results From Generalized Linear Model Analyses

Temporal items								
N = 123	**Odds**[a]	***B***	X^2_{LR}	***df***	***p***	**Model** X^2_{LR}	***df***	***p***
***DV* = Solving the problem**						*7.47*	*2*	*.126*
Temporal Order	*1.31*	*0.27*	*0.52*	*1*	*.472*			
Instructor	*0.48*	*-0.74*	*3.93*	*1*	*.048*			
***DV* = Coding conditional probability**						*9.16*	*2*	*.010*
Temporal Order	*2.66*	*0.98*	*6.02*	*1*	*.014*			
Instructor	*2.31*	*0.84*	*4.47*	*1*	*.034*			
Causal items								
N = 122	**Odds**	***B***	X^2_{LR}	***df***	***p***	**Model** X^2_{LR}	***df***	***p***
***DV* = Solving the problem**						*5.65*	*2*	*.059*
Causal Order	*1.63*	*0.49*	*1.21*	*1*	*.272*			
Instructor	*2.45*	*0.89*	*3.90*	*1*	*.048*			
***DV* = Coding conditional probability**						*17.08*	*2*	*.001*
Causal Order	*3.70*	*1.31*	*11.71*	*1*	*.001*			
Instructor	*2.17*	*0.78*	*3.98*	*1*	*.046*			

[a]*Reference group = inverted*

For causal items, the model including order and instructor predict expressing CP correctly also fit significantly better than an intercept-only model, X^2_{LR} (2, N = 122) = 17.08, p < .001. Controlling for instructor effect, causal order significantly predicted coding CP correctly, X^2_{LR} (1, N = 122) =11.71, p < .001. The odds of expressing CP correctly for order intact items is 3.70 (p < .001) relative to order inverted items. Controlling for causal order, the effect of instructor was not significant.

Error analyses

Types of errors by condition are summarized in Table 17. Simple probability substitution for conditional errors did not occur in Study 3, most likely since the problems were redesigned to require different calculation procedures. Statistical tests were conducted on observations with incorrect CP encoding only.

Table 17

Study 3: Distribution of Prototypical Encoding Errors by Condition

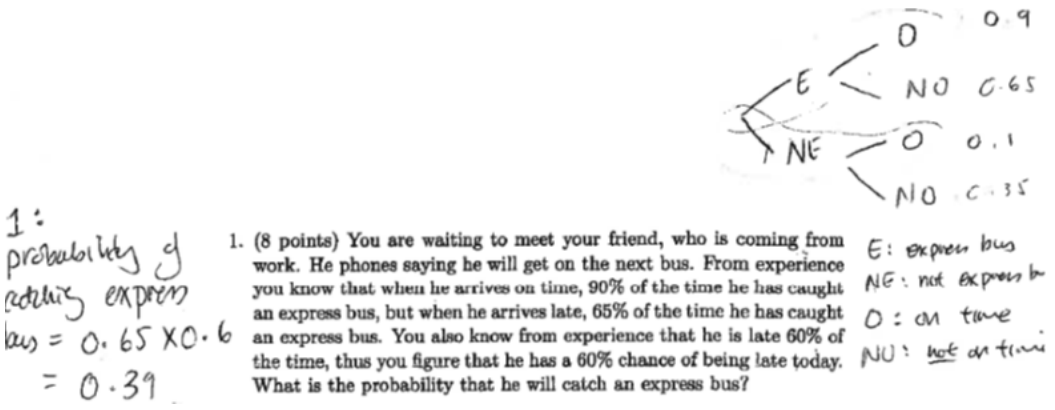
Temporal Items	**Intact**	**Inverted**	**Total Errors**
n errors per condition (error type)	*14 (32.6)*	*29 (67.4)*	*43 (35.0)*
Transposed Conditional	*0 (0)*	*15 (51.7)*	*15 (34.9)*
Compound Substitution	*1 (7.1))*	*2 (6.9)*	*3 (7.0)*
Unclassified	*13 (92.9)*	*12 (41.4)*	*25 (58.1)*
Causal Items	**Intact**	**Inverted**	**Total Errors**
n errors per condition (error type)	*23 (39.7)*	*35 (60.3)*	*58 (49.2)*
Transposed Conditional	*2 (8.7)*	*11 (31.4)*	*13 (22.4)*
Compound Substitution	*14 (60.9)*	*9 (25.7)*	*23 (39.7)*
Unclassified	*7 (30.4)*	*15 (42.9)*	*22 (37.9)*

No participants in the NT condition made the transposed conditional errors. GZLM could not be used to analyze these data as zero-cells yield complete separation in the data and unreliable estimates. Statistical significance of a chi-square test was therefore evaluated using Fisher's exact test, χ^2 (1, N = 43) = 11.12, p < .001.

Analyses on transposed conditional errors in causal items controlled for effect of instructor using GZLM. Although more transposed conditionals occurred in the causal inverted group, the difference was significant only at the p < .10 level, χ^2_{LR} (1 , N = 58) = 4.81, p < .057 . Students in the NT condition were significantly more likely to commit joint substitution errors, χ^2_{LR} (1 , N = 58) = 6.66, p < .010, with an odds ratio of 4.3. Figures 5 and 6 include examples of transposed conditional errors for both temporal and causal items.

