ABSTRACT

Title of thesis: CAN IRRELEVANT ALTERNATIVES AFFECT PROBABILITY JUDGMENT? THE DISCRIMINATION BIAS

Amber Sprenger, Master of Science, 2005

Thesis directed by: Professor Michael Dougherty
Department of Psychology

In this paper we used a proactive interference (PI) paradigm to examine the effect of generating irrelevant alternative hypotheses on probability judgments. Two possible effects of generating irrelevant alternative hypotheses on probability judgment were tested: discrimination failure and inhibition failure. The discrimination failure account predicted that participants would fail to identify irrelevant alternatives as irrelevant, causing them to include irrelevant alternatives in their judgments. Then, the magnitude and relative accuracy of participants' probability judgments would decrease as PI increases. The inhibition failure account predicted that participants would identify irrelevant alternatives as irrelevant, but would fail to inhibit them from working memory. Then, the magnitude of participants' probability judgments would increase as PI increases, but that the relative accuracy of the probability judgments would be unaffected by the build-up of PI. Three experiments support the discrimination failure account of the effect of PI on probability judgment.

CAN IRRELEVANT ALTERNATIVES AFFECT PROBABILITY JUDGMENT? THE DISCRIMINATION BIAS

by

Amber Sprenger

Thesis submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Master of Science
2005

Advisory Committee:

Professor Michael Dougherty, Chair
Professor David Huber
Professor Thomas Wallsten

ACKNOWLEDGMENTS

I thank Michael Dougherty, David Huber and Thomas Wallsten for their time and helpful feedback.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

Most theories of probability judgment posit that people compare a focal hypothesis with at least one alternative hypothesis (Tversky & Koehler, 1994; Dougherty, Gettys, & Ogden, 1999; Windschitl & Wells, 1998). For instance, Tversky and Koehler proposed that when judging the likelihood of a given hypothesis, people compare the support for a focal hypothesis with the support for alternative hypotheses. Dougherty et al.'s Minerva-DM model of probability judgment assumed that people compare the memory activation for the focal hypothesis with the memory activation for all explicitly considered alternative hypotheses. Windschitl and Wells proposed that people compare the support for a focal hypothesis with the support for the strongest (most likely) alternative hypothesis. Although slightly different in form, all three theories assume that the to-be-judged hypothesis is compared to a set of alternative hypotheses.

Prescriptively, accurate probability judgments require that people generate and include relevant alternative hypotheses in the comparison process. However, the ability to exclude irrelevant alternatives is also required for accurate probability judgment. Relevant alternative hypotheses can be defined as hypotheses that have some probability of occurring. Irrelevant hypotheses, on the other hand, are hypotheses that should not be included in the probability judgment because they have no possibility of occurring in the context of interest. Excluding irrelevant hypotheses may, in fact, be quite difficult when the set of relevant alternatives shares either surface-level or deep-level features with potential irrelevant alternative hypotheses, or when a previously relevant alternative is deemed inconsistent with available data. For example, it makes no sense for a physician to continue to consider an alternative disease hypothesis that has been eliminated on the basis of a blood test. However, if the patients' symptomology, other than the blood test result, resembles the disconfirmed diagnosis, associative memory processes alone may prevent the physician from being able to completely ignore the disconfirmed diagnosis.

The present research examined the extent to which generating irrelevant alternative hypotheses affects the accuracy of probability judgments. Our research addressed two general questions. First, can judgments of probability be affected by interference from irrelevant information? Prescriptively, when one estimates the probability of a particular event, one's judgment should incorporate only judgment-relevant information. However one could generate a judgment-irrelevant alternative when making a probability judgment. If one fails to discriminate the irrelevant alternative as irrelevant or if one discriminates the irrelevant alternative but fails to inhibit it, it can bias the probability judgment. Second, what cognitive processes underlie how people make probability judgments when they must discriminate between judgment-relevant and judgment-irrelevant information? We were interested in the extent to which people are able to discriminate and/or suppress irrelevant information from influencing judgment when it is retrieved.

The effect of irrelevant alternatives on probability judgments was examined using the proactive interference (PI) paradigm. In the PI paradigm, people learn and recall several lists of different words which are related by category membership, the buildup of PI phase. Typically, memory recall decreases as PI increases. (Underwood, 1957, 1945; Postman & Keppel, 1977; Wickens, Born, & Allen, 1963; Wickens, 1970). Then, participants learn and recall a list of words from a different category, the release from PI phase. Typically, at release from PI memory recall increases back to levels of recall before PI was introduced (Wickens, Born, & Allen, 1963). The finding of release from PI when a new category of information is learned and recalled reveals that the effects of PI are not due to a general inability to store and learn multiple sets of information or due to fatigue, but rather that PI is due to interference from similar irrelevant previous information.

The PI paradigm was chosen to test the effect of irrelevant alternatives on judgments for two reasons: First, it is well established that within the PI paradigm irrelevant information (words from prior lists) interferes with the retrieval of relevant information (words from the current list). Second, the PI paradigm provides a well-specified context for testing whether irrelevant alternatives can affect judgment because the separation between relevant and irrelevant information is easily conceptualized. When making probability judgments, participants should generate and consider alternatives from the relevant study trial (i.e. the current list) but ignore alternatives from the irrelevant study trials (i.e. lists studied prior the current list). To the extent that participants are able to accurately discriminate between relevant list and irrelevant list information and consider only the relevant information, their judgments should be relatively accurate. However, to the extent that participants cannot completely discriminate between the relevant and irrelevant lists of information, their judgments should be biased.

1.1 Effect of PI on Probability Judgments

One can conceptualize the impact of irrelevant information on probability judgment within the framework of Support Theory. Tversky and Koehler (1994) proposed that to make probability judgments, people compare the support for a focal hypothesis with the support for a set of alternative hypotheses:

$$P(A,B) = \frac{s(A)}{s(A)+s(B)} \tag{1.1}$$

where A refers to the focal hypothesis; B refers to the set of alternative hypotheses; s(A) and s(B) represent the support for A and B respectively; and P(A,B) represents the probability of hypothesis A versus an alternative hypothesis in set B occurring. Consider the situation in which one estimates the likelihood that basketball team A will win its division championship. To make an accurate probability judgment, one should retrieve and consider all alternative hypotheses contained in the residual hypothesis B, where B represents an implicit disjunction of other teams in the division. But in addition, one should exclude those alternative hypotheses that are similar but irrelevant to the given sample space. When judging the likelihood of Team A winning the basketball division title, for instance, relevant alternative hypotheses include all other teams in the division. Irrelevant hypotheses include any teams not in the given division. If irrelevant alternative hypotheses are retrieved, they may bias probability judgment.

To illustrate how PI might affect judgment, consider the scenario used in Experiment 1 in which participants imagined making repeated trips to the grocery store to purchase products for an upcoming party. On the first three trips to the store participants imagined purchasing various amounts (between 2 and 16) of produce items, such as bananas, broccoli, and apples. A specific kind of item (such as bananas) was purchased on only one trip to the store, and was not purchased on any other trip to the store. The first three trips to the store were considered the build-up of PI phase because produce items from previous trips could potentially interfere with the retrieval of produce items bought on the current shopping trip. On the fourth and final trip to the grocery store, participants imagined buying various amounts of beverage items, such as milk, water, and cola. This trip was considered the release from PI trip because items from previous trips were dissimilar to the items purchased on the current trip, and thus should not interfere with the retrieval of current-trip items. After each trip, participants judged the proportion of each kind of item (bananas, broccoli, apples) in their shopping bag. Participants also judged the proportion of four items that were not in the current shopping bag, but that had been bought on the previous trip to the grocery store. These four judgments measured participants' ability to discriminate items bought on the current trip from items bought on previous trips to the grocery store. Any nonzero judgments of these previous-trip items suggested a failure to discriminate that the items had been purchased on the previous shopping trip, and were not in the current shopping bag.

If participants retrieved an irrelevant, previous list produce item when making the proportion judgment about the current list, several possible outcomes could occur. First, the irrelevant hypothesis could be identified as irrelevant and inhibited so that one could continue to generate further hypotheses for inclusion in the comparison. Second, the irrelevant hypothesis could fail to be identified as irrelevant, and consequently be included in the probability judgment. I refer to the failure to identify an alternative hypothesis as irrelevant as discrimination failure. Third, one could identify the hypothesis as irrelevant but be unable to inhibit the hypothesis. Then, the irrelevant alternative could occupy space in working memory (WM) otherwise used for further hypothesis generation. I refer to the failure to inhibit an irrelevant alternative hypothesis as inhibition failure. Discrimination failure or inhibition failure should have distinct effects on probability judgment, and both could result in biased probability judgments. In the next section, the three possible effects of PI on probability judgment are described in more detail.

1.1.1 Optimal Case – No Bias

Consider the optimal case in which cognitive limitations played no role. People would then be able to both discriminate and inhibit irrelevant alternative hypotheses when retrieved, and therefore PI would not lead to biases in probability judgments. Further, in the optimal case people would completely unpack the set of alternative hypotheses into its subcomponents for inclusion in the comparison. Let subscripts represent the frequency with which each alternative hypothesis was bought when participants imagined going to the store and assume that the memory strength for each alternative is equivalent to its frequency of presentation in the learning phase of the experiment. Let A12 represent the focal hypothesis, an alternative that was presented 12 times in the learning phase of the experiment. The implicit alternative hypothesis, B, represents all other possible alternatives from the current list. When judging the likelihood of the focal hypothesis, the alternative hypothesis B would be optimally unpacked as follows: $B_2, B_2, B_6, B_6, B_{12}, B_{16}, B_{16}$. Then, the optimal comparison would be:

$$P(A, B) = \frac{s(A_{12})}{s(A_{12}) + s(B_2) + s(B_2) + s(B_6) + s(B_6) + s(B_{12}) + s(B_{16}) + s(B_{16})} \tag{1.2}$$

Here we see that the relative support for focal alternative A_{12} would be 12/72 (assuming that support is equivalent to the frequency of occurrence). In the next sections, two non-optimal effects of retrieving irrelevant alternatives are discussed: discrimination bias and dysinhibition bias.