Figure 5

Study 3: Example of Reverse Error in Temporal Inverse Item



Discussion

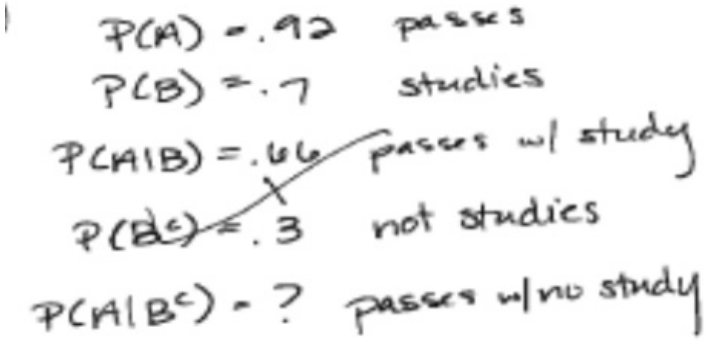
The rationale behind Study 3 was to demonstrate the effects of temporal and causal order on encoding conditional probability. Revising the hypotheses from Study 2 highlighted that order affected encoding CPs which were given in a probability problem, rather than the order of the CP to solve.

Temporal order significantly affected encoding CP. The odds ratio for encoding CP correctly in temporal items was 2.66, indicating that, controlling for instructor, a student receiving an NT item was 2.66 times as likely to encode conditional probability correctly than a student receiving an NV item. Further, of students who did not encode CP correctly for temporal items, only students receiving NV items transposed the order of events. Order effects were also demonstrated in the causal items.

Participants were more than three times as likely to encode CP correctly for NT items than for NV items, controlling for instructor. Error analyses using GZLM indicated that a

Figure 6

Study 3: Example of Reverse Error in Causal Inverse Item



marginally significant proportion of students made the transposed conditional error in the NV condition, controlling for instructor.

The results from these error analyses indicate that there may be schema-specific effects related to temporal and causal order. Recall that that the likelihood of committing a compound probability error in causal items was significantly more likely in the NT condition. Since types of errors coded were mutually exclusive, it makes sense that in absence of making a transposed conditional error, students erring in encoding CP may be more likely to make another systematic error. Compound substitution errors are fairly common when interpreting conditional probability, and these data do not necessarily indicate that mistaking compound for conditional probability is specifically affected by temporal or causal order. In other words, error analyses may demonstrate that transposed conditional errors are sensitive to order effect, while compound probability errors are *not* specific to order effect.

Study 3 findings support a phenomenon demonstrated by Tversky and Kahneman (1980). It was shown that, in cases in which $P(A) = P(B)$, and $P(A|B) = P(B|A)$, individuals were more likely to judge $P(B|A)$ as greater than $P(A|B)$ if they believed that A was a cause of B. The results of Study 3 demonstrated that, possibly as a result of the strength of causal direction, students exhibited a preference for expressing conditional probability preserving forward direction of cause to effect that was not demonstrated with respect to the events depicted in the Study 2 items.

CHAPTER VI: GENERAL DISCUSSION

Summary

Results from the three studies described in this paper support the idea that there are differential schematic effects on probability problem solving. Probability problems were written with cover stories depicting objects (e.g., randomizing devices) as well as social situations, keeping the formal problem structure and solution processes consistent between conditions. Systematically altering the schema depicted in a probability problem's cover story was shown to affect the difficulty of the item as well as the type of errors. This effect was demonstrated with respect to social schemas (Study 1) as well as temporal and causal schemas (Studies 2 and 3) for a variety of participants.

Specifically, Study 1 illustrated that when randomness is unexpected in a social schema, individuals were less likely to correctly solve a probability problem relative to problems in which randomness was appropriate to the schema depicted. An individual was more than one-and-a-half times as likely to correctly solve a problem in which randomness was appropriate to the problem schema. Our results did not support that it would be easier to solve a randomness-appropriate item if it depicted objects rather than people in a lottery-type situation. For situations in which random ordering was not appropriate, having an explanation for imposing randomness did not affect formal probability calculations.

Schematic effects were also shown with respect to temporal and causal direction in Studies 2 and 3. Results from Study 2 specified that, in solving problems requiring Bayes's formula, difficulty in encoding conditional probability affected an item's difficulty with respect to the given CPs, rather than the direction of the CP to be solved. The difficulty that novice

statistics students have in applying Bayes's formula was also supported in the data by the scarcity of correct solutions.

Study 3 was designed to revisit the hypotheses addressed in Study 2, with refinements to isolate schematic order effects. It was shown that for both temporally- and causally-related events, participants were more likely to encode given conditional probabilities incorrectly when the events are expressed in inverse order. In addition, those errors were more likely to transpose the events from inverse order to intact order than from intact to inverse order. For problems depicting causally-related events, participants were also more likely to incorrectly encode CP when solving for CP expressing events in inverse order.