1.1.2 Discrimination Bias

Consider the non-optimal case in which PI causes one to retrieve an irrelevant alternative from the previous list, but one fails to distinguish that the irrelevant alternative hypothesis is from the previous list. Then, that hypothesis might be included in the judgment in place of a relevant, current list hypothesis. One can conceptualize discrimination failure in terms of Support Theory. The set of alternative hypotheses, B, in equation 1 can be divided into two subsets: B, the set of relevant alternatives; and B', the set of irrelevant alternatives. The B' set should be ignored when judging the likelihood of A. However, when discrimination failure occurs, the B' set is not identified as irrelevant and is consequently included in the comparison in place of competitors from the B set:

$$P(A,B) = \frac{s(A_{12})}{s(A_{12}) + s(B') + s(B_2) + s(B_6) + s(B_6) + s(B_{12}) + s(B_{16}) + s(B_{16})} \tag{1.3}$$

In this equation, relevant alternative B_2 was replaced by an irrelevant alternative B'. If a strong irrelevant alternative from a previous list was included in the comparison in place of a weaker relevant current list alternative, the relative support for focal alternative A_{12} would decrease. It is not assumed that an interfering alternative hypothesis must be the most frequently presented alternative from the previous list, nor is it necessary for the interfering alternative to replace the least frequent current list alternative. It is assumed, however, that most often irrelevant alternatives are more frequent than the relevant alternative hypotheses they replace, because the alternatives that were more frequently presented are more likely to be retrieved. Thus, one possible effect of PI on probability judgments would be to decrease participants' probability judgments, because when stronger (more frequent) alternative hypotheses are considered in place of weaker (less frequent) alternative hypotheses, the ratio of support for the focal hypothesis to the set of alternative hypotheses decreases. It is possible that in some cases interfering items are weaker or equal to the relevant items they replace. We only assume that most often, stronger irrelevant alternatives will replace weaker relevant alternatives.

Discrimination failure could lead to two other effects on judgment. First, when participants were directly asked to judge irrelevant items from previous lists that had no probability of occurring on the current list, participants would give judgments greater than zero to some of the irrelevant, previous lists items. Second, as PI increased, relative accuracy would decrease. Relative accuracy was measured by the correlation between subjective judgments and objective probabilities. Relative accuracy would decrease because the process by which non-discriminated irrelevant alternatives supplant relevant alternatives is a process with at least three stochastic components. First, with some probability an irrelevant alternative could be retrieved, and with some probability an irrelevant alternative would not be retrieved. Second, with some probability, a retrieved irrelevant alternative would fail to be discriminated as irrelevant, and with some probability a retrieved irrelevant alternative would be correctly discriminated as irrelevant. Third, with some probability a retrieved, non-discriminated, irrelevant alternative could be stronger in support than the relevant alternative it supplants, and with some probability the alternative could be weaker or equal in support to the alternative it supplants. The stochastic nature of these processes causes variability in assessing the support for the set of alternatives to be larger when discrimination failure occurs than when discrimination is possible. Increasing the amount of variability in assessing the support for the set of alternatives, in turn, would cause subjective judgments to vary more. Consequently, subjective judgments would be less correlated with objective probability as PI increased. Thus, if discrimination failure occurs, as PI increases relative accuracy would decrease.

What support for the discrimination failure hypothesis exists? One of the most widely accepted accounts of the effect of PI on recall is the temporal discrimination theory (Baddeley, 1990; Underwood, 1945) in which the buildup of PI reflects a growing impairment in the ability to distinguish items that appeared on the most recent list from those that appeared on previous lists. Postman and Keppel (1977) had participants learn and recall 8 lists of word pairs. At the end of the experiment they presented participants with all of the word pairs from the entire experiment, and had participants identify on which list each word pair was learned. Postman and Keppel found that participants' ability to identify which items were learned on which lists was quite low. For instance, participants made fewer than 30% correct identifications (subtracting false alarms) even on the final list that was most recently learned. According to this temporal discrimination theory, when PI is present participants do not restrict their search to only the most recent list of items. Rather, participants search the entire set of category-specific items that have been presented because they fail to distinguish current list from previous list items. Wixted and Rohrer (1993) examined recall latency in a PI experiment. Based on the latency function and latency onset observed, they argued that as PI builds, a larger area of memory is searched. They theorized that the contents of a search set on a given trial are established by a retrieval cue (such as a category name). As PI increases, more items are activated by the retrieval cue. The more items activated by the retrieval cue, the lower the probability of retrieval associated with any individual item. This argument supports the possibility of discrimination failure, in that it argues that people are not searching only the relevant list of items, and consequently could retrieve irrelevant alternatives.

One argument against discrimination failure, though, is that in typical PI experiments, people report few intrusions from previous irrelevant lists (Kane &

Engle, 2000; Postman & Keppel, 1977), and when intrusions occur, participants have little difficulty identifying which of their own responses were correct and which were words from previous trials (Dillon & Thomas, 1975). These results suggest that people can identify items retrieved from previous lists as irrelevant, making the discrimination failure hypothesis rather unlikely. However another explanation of the low number of intrusions is that people have a bias against reporting intrusions. One can conceptualize this possibility in terms of a signal detection model, with one distribution representing old, irrelevant items and the other distribution representing new, relevant items. Then when people experience more interference, they shift their decision criterion to a more conservative location in order to reduce the number of false alarms (intrusions) they report. This would cause people to have both few intrusions as well as to report fewer relevant current list words. If few intrusions occur due to criterion shifting, discrimination failure is a viable possible effect of PI on probability judgments. Donaldson and Murdock (1968) argued that the decision criterion during recall is generally high, which might lend support for this explanation of low intrusion rates.