This finding may be considered in light of the phenomenon of the *fallacy of the time axis*, as illustrated by Falk (1986). In a within-subjects study, participants were asked to consider the events of drawing two marbles from an urn containing two black and two white marbles. Asked first to evaluate the probability of drawing a second white marble after having drawn a first white marble without replacement, $P(W_2|W_1)$, most participants provide the correct answer with relative ease. Next, asking the same participants to consider $P(W_1|W_2)$, a significant proportion of participants reply that the question is meaningless. Of those who attempt to solve the problem, many indicate that the probability is 1.00 or incorrectly solve the item without considering the probability of the conditioning event. A similar phenomenon was also observed in the data from Studies 2 and 3, in that only participants asked to evaluate and solve for CP in inverted order answered the problem with extreme values of 1.00 or the given simple probability. For example, given event A which precedes event *B*, when asked the likelihood of *A* given *B*, only students in the inverted-order condition answered $P(A)$ rather than $P(A|B)$. In

Study 2, one participant in the temporal-item NV condition reasoned that "P(E) = 1 because you know he took an express bus".

Although the results from Studies 2 and 3 parallel Falk's (1986), only the current studies illustrated that the fallacy persists when encoding formal CP, typically considered a Type 2 process. There has been little evidence or explanation of relevant processes influencing the occurrence of the fallacy of the transposed conditional. Villejoubert and Mandel (2002) illustrated that while frequency formats reduced the number of transposed conditional errors, their only explanation offered is that "people simply confuse p(H|D) with p(D|H) because the latter *sounds* a lot like the former" (their emphasis).

Krynski and Tenenbaum (2007) speculated that transposed conditionals are more likely when $P(A|B)$ is estimated as roughly equivalent to $P(B|A)$, as defined by Tversky and Kahneman (1980). Krynski and Tenenbaum's interpretation overgeneralizes the Tversky and Kahneman finding; neither the 1980 data nor the present findings support that the transposed conditional occurs specifically in cases when $P(A|B)$ and $P(B|A)$ are roughly equivalent, but rather persists *despite* the two CPs being defined as equivalent. Results from Study 3 further illustrated that participants exhibited a significant bias in transposing CP to forward causal or temporal direction when those events in a problem were depicted in inverse temporal or causal order, but rarely from intact to inverse order.

Limitations

The current research was undertaken to examine schematic effects on probability problem solving in a general population, but with particular emphasis on training social scientists expected to produce and consume research. The pool of participants for Study 1 was

different from the participants in Studies 2 and 3 in a number of ways worth noting. Information about Study 1's participants' fields of study or profession was collected with other demographic items. No differences were shown affecting performance or assigned condition among groups defined by academic field, so it was not included in analyses. In addition, almost 80% of in-person volunteers for Study 2 indicated that they had never heard of the online platform used to recruit Study 1 participants. Anecdotal evidence may inform that volunteers on an online testing platform have more experience in technical fields and, by extension, more experience with training in mathematics and statistics. These participants also exhibited differential effects depending on the countries in which they were educated, although the variable used to group them by region did not specifically assess English proficiency and should not be interpreted as such.

The experimental data collected for all three studies were limited to performance on a numeric calculation; participants were not required to provide any interpretation of results or put their numeric answers in the problem context. This may have been where additional schematic effects would have been demonstrated. Further study into the additional conditions for Study 1 will likely require more complicated problems to illustrate the more nuanced effects typically attributed to Type 1 processes.

Analyses on problem-solving processes were possible to a degree on data collected for Studies 2 and 3 since the data included written protocols, but these were unavailable for about ten participants who used separate materials on which to work. Most of these participants produced correct solutions to the problems, but error analyses were not possible for the few who did not. Further, in those cases it was not possible to examine written work to inform whether their solutions included a tree or other graphical representation of the problem.

Implications

How do schematic effects of probability problem solving inform processes by which individuals make valid statistical inferences about human behavior? Allowing for the limitations of the studies presented in this paper, it is still apparent that stochastic influence on human behavior is not easily conceptualized. People more readily attribute outcomes to unobserved or unknown variables rather than stochasticity in most situations (Luhmann & Ahn, 2005). When a problem's situational or schematic content refers to human behavior, variation is regularly attributed to human intent (Cheng & Holyoak, 1985; Cheng, et al., 1993; Falk & Lann, 2008; Fong, et al., 1993; Nisbett, et al., 1993; Schwartz & Goldman, 1996).

Krynski & Tenenbaum (2007) showed improved performance on probability problems when the *conditions* assumed in conditional probabilities function as causal roles or explanations over those problems in which conditions are presented with no explanatory role. For example, the temporal items from Studies 2 and 3 might include an explanation that the friend is unlikely to catch an express bus because they are typically overcrowded. The friend's preference for local buses may then account for his 75% overall likelihood to arrive late. Such details have been shown to highlight problem elements which may otherwise be overlooked. In statistics instruction in the social sciences, providing explanatory roles for conditions affecting stochastic human behavior may improve both performance on a problem as well as its interpretability for students.