1.1.3 Inhibition Bias

Consider the case in which a person retrieves an irrelevant previous list alternative and recognizes that the item is irrelevant, but is unable to inhibit the irrelevant alternative. Then, the irrelevant alternative may act as a placeholder in WM, allowing fewer attentional resources to be available to retrieve, temporarily store, and compare other relevant alternatives. Such an effect of interference would cause people to consider fewer relevant alternative hypotheses. Consequently, probability judgments would increase as PI increased. If irrelevant alternatives are retrieved and not inhibited causing people to retrieve and compare fewer relative alternatives, the relative support for the focal alternative would increase. Consider the case in which two weak alternatives are not included in the judgment:

$$P(A,B) = \frac{s(A_{12})}{s(A_{12}) + s(B_6) + s(B_6) + s(B_{12}) + s(B_{16}) + s(B_{16}) + [s(B')]} \qquad (1.4)$$

Here we see that if two weak alternatives are replaced by B' in WM, but B' is discriminated as irrelevant and therefore is not included in the judgment, the relative support for A_{12} increases from 12/72 (in the optimal case) to 12/68. The ratio of support for the focal hypothesis in comparison with the alternative hypotheses would increase as the support for the set of alternative hypotheses decreased due to considering fewer alternative hypotheses.

The inhibition failure hypothesis makes different predictions about participants' judgments than did the discrimination failure hypothesis. First, when participants are asked to judge the proportion of an irrelevant, previous-trip item in their current shopping bag, the inhibition failure hypothesis predicts that participants would correctly identify that item as irrelevant. Thus, participants would judge the proportion of a previous list item in their shopping bag as zero. Second, the inhibition account predicts that judgments would increase as PI increases. Third, the inhibition failure account predicts that relative accuracy would remain unchanged as PI increases. Relative accuracy would remain unchanged because although fewer items would be compared in the probability judgment, this would cause all judgments to increase, but their relative relationship to objective frequencies would stay the same.

What support for the dysinhibition bias account exists? Anderson and Neely (1996) argued that overcoming PI requires active suppression of competitors at retrieval. Postman and Hasher (1972) argued that decreased recall from PI could be due to output interference, which "refers to the detrimental effects of recall on further recall . . . Recall of items from List A will militate against the reproduction of items from List B because a change in response set or a shift from one set of retrieval cues to another is required as S moves from one group of words to the other" (p. 276). This account of PI suggests that decreased recall from PI is due to interference from previous lists blocking the retrieval of current list items, not due to discrimination failure. Kane and Engle (2000) argued that participants who are less susceptible to the effects of PI are better able to inhibit competition from prior list items. If this account of PI is correct, when participants make judgments irrelevant, previous list items would be generated and participants would be unable to suppress those competitors when trying to retrieve further alternative hypotheses. In consequence, those irrelevant items would act as placeholders in WM and would cause fewer relevant alternative hypotheses to be retrieved and compared with the focal hypothesis. Then, subadditivity would increase due to fewer alternative hypotheses being compared with the focal hypothesis.

1.2 Role of WM/Attentional Capacity in PI

WM span may be related to the ability to inhibit irrelevant information. If inhibition failure occurs as a result of retrieving irrelevant alternative hypotheses, those better at inhibiting irrelevant information would make less biased probability judgments than those worse at inhibiting irrelevant information. Researchers currently argue that tasks that are intended to measure individual differences in WM include a component that measures one's ability to monitor and suppress interfering, task-irrelevant information (Lustig, May, & Hasher, 2001; May, Hasher, & Kane, 1999). May et al. (1999) noted that measures of WM capacity examine not only how much information one can store and process simultaneously, but also one's ability to suppress interference. For instance in Daneman and Carpenter's (1980) reading span measure of WM participants must read sets of between 2 to 6 sentences and must remember the final word of each sentence. The number of sentences in each set increases throughout the task. WM span is scored as the number of words recalled in the correct order. May et al. noted that in this task, participants must not only process the sentences and store the words temporarily, but must also ignore words from previous sets of sentences. Because set size increases throughout the task, participants have more interference from previous sets when trying to recall the largest set sizes at the end of the task. Therefore, scores are most likely to be negatively affected by susceptibility to PI in the larger sets, because the opportunity to score higher by doing well on these larger set sizes occurs when PI is greatest. Thus, tasks which measure WM entail a PI component, and thus those who score better on these tasks not only can store and process more information simultaneously, but also can better inhibit previous irrelevant information. In fact, Lustig et al. (2001) found that when they gave participants an interference-reduced version of the reading span in which the larger set sizes occurred earlier and the smaller set sizes occurred later, significant correlations that normally existed between WM span and prose recall were no longer obtained. This finding suggests that one's ability to inhibit irrelevant previous information is an important component of WM.

Other support that differences in WM capacity reflect differential capacity to inhibit irrelevant information has been found. Carretti, Cornoldi, and De Beni (2004) compared participants with high WM capacity, high spans, with participants with low WM capacity, low spans, on a short-term memory task and found that high spans had fewer intrusions and a lower activation level of irrelevant information. Engle, Conway, Tuholski, and Shisler (1995) found that inhibition requires controlled-attention, WM resources. Engle et al. had participants engage in an attention-demanding task while performing a negative priming task, and found that unlike usual findings in negative priming experiments, response times were not slowed when responding to to-be-ignored letters on the previous trial. This finding suggests that inhibition requires attentional resources, and is not an automatic process. When attention is being used for another task, participants do not have enough attentional resources to inhibit task-interfering information. Rosen and Engle (1997) found support that interfering information blocks further item generation. Rosen and Engle compared high and low-span participants' performance on a verbal fluency task, and found that high spans generated more words than did low spans.

When participants were allowed to state repetitions, high and low spans retrieved an equal number of words, but low spans generated more repetitions than high spans. Rosen and Engle posited that failing to suppress repetitions prevented participants (especially low-span participants) from being able to generate new retrieval cues to retrieve further items. Rosen and Engle's findings suggest that unsuppressed, interfering information can impede the generation of alternatives, and that high spans are better at suppressing interfering information.

Kane and Engle (2000) found direct evidence relating WM to PI susceptibility. Participants engaged in a PI task in which they learned and recalled three lists of words from one category. High spans and low spans were compared in their susceptibility to PI, and Kane and Engle found that high spans were less susceptible to the negative effects of PI on recall than were low spans. Furthermore, dividing attention had little to no effect on recall during PI buildup for low spans, but it increased deficits in recall during PI buildup for high spans. It was concluded that dividing attention affected high spans more so than low spans because under normal conditions high spans use attentional capacity to suppress PI. When attention was divided, however, they had fewer attentional resources available for suppressing interference, and thus PI caused their recall to decrease. Low spans, contrarily, do not have as much controlled attention capacity to counteract the effect of PI on recall from long-term memory under normal conditions and therefore they are equally affected by PI when their attention is divided as when it is not divided. Thus individual differences in WM relate to differences in the ability to activate relevant and suppress irrelevant information.

Previous research has examined the role of WM in making probability judgments. Dougherty and Hunter (2003) found that differences in the subadditivity of probability judgments was correlated with individual differences in WM. They compared high spans and low spans on the subadditivity[1] of their probability judgments, and found that high spans were less subadditive than low spans. It has been posited that subadditivity arises due to a failure to consider all possible alternative hypotheses when making a probability judgment (Tversky & Koehler, 1994; Dougherty & Hunter, 2003). The judged probability of a focal alternative has been found to decrease as the number of alternatives considered increases (Dougherty, Gettys, & Thomas, 1997). In fact, Dougherty and Hunter found that high WM spans recalled more alternative hypotheses, made lower probability judgments, and were less subadditive than low spans. It was argued that that high spans' greater capacity allowed them to consider more alternative hypotheses at one time when comparing the evidence for a given focal hypothesis relative to all other potential hypotheses.

Although Dougherty and Hunter (2003) speculated that irrelevant alternatives may affect judgment by interfering with the retrieval of relevant alternatives, they did not directly examine this possibility. Indeed, research has yet to examine whether the relationship between WM capacity and probability judgment is due to differential susceptibility to interference. If discrimination failure occurs and partici-

[1] Subadditivity occurs when people judge the likelyhood of an event occurring (such as cancer) lower than the summed judgments of specific subcomponents of the event occuring (such as specific types of cancer) (Tversky & Koehler, 1994).

pants are unable to distinguish one list from another list, high WM span participants cannot use their attentional capacity to suppress irrelevant items. Thus, if PI causes discrimination failure, it was predicted that no relationship between PI and WM span would be found. Contrarily, if people are able to distinguish between items from different lists, and consequently inhibition failure occurs, it was hypothesized that WM span would moderate the effect of PI on probability judgments because high spans can use their attentional resources to inhibit irrelevant alternatives. Thus, it was predicted that if inhibition failure occurred, PI would lead to more bias in judgments for low spans than for high spans.

In sum, a key part of making accurate probability judgments is retrieving relevant alternative hypotheses to include in the comparison of support for focal hypotheses to support for alternative hypotheses. Irrelevant alternative hypotheses could be retrieved when participants experience interference. Irrelevant alternatives could bias probability judgments either by replacing relevant alternative hypotheses in the probability judgment, discrimination failure, or by utilizing WM space and causing fewer relevant alternatives to be retrieved, inhibition failure. If discrimination failure occurred, probability judgments would decrease as PI increased. On the other hand, if the dysinhibition bias occurred, probability judgments would increase as PI increased. Individual differences in WM relate to individual differences in susceptibility to interference. High spans are less susceptible to the effects of PI than low spans. If participants are able to discriminate relevant from irrelevant alternatives, the differences in susceptibility to interference for high versus low spans may transfer to individual differences in probability judgment bias under interference.

Chapter 2

Experiment 1

Experiment 1 determined the effect of PI on proportion judgments, and examined the role of WM capacity in such effects of PI. More specifically, Experiment 1 tested how PI affected the subadditivity and relative accuracy (the relationship between objective frequencies and participants' subjective probability judgments) of judgments.

2.1 Methods

The method for Experiment 1 was based on the multiple-trial, free-recall PI paradigm. Participants learned words, each presented a varied number of times, engaged in a distractor task, recalled the words learned in the current list, and then made proportion judgments about the items in that list. Participants learned and judged three lists of words from one category (buildup of PI) and then learned and judged a list of words from a new category (release from PI). Finally, WM span was measured using Turner and Engle's (1989) Operation-Span (O-span) task.

2.1.1 Participants

One-hundred twelve University of Maryland students participated and received monetary compensation for their time. Participants were classified into WM capacity tertials based on their O-span score (high: $n = 38$; middle: $n = 35$; low: $n = 33$). The tertial classification levels were determined from a separate experiment of over 150 participants (Sprenger & Dougherty, 2005).

2.1.2 Materials

The experiment was conducted on the computer using a Java program for the PI task and DirectRT software (Jarvis, 2004) for the O-span task. A tape recorder was used to record verbal memory recall. For the PI learning task, 48 words from the categories fruits, vegetables, and alcoholic and nonalcoholic beverages from Battig and Montague's (1969) category word norms were used. Words used were no longer than 10 letters. Further, eight color words were used for the practice trial of the task. For the O-span task, 60 concrete words from the Toronto word pool, paper and a pencil were used.