Parsing how conditional probability is interpreted may inform a source of bias produced by the order of events depicted in probability problems. If, in the cover story of a problem, event *A* occurs before event *B* or causes event *B*, the conditional probability $P(B|A)$ reflects a schema-

consistent, intact order of events, translated as "the probability that *B* occurs given that *A* has occurred," reflecting a deterministic, forward-looking time perspective. In contrast, translating $P(A|B)$ as "the probability that *A occurs* given that *B* has occurred" is nonsensical given the problem script. Representing $P(A|B)$ demands consideration of events in inverted chronological order and may be validly translated as only "the probability that *A has occurred* given that *B* has occurred," with retrospective time perspective.

These two representations of conditional probability reflect two different types of reasoning about uncertainty (Hacking, 1975; Fox & Ülkümen, 2011). While $P(B|A)$ may be addressed in the predictive, aleatory sense, $P(A|B)$ must be considered with epistemic evaluation postdictively.

Reasoning from signs to sources has an extensive history distinct from deductive reasoning from causes to effects. The historical and philosophical contexts of reasoning about an event's likelihood versus estimating one's confidence in an assertion has been thoroughly discussed as the *aleatory-epistemic* distinction by Hacking (1975), but the dichotomy has also been explicitly addressed as *probable-plausible* with respect to mathematical induction (Pólya, 1941, 1954b), *predictive-diagnostic* in the heuristics and biases debate (Cohen, 1981; Mackie, 1981), as well as *objective-subjective* in literature on decision making (Fox & Ülkümen, 2011).

Schematic effects on probability problem solving may inform instruction on statistical inference by reflecting on how the two dimensions of probability judgments are theoretically distinct. Evidence from Study 3 supports an interpretation which addresses this difference in reasoning as parallel to the process of statistical inference and has implications for statistical

education. In evaluating evidence from a sample, the process of hypothesis testing requires considering the conditional probability of finding a result in light of the null hypothesis H_0, a preexisting fact in the world, which *in fact* either is, $P(H_0) = 1$, or is not, $P(H_0) = 0$. So while results from Studies 2 and 3 indicate that it is easier to conceptualize $P($ *some observed phenomenon* $\mid H_0)$, hypothesis testing evaluates $P(H_0 \mid$ *some observed phenomenon*$)$. In addition, in evaluating the validity of statistical inferences, students are taught to consider the likelihood of *Type I* and *Type II* errors, terms which have become shorthand for particular conditional probabilities. The convention in statistics education of not representing Type I and Type II errors as explicitly conditional probabilities has been criticized for oversimplifying students' concept of statistical inference and leads to inappropriate levels of confidence in results (Neath, 2010). Training in statistical inference demands a level of mastery of understanding CP in inverted order, which is inconsistent with causal schemas and may warrant more classroom discussion with respect to the nuances and implications for hypothesis testing.

In the case of social scientific research, quantitative evidence supporting claims about human behavior relies on the convention of hypothesis testing, which requires a degree of sophistication in understanding probability and likelihood. Judgments about social phenomena are mediated by the interaction of how one interprets human behavior through personal experience while accounting for patterns which may not be apparent except in aggregate. The combination of the salience of personal experience with the relative opacity of statistical inference regularly leads to discounting or misinterpreting empirical evidence when making high-stakes educational or policy decisions (Epstein, 1986; Milton, 2006; Paris & Luo, 2010; Slavin, 2004).

Heuristics and biases in evaluating quantitative information, including those addressed in this paper, have been shown in both laypeople as well as persist in individuals with extensive graduate training in probability and statistics. Conditions and assumptions which are relevant in applying statistical analyses to human behavior are regularly violated or overlooked, in part, by these biases, and impede nuanced evaluation of empirical results. Statistical training may explicitly address the demonstrated schematic effects, so that social scientists' are better able to contextualize statistical evidence having real-world consequence.

Future Directions

Further investigation into schematic effects on probability problem solving will incorporate additional conditions which have been demonstrated in the literature. The robust effects facilitating problem-solving success when probabilities are expressed in frequency formats (Gigerenzer & Hoffrage, 1995, 2007; Krynski & Tenenbaum, 2007) would be expected to affect problem solving on the items designed for this paper. The salience of a causal relationship was shown to possibly mediate order effects with respect to coding conditional probability in Studies 2 and 3, and variants of causal items in future studies will be written to address this factor more explicitly, in the model of Krynski and Tenenbaum (2007).

Interview data may contribute particularly relevant details about the probability problem-solving process. A few sessions with students instructed to "think aloud" while solving problems may add to a future research program. Analyses on interview data should inform critical points as sources of common misconceptions and biases in problem-solving performance.

REFERENCES

Agnoli, F. & Krantz, D. H. (1989). Suppressing natural heuristics by formal instruction: The case of the conjunction fallacy. *Cognitive Psychology, 21*, 515-550.

Anderson, J. R. & Thompson, R. (1989). Use of analogy in a production system architecture. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 267-297). New York: Cambridge University Press.

Bar-Hillel, M. & Falk, R. (1982). Some teasers concerning conditional probabilities. *Cognition, 11*, 109-122.