2.1.3 Design and Procedure

The design was a 3 (WM span: high, middle, low) x 4 (List: 1-4) x 5 (Alternative Frequency: 0, 2, 6, 12, 16) mixed factorial, with WM span measured between subjects, and List and alternative frequency manipulated within subjects. The entire experiment, including collecting individual difference data, was conducted in approximately 50 minutes. Participants engaged in two experimental tasks: a PI task and the O-span task. For the PI task, participants first engaged in a practice trial of the task, then engaged in three PI buildup trials and one PI release trial. Each trial consisted of 4 parts: participants first engaged in a learning phase, then engaged in a rehearsal-prevention task to eliminate recency effects, then verbally recalled the items from the current trial, and finally made proportion judgments on the items from the current trial. Participants cycled through these four phases five times over the course of the experiment: Once for the practice trial and four additional times for the main portion of the experiment.

PI Task—Learning Phase

For the learning phase, participants were instructed to imagine that they needed to buy some items from a futuristic grocery store (which sold only cans of items) for a party they were planning. They were told to imagine that they had an empty shopping cart, and that they would see words representing cans of items that they would pick up from the shelf of the grocery store and place into their shopping cart. Participants were informed that they would be getting more than one can of each item, and that the items would be gathered from the grocery store in a random order. Each item that was purchased was represented by a word on the computer screen. In order to add an item to their shopping cart, participants had to press the first letter of the name of the item they just picked up. Upon pushing the first letter, the item disappeared and the next item appeared in its place on the computer screen. On each list participants saw eight alternatives. Each alternative was assigned to one of the following presentation frequencies: $16 - 16 - 12 - 12 - 6 - 6 - 2 - 2$. For example, one participant might be presented with 16 cans of “limes”, 16 cans of “broccoli”, 12 cans of “apples”, 12 cans of “celery”, 6 cans of “lettuce”, 6 cans of “peaches”, 2 cans of “strawberries”, and 2 cans of “tomatoes”. Assignment of alternatives to List Frequency was determined randomly for each participant. For the first three trips to the store (PI build-up) participants bought produce items. For the fourth and final trip to the store (PI release) participants bought beverage items.

PI Task—Rehearsal Prevention

A rehearsal prevention task was implemented in order to reduce possible recency effects. Participants counted backwards by threes for 20 seconds.

PI Task—Memory Recall

Participants heard a series of 12 beeps (each 1350 milliseconds apart), and were instructed to recall aloud one kind of can currently in their shopping bag after each beep. Participants were instructed to respond with any item they thought of, even if they knew the recalled item was incorrect. They were further instructed that if they could think of nothing, they should not respond. The short amount of time between beeps and the instructions to recall whatever came to mind were chosen to encourage participants to respond with intrusions if intrusions were generated during the recall session. Responses were tape-recorded.

PI Task—Proportion Judgment

In this part of the PI task, participants made proportion judgments for each of the eight items they bought on a given shopping trip, as well as for four items (one item of each possible frequency) that they bought on previous shopping trips. For the first trip, they made judgments of 4 other fruits and vegetables not appearing in that section or future sections. We asked participants to judge items from previous shopping trips for two main reasons. First, PI is maximized when the irrelevant (prior list) information and the relevant target list information both gain access to WM at the same time (Postman & Hasher, 1972). Second, we were interested in whether participants could discriminate between list-relevant and list-irrelevant items. Giving non-zero judgments to items bought on previous shopping trips suggests that participants experienced discrimination failure. Participants were instructed to imagine that they were now back at home with their newly-bought items all in one shopping bag. For each item, they were asked, "Out of all of the kinds of items in your current shopping bag, what proportion are [item]?" Participants were cautioned that they may also be asked to make judgments about items that were not actually in their shopping bag. Participants typed their judgments into a textbox on the computer. One relevant frequency-12 item was always judged first and the other was always judged last, in order to determine if participants' proportion judgments changed as they made other judgments. All other items were judged in a random order.

Operation-Span Task

After completing the PI task, participants completed the Operation-span (O-span) task as a measure of WM span (Turner & Engle, 1989). The O-span task required participants to retain a growing list of words while solving mathematical problems. For example, on one presentation participants would be shown $(4 * 3) - 3 = 9$? DOOR. Participants were required to read the equation aloud, verify whether the equation was true or false, and then read the word aloud. After saying the word, the experimenter advanced to the next operation-word pair. This continued until the participant was prompted to recall the words from that set in the order in which they were presented. Participants were presented with 15 sets of equation-word pairs with set sizes ranging from 2 to 6. Each set size occurred three times in random order (thus 60 total operation-word pairs were presented). Performance on the O-span task was measured by summing the number of words recalled in the correct order and for which the participant had correctly verified the math equation. Scores could range from 0 to 60. Data from participants correctly answering fewer than 85% of the math problems was not included in the analyses. A detailed description of the operation-span task is presented in Turner and Engle (1989).

2.2 Results and Discussion

The data from six participants were not included in the analyses. Three participants' data were not included because they scored less than 85% correct on the O-span math problems. Computer failure caused the loss of one participant's data. One participant's data was excluded because s/he gave judgments of only 0. One participant's subadditivity was greater than 2.5 standard deviations above the mean, and therefore the data was excluded from the analyses. The participant excluded due to being an outlier was in the high span group. The mean subadditivity for high spans for each list was: 148.05 (SD=83.64) for list 1; 128.37 (SD= 69.64) for list 2; 110.87 (SD=53.53) for list 3; and 117.63 (SD=58.29) for list 4. The participant's subadditivity for each list was: 450 for list 1; 450 for list 2; 440 for list 3; and 430 for list 4.