Barbey, A. & Sloman, S. (2007). Base-rate respect: From ecological rationality to dual processes. *Behavioral and Brain Sciences, 30*, 241.

Bassok, M., Wu, L.-L., & Olseth, K. L. (1995). Judging a book by its cover: Interpretative effects of content on problem-solving transfer. *Memory & Cognition, 23*, 354-367.

Batanero, C., Henry, M., & Parzysz, B. (2005). The nature of chance and probability. In G. A. Jones (Ed.), *Mathematics education library: Vol 40. Exploring probability in school* (pp. 15-37). New York: Springer.

Bower, G. H., Black, J. B., & Turner, T. J. (1979). Scripts in memory for text. *Cognitive psychology, 11*, 177-220.

Brun, W. & Teigen, K. H. (1990). Prediction and postdiction preferences in guessing. *Journal of Behavioral Decision Making, 3*, 17-28.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk. *Perspectives on Psychological Science*, 6, 3-5.

Cheng, P. W. & Holyoak, K. J. (1985). Pragmatic reasoning schemas. *Cognitive Psychology, 17*, 391-416.

Cheng, P. W., Holyoak, K. J., Nisbett, R. E., & Oliver, L. M. (1993). Pragmatic versus syntactic approaches to training deductive reasoning. In R. E. Nisbett (Ed.), *Rules for reasoning* (pp. 165-203). Hillsdale, NJ: Erlbaum.

Cheng, P. W. & Nisbett, R. E. (1993). Pragmatic constraints on causal deduction. In R. E. Nisbett (Ed.), *Rules for reasoning* (pp. 207-227). Hillsdale, NJ: Erlbaum.

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.

Chi, M. T. H., Glaser, R., & Rees, E. (1982). Expertise in problem solving. In R. J. Sternberg (Ed.), *Advances in the psychology of human intelligence* (Vol. 1, pp. 7-75). Hillsdale, NJ: Erlbaum.

Cohen, L. J. (1981). Can human irrationality be experimentally demonstrated? *Behavioral and Brain Sciences*, *4*, 317-370.

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

Diaconis, P. & Freedman, D. (1981). The persistence of cognitive illusions. *Behavioral and Brain Sciences*, 4, 333-334.

Díaz, C. & Fuente, I. de la (2007). Assessing students' difficulties with conditional probability and Bayesian reasoning. *International Electronic Journal of Mathematics Education*, *2*, 128-148.

Epstein, S. (1986). Does aggregation produce spuriously high estimates of behavior stability? *Journal of Personality and Social Psychology*, *50*, 1199-1210.

Epstein, S., Pacini, R., Denes-Raj, V., & Heier, H. (1996). Individual differences in intuitive–experiential and analytical–rational thinking styles. *Journal of Personality and Social Psychology*, *71*, 390-405.

Evans, J. S. B. T. & Frankish, K. (Eds.). (2008). *In two minds: Dual processes and beyond*. New York: Oxford University Press.

Falk, R. (1986). Of probabilistic knights and knaves. *The College Mathematics Journal*, *17*, 156-164.

Falk, R. (1989). Judgment of coincidences: Mine versus yours. *The American Journal of Psychology*, *102*, 477-493.

Falk, R. & Lann, A. (2008). The allure of equality: Uniformity in probabilistic and statistical judgment. *Cognitive Psychology*, *57*, 293-334.

Fantino, E. & Stolarz, F. S. (2007). Enhancing sensitivity to base-rates: Natural frequencies are not enough. *Behavioral and Brain Sciences*, *30*, 262.

Fischhoff, B. (1975). Hindsight is not equal to foresight: The effect of outcome knowledge on judgment under uncertainty. *Journal of Experimental Psychology: Human Perception and Performance*, *1*, 288-299.

Fischhoff, B. (1976). The effect of temporal setting on likelihood estimates. *Organizational Behavior and Human Performance*, *15*, 180-194.

Fong, G. T., Krantz, D. H., & Nisbett, R. E. (1993). The effects of statistical training on thinking about everyday problems. In R. E. Nisbett (Ed.), *Rules for reasoning* (pp. 91-135). Hillsdale, NJ: Erlbaum.

Fox, C. R. & Levav, J. (2004). Partition-edit-count: Naive extensional reasoning in judgment of conditional probability. *Journal of Experimental Psychology: General*, *133*, 626-642.

Fox, C. R. & Ülkümen, G. (2011). Distinguishing two dimensions of uncertainty. In W. Brun, G. Keren, G. Kirkebřen, & H. Montgomery (Eds.), *Perspectives on thinking, judging and decision-making* (pp. 21-35). Oslo: Universitetsforlaget.

Gal, I. (2005). Towards "Probability Literacy" for all citizens: Building blocks and instructional dilemmas. In G. A. Jones (Ed.), *Mathematics education library: Vol 40. Exploring probability in school* (pp. 39-63). New York: Springer.

Gick, M. L. & Holyoak, K. J. (1980). Analogical problem solving. *Cognitive Psychology*, *12*, 306-355.