2.2.1 Recall

We hypothesized that as PI increased over the first three lists, recall would decrease, and then on the fourth release from PI list that recall would increase again. Further, we hypothesized that WM span would interact with List, such that when comparing high and low spans' recall, high spans would be less affected by PI on lists 2 and 3 than low spans (replicating Kane and Engle's (2000) study).

The left part of Table 2.1 presents the mean number of words recalled (out of eight) for each of the four lists as a function of WM span. There was a significant main effect of List on the number of words recalled such that as PI increased, recall decreased, $F(3, 309) = 15.03, p < 0.0001, \omega^2 = .09$. Further, trend analyses revealed that a quadratic trend fit the changes in recall well, $F(1, 103) = 48.29, p < 0.0001$. This finding replicates typical findings in PI experiments that recall decreases as PI increases and then increases again at release from PI. Although there was a main effect of WM span on the number of items recalled, $F(2, 103) = 8.39, p = 0.0004, \omega^2 = .03$, there was no interaction between List and WM span, contrary to the findings of Kane and Engle (2000). Overall, high spans recalled more words than did low or middle spans. Our method for measuring recall was not typical of other experiments using the PI paradigm because participants were required to recall under time pressure. Johnson, Kounios, and Reeder (1994) found that the time course for source monitoring is slower than the time course of recognition

Table 2.1: Mean Recall as a Function of List

	Experiment 1			Experiment 2		Experiment 3	
	WM Span			List Frequency		List Frequency	
	High	Middle	Low	Ascending	Descending	Ascending	Descending
List 1	6.9 (.15)	5.88 (.21)	5.88 (.20)	6.57 (.18)	6.82 (.26)	6.79 (.22)	6.79 (.22)
List 2	6.02 (.28)	5.77 (.24)	5.69 (.20)	6.49 (.20)	6.06 (.27)	7.36 (.25)	6.46 (.25)
List 3	6.29 (.31)	5.31 (.23)	4.97 (.30)	6.34 (.21)	5.60 (.28)	7.14 (.22)	6.46 (.22)
List 4	6.95 (.22)	6.45 (.19)	6.30 (.19)	7.17 (.14)	6.90 (.19)	7.00 (.28)	6.71 (.28)

memory. Thus, people can quickly detect whether or not they have seen items before, but require more time to determine from where exactly they saw items. Perhaps by imposing time pressure on probability judgments under PI, participants could not discriminate relevant from irrelevant items and therefore their recall was more affected by PI than it would have been under normal conditions without time pressure.

We also examined the number of intrusions of previous list items participants reported during recall. The left part of Table 2.2 presents the mean number of recall intrusions for each of the four lists as a function of WM span. There was a significant main effect of List on the number of intrusions reported, such that more intrusions were reported on the second and third lists when PI was greatest than on the first and fourth lists when PI was least present, $F(3, 309) = 29.85, p < .0001, \omega^2 = .17$. No main effect of WM and no interaction between WM and List on the number of intrusions reported were found. Overall the number of intrusions reported was low (less than 1), but more intrusions were reported when PI was greatest. Intrusions reflect a failure to discriminate between lists. The finding that participants reported intrusions of previous list items suggests that they experienced list discrimination failure. However, time pressure may have impaired participants' ability to monitor if their recall output was a current list item or an intrusion. Perhaps when participants have more time, they are better able to monitor their recall, and thus the mean number of intrusions reported in this experiment may be biased upward.

Table 2.2: Mean Number of Verbal Intrusions during Recall as a Function of List

	Experiment 1			Experiment 2	
	WM Span			List Frequency	
	High	Middle	Low	Ascending	Descending
List 1	0.00 (.00)	0.00 (.00)	0.00 (.00)	0.00 (.00)	0.03 (.02)
List 2	0.32 (.14)	0.49 (.15)	0.67 (.15)	0.28 (.11)	0.17 (.11)
List 3	0.92 (.25)	0.69 (.26)	1.09 (.26)	0.36 (.11)	0.25 (.11)
List 4	0.05 (.03)	0.00 (.00)	0.00 (.00)	0.00 (.00)	0.03 (.02)

2.2.2 Subadditivity

We hypothesized that if discrimination failure occurred, as PI increased over the first three lists, subadditivity would decrease, and then on the fourth, release from PI list subadditivity would increase again. If discrimination failure occurred, we predicted WM would not relate to differences in subadditivity across lists. In contrast, if inhibition failure occurred, we hypothesized that as PI increased over the first three lists, subadditivity would increase, and then on the fourth release from PI list subadditivity would decrease again. If inhibition failure occurred, we hypothesized that WM would interact with List such that subadditivity would increase across the first three PI build-up lists more for high spans than for low spans. Subadditivity was measured by summing each participant's judgments of the eight relevant items for each list. Judgments of list irrelevant items were excluded from the subadditivity measure. A 3 (WM span: high, middle, low) x 4 (List: 1-4) mixed factorial ANOVA was conducted to determine whether subadditivity changed as a function of PI, and whether WM interacted with changes in subadditivity due to PI.