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

Gigerenzer, G. & Hoffrage, U. (2007). The role of representation in Bayesian reasoning: Correcting common misconceptions. *Behavioral and Brain Sciences*, *30*, 264.

Girotto, V. & Gonzalez, M. (2007). How to elicit sound probabilistic reasoning: Beyond word problems. *Behavioral and Brain Sciences*, *30*, 268.

Gras, R. & Totohasina, A. (1995). Conceptions d'élèves sur la notion de probabilité conditionnelle révélées par une méthode d'analyse des donn ées: Implication - similarité - corrélation. *Educational Studies in Mathematics*, *28*, 337-363.

Heath, C. & Tversky, A. (1991). Preference and belief - ambiguity and competence in choice under uncertainty. *Journal of Risk and Uncertainty*, *4*, 5-28.

Henson, R. K., Hull, D. M., & Williams, C. S. (2010). Methodology in our education research culture. *Educational Researcher*, *39*, 229-240.

Hinsley, D. A., Hayes, J. R., & Simon, H. A. (1977). From words to equations: Meaning and representation in algebra word problems. In M. A. Just & P. A. Carpenter (Eds.), *Cognitive processes in comprehension* (Vol. 329). Hillsdale, NJ: Erlbaum.

Holyoak, K. J. & Thagard, P. R. (1989). A computational model of analogical problem solving. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 242-266). New York: Cambridge University Press.

Howell, W. C. & Burnett, S. A. (1978). Uncertainty measurement: A cognitive taxonomy. *Organizational Behavior and Human Performance*, *22*, 45-68.

Kahneman, D. & Tversky, A. (1979). On the interpretation of intuitive probability: A reply to Jonathan Cohen. *Cognition*, *7*, 409-411.

Krynski, T. R. & Tenenbaum, J. B. (2007). The role of causality in judgment under uncertainty. *Journal of Experimental Psychology: General*, *136*, 430-450.

Larkin, J. H., Heller, J. I., & Greeno, J. G. (1980). Instructional implications of research on problem solving. *New Directions for Teaching and Learning, 1980*, 51-65.

Larkin, J. H., McDermott, J., Simon, D. P., & Simon, H. A. (1980). Models of competence in solving physics problems. *Cognitive Science, 4*, 317-345.

Larkin, J. H. & Simon, H. A. (1987). Why a diagram is (sometimes) worth ten thousand words. *Cognitive Science, 11*, 65-100.

Lehman, D. R., Lempert, R. O., & Nisbett, R. E. (1993). The effects of graduate training on reasoning: Formal discipline and thinking about everyday life events. In R. E. Nisbett (Ed.), *Rules for reasoning* (pp. 315-339). Hillsdale, NJ: Erlbaum.

Luhmann, C. C. & Ahn, W.-k. (2005). The meaning and computation of causal power: Comment on Cheng (1997) and Novick and Cheng (2004). *Psychological Review, 112*, 685-692.

Mackie, J. L. (1981). Propensity, evidence, and diagnosis. *Behavioral and Brain Sciences, 4*, 345-346.

Marshall, S. P. (1993). Assessing schema knowledge. In N. Frederiksen, R. J. Mislevy, & I. I. Bejar (Eds.), *Test theory for a new generation of tests* (pp. 155-180). Hillsdale, NJ: Erlbaum.

Martin, S. & Bassok, M. (2005). Effects of semantic cues on mathematical modeling: Evidence from word-problem solving and equation construction tasks. *Memory & Cognition, 33*, 471-478.

Mason, W. & Suri, S. (2012). Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods, 44*, 1-23.

Milton, P. (2006). Opening minds to change the role of research in education. *Education Canada, 47*, 39.

Neath, A. A. (2010). *Statistical Inference, Statistics Education, and the Fallacy of the Transposed Conditional.* Paper presented at the Joint Statistical Meetings, Vancouver, British Columbia.

Newell, A. & Simon, H. A. (1972). *Human problem solving.* Englewood Cliffs, NJ: Prentice-Hall.

Nisbett, R. E. (1993). *Rules for reasoning.* Hillsdale, NJ: Erlbaum.

Nisbett, R. E., Fong, G. T., Lehman, D. R., & Cheng, P. W. (1987). Teaching reasoning. *Science, 238*, 625-631.

Nisbett, R. E., Krantz, D. H., Jepson, C., & Kunda, Z. (1993). The use of statistical heuristics in everyday inductive reasoning. In R. E. Nisbett (Ed.), *Rules for reasoning* (pp. 15-54). Hillsdale, NJ: Erlbaum.

Olani, A., Harskamp, E., Hoekstra, R., & van der Werf, G. (2010). The roles of self-efficacy and perceived teacher support in the acquisition of statistical reasoning abilities: A path analysis. *Educational Research and Evaluation*, *16*, 517-528.

Pallas, A., M. (2001). Preparing education doctoral students for epistemological diversity. *Educational Researcher*, *30*(3), 6-11.

Paris, S. G. & Luo, S. W. (2010). Confounded statistical analyses hinder interpretation of the NELP report. *Educational Researcher*, *39*, 316-322.

Pólya, G. (1941). Heuristic reasoning and the theory of probability. *The American Mathematical Monthly*, *48*(7), 450-465.

Pólya, G. (1954a). Mathematics and plausible reasoning: Induction and analogy in mathematics. Princeton, NJ: Princeton University Press.

Pólya, G. (1954b). Mathematics and plausible reasoning: Patterns of plausible inference. Princeton, NJ: Princeton University Press.

Pólya, G. (2004). How to solve it: A new aspect of mathematical method. Princeton, NJ: Princeton University Press.

Reusser, K. (1988). Problem-solving beyond the logic of things - Contextual effects on understanding and solving word-problems. *Instructional Science*, *17*, 309-338.

Rips, L. J. (1994). The psychology of proof : deductive reasoning in human thinking. Cambridge, MA: MIT Press.

Ross, B. H. (1984). Remindings and their effects in learning a cognitive skill. *Cognitive psychology*, *16*, 371-416.

Ross, B. H. (1989). Remindings in learning and instruction. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 438-469). New York: Cambridge University Press.

Rothbart, M. & Snyder, M. (1970). Confidence in the prediction and postdiction of an uncertain outcome. *Canadian Journal of Behavioural Science/Revue canadienne des sciences du comportement*, *2*, 38-43.

Schank, R. & Abelson, R. (1977). *Scripts, plans, goals, and understanding: An inquiry into human knowledge structures*. Hillsdale, NJ: Erlbaum.

Schurr, A. & Erev, I. (2007). The effect of base rate, careful analysis, and the distinction between decisions from experience and from description. *Behavioral and Brain Sciences*, *30*, 281.

Schwartz, D. L. & Goldman, S. R. (1996). Why people are not like marbles in an urn: An effect of context on statistical reasoning. *Applied Cognitive Psychology*, *10*, 99-112

Slavin, R. E. (2003). A reader's guide to scientifically based research. *Educational Leadership, 60*(5), 12.

Slavin, R. E. (2004). Education research can and must address "What works" questions. *Educational Researcher, 33*(1), 27-28.

Slavin, R. E. (2008). Perspectives on Evidence-based research in education—what works? Issues in synthesizing educational program evaluations. *Educational Researcher, 37*, 5-14.

Sloane, F. C. (2008). Randomized trials in mathematics education: Recalibrating the proposed high watermark. *Educational Researcher, 37*, 624-630.

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin, 119*, 3-22.

Sloman, S. A. & Rips, L. J. (1998). *Similarity and symbols in human thinking.* Cambridge, MA: MIT Press.

Smith, E. R. & Collins, E. C. (2009). Dual-process models: A social psychological perspective. In J. S. B. T. Evans & K. Frankish (Eds.), *In two minds: Dual processes and beyond* (pp. 197-216). New York: Oxford University Press.

Stanovich, K. E., Toplak, M. E., & West, R. F. (2008). The development of rational thought: A taxonomy of heuristics and biases. In R. V. Kail (Ed.), *Advances in child development and behavior: Vol 36* (pp. 251-285). San Diego: Elsevier.

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science, 12*, 257-285.

Teigen, K. H. & Keren, G. (2007). Waiting for the bus: When base-rates refuse to be neglected. *Cognition, 103*, 337-357.

Tversky, A. & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*, 1124-1131.

Tversky, A. & Kahneman, D. (1980). Causal schemas in judgments under uncertainty. In M. Fishbein (Ed.), *Progress in social psychology.* (pp. 49-72). Hillsdale, N J: Erlbaum.

Tversky, A. & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review, 90*, 293-315.

Villejoubert, G. & Mandel, D. (2002). The inverse fallacy: An account of deviations from Bayes's theorem and the additivity principle. *Memory & Cognition, 30*, 171-178.

Vosniadou, S. (1989). Analogical reasoning as a mechanism in knowledge acquisition: A developmental perspective. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 413-437). New York: Cambridge University Press.

Wason, P. C. (1968). Reasoning about a rule. *Quarterly Journal of Experimental Psychology, 20*, 273-281.

Windschitl, P. D. & Krizan, Z. (2005). Contingent approaches to making likelihood judgments about polychotomous cases: the influence of task factors. *Journal of Behavioral Decision Making, 18*, 281-303.

Windschitl, P. D. & Wells, G. L. (1998). The alternative-outcomes effect. *Journal of Personality and Social Psychology, 75*, 1411-1423.

Wright, G. (1982). Changes in the realism and distribution of probability assessments as a function of question type. *Acta Psychologica, 52*, 165-174.

Wright, J. C. & Murphy, G. L. (1984). The utility of theories in intuitive statistics: The robustness of theory-based judgments. *Journal of Experimental Psychology: General, 113*, 301-322.

Zahner, D. & Corter, J. (2010). The process of probability problem solving: Use of external visual representations. *Mathematical Thinking and Learning, 12*, 177-204.

APPENDIX A

Table A1: Study 1 Participants' Self-Reported Countries of Education

One of the demographic items asked in which country or countries participants were educated. Each participant was assigned to one of three following categories for the variable *Region* based on the response to this question.

Table A1

Study 1: Region Codes by Participants' Self-Reported Countries of Education

Region	***N***	**Countries of Education[a]**
English-speaking countries	*275*	United States, Canada, United Kingdom, Ireland, Australia, Jamaica, and New Zealand
Asia	*73*	India, Singapore, Bahrain, China, Japan, Iran, Mongolia, Nepal, Pakistan, South Korea, Sri Lanka, and Thailand
Europe, Latin America, Non-English-speaking countries	*46*	Spain, Chile, France, Romania, Russia, Austria, Brazil, Germany, Serbia, Ukraine, Venezuela, Bosnia, Brazil, Bulgaria, Costa Rica, Croatia, Czech Republic, Greece, Israel, Italy, Macedonia, Mexico, Poland, Portugal, Sweden, Switzerland, The Netherlands

[a]Countries listed in order of most to least frequent

Figure A1: Probability Formulas Provided to Participants

Probability Formulas

$0 \leq P(e_i) \leq 1$

$\sum_{i=1}^{n} P(e_i) = 1$

$P(A) = \sum_{i=1}^{k} P(a_i)$

$P(S) = 1,\ P(\varnothing) = 0$

$0 \leq P(A) \leq 1$

$P(A^C) = 1 - P(A)$

$P(A) + P(A^C) = 1$

$P(A \cup B) = P(A \text{ or } B) = P(A) + P(B) - P(A \cap B)$

$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$ $\qquad P(A \cap B) = P(B) \cdot P(A \mid B)$

$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$

Combinatorics formulas:

$P^n = n! = n(n-1) \ldots (2)(1)$

$P_k^n = \frac{n!}{(n-k)!} = \frac{n(n-1)..(k+1)}{(n-k)(n-k-1)..(2)(1)}$

$C_k^n = \frac{n!}{(n-k)!\,k!} = \frac{n(n-1)..(n-k+1)}{k!} = \frac{n(n-1) \ldots (n-k+1)}{(k)(k-1) \ldots (2)(1)}$

Table A2: Detailed Regional Outcomes and descriptives

Table A2

Study 1: Descriptive Statistics by Region

Region		Correct solution	Time on task (minutes)	Age (years)
	N (% total)	*N (% region)*	*M (SD)*	*M (SD)*
United States	*248 (62.9)*	*151 (60.9)*	*3.28 (2.3)*	*29.1 (9.9)*
India	*58 (14.7)*	*20 (34.5)*	*3.56 (2.7)*	*26.9 (6.6)*
Europe, not English-speaking	*36 (9.1)*	*28 (77.8)*	*4.84 (2.5)*	*30.2 (9.2)*
English-speaking countries outside North America[a]	*17 (4.3)*	*11 (64.7)*	*3.35 (2.2)*	*25.2 (5.3)*
Asia, not India	*14 (3.6)*	*9 (64.3)*	*2.54 (1.7)*	*26.5 (8.4)*
Canada	*10 (2.5)*	*5 (50.0)*	*2.77 (1.5)*	*24.4 (4.6)*
Mexico and South America	*10 (2.5)*	*9 (90.0)*	*4.71 (1.9)*	*27.1 (6.0)*
Africa	*1 (0.3)*	*1 (100)*	*5.38 (0)*	*23 (0)*
Total	*394*	*234 (59.4)*	*3.47 (2.4)*	*28.4 (9.1)*

[a] *Australia, Ireland, Jamaica, New Zealand, United Kingdom*

Table A3: Sensitivity Analysis

Table A3

Study 1: Sensitivity Analyses Using Generalized Linear Models: Effect by Region

ST1: Random appropriate versus Random inappropriate

Region	*N*	Odds	*B*	X^2_{LR}	*df*	*p*
Asia	73	*3.06*	*1.12*	*5.28*	*1*	*.022*
Europe and Latin America	46	*13.14*	*2.58*	*8.32*	*1*	*.004*
English-speaking Countries	275	*1.19*	*0.17*	*0.48*	*1*	*.489*

APPENDIX B

Table B1: Study 2 Correlation Among Dependent Variables

Table B1

Study 2: Intercorrelations Among Dependent Variables by Cohort

In-Person (*N* = 19)

	Temporal solution	Causal solution	Temporal encoding	Causal encoding
Temporal solution	-	.344	.309	.286
Causal solution		-	.268	.899**
Temporal encoding			-	.328
Causal encoding				-

Online (*N* = 59)

	Temporal solution	Causal solution	Temporal encoding	Causal encoding
Temporal solution	-	.247	.376**	.295**
Causal solution		-	.105	.705**
Temporal encoding			-	.135
Causal encoding				-

** *Statistically significant at* p < *.001*

APPENDIX C

Table C1: Study 3 Correlation Among Dependent Variables

Table C1

Study 3: Intercorrelations Among Dependent Variables

N = 122	Temporal solution	Causal solution	Temporal encoding	Causal encoding
Temporal solution	-	.278	.565	.316
Causal solution		-	.185	.481
Temporal encoding			-	.140
Causal encoding				-

***Statistically significant at* p < *.002*