Figure 2.1 presents mean subadditivity for each of the four lists as a function of WM span. A main effect of List was found for subadditivity, $F(3, 309) = 18.63, p = 0.0001, \omega^2 = .11$. Trend analyses revealed that both a linear trend, $F(1, 105) = 24.73, p < 0.0001$, and a quadratic trend, $F(1, 105) = 21.24, p < 0.0001$, fit the changes in subadditivity well suggesting that as PI increased, subadditivity decreased and that at release from PI subadditivity increased again although not back to pre-PI levels. This pattern of results supports the discrimination failure hypothesis. It also appears that subadditivity was affected by some form of learning. Subadditivity increased at release from PI, but did not increase back to pre-PI levels. Perhaps participants realized that their judgments on the first list were not additive, and attempted to make lower judgments on successive lists in order to be additive. Or, perhaps participants learned how many alternatives were on each list and better distributed their support across more alternatives. The decrease in participants' subadditivity across lists was not due solely to learning, however, because at release from PI participants' subadditivity increased again as indicated by the significant quadratic trend. Although the release from PI subadditivity was not equal to levels of subadditivity before PI was introduced, it did increase significantly, suggesting that PI affected subadditivity above and beyond learning. Thus, it appears that both discrimination failure and learning caused participants' subadditivity to decrease as PI increased.

The subadditivity of low and high spans' judgments were equally affected by PI; neither a main effect of WM span on subadditivity, $F(2, 103) = 1.97, p > 0.05$,

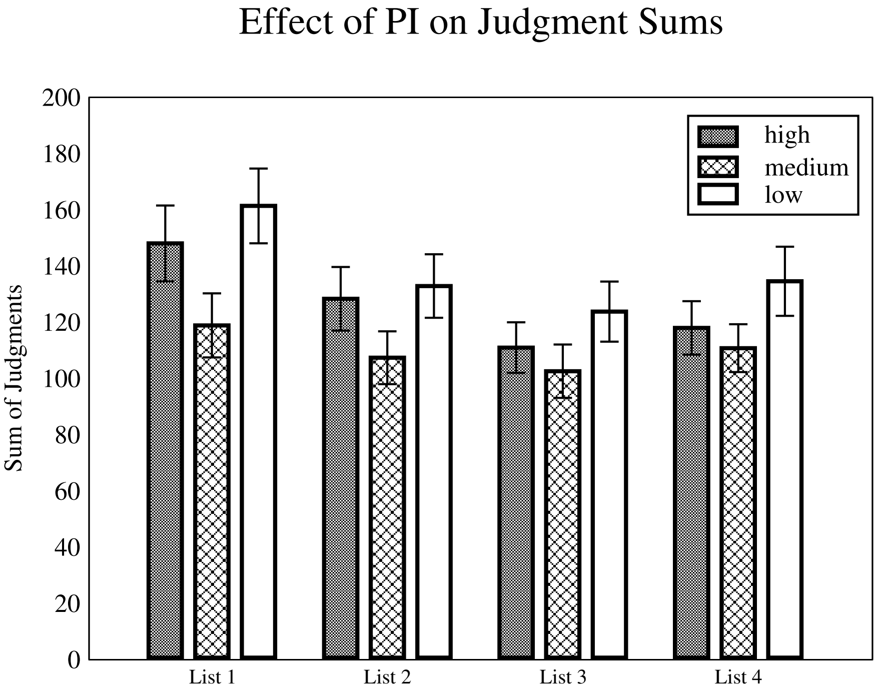

Figure 2.1: Change in judgment sums as a function of PI.

nor an interaction between List and WM span was found, $F(6, 309) = 1.48, p > 0.05$. This suggests that discrimination failure affected high and low spans equally. Note that on the practice list, no significant effect of WM on subadditivity was found. This finding contrasts Dougherty and Hunter's (2003) findings that high spans were less subadditive than low spans. An explanation of these disparate results is suggested in the General Discussion.

2.2.3 Focal Judgments: Learning Effects

Our hypotheses for the focal judgments paralleled those for subadditivity. Each list had two focal items, each of which had been presented 12 times in the learning phase of the task. One focal item was always judged first and the other focal item was always judged last, after making all other judgments. Examining those two judgments allowed an exploratory examination of changes in judgment as participants made their other 10 judgments.

Figure 2.2 presents the mean of participants' focal judgments judged first, and Figure 2.3 presents the mean of participants' focal judgments judged last for each of the four lists as a function of WM span. A marginally significant main effect of List was found for focal judgment judged first, $F(3, 309) = 2.26, p = 0.081$, and a significant main effect of List was found for focal judgments judged last, $F(3, 309) = 4.27, p = 0.006, \omega^2 = .02$. For the focal judgments judged first, a marginally significant quadratic trend was found, $F(1, 103) = 3.75, p = 0.056$, and for the focal judgments judged last significant linear, $F(1, 103) = 8.46, p < 0.005$, and quadratic, $F(1, 103) = 4.08, p < 0.05$, trends were found. Thus, as PI increased in lists 2 and 3, focal judgments decreased, irrespective of whether focal items were judged first or last. Further, the quadratic trends suggest that at release from PI, focal judgments increased again. However, the linear trend for the focal judgments judged last suggests perhaps that participants give even lower judgments to the items in the last position because they had already made 11 judgments at that point and realized that to be additive they needed to make lower proportion judgments for that last item. This finding suggests that judgments were affected by learning. Note in Figure 2.3 that for high and middle span participants, focal judgments judged last did not increase on the fourth, release from PI list. For the focal items judged last, participants may have realized that their other judgments sum to something greater than 100%, and consequently attempted to adjust their judgments downward to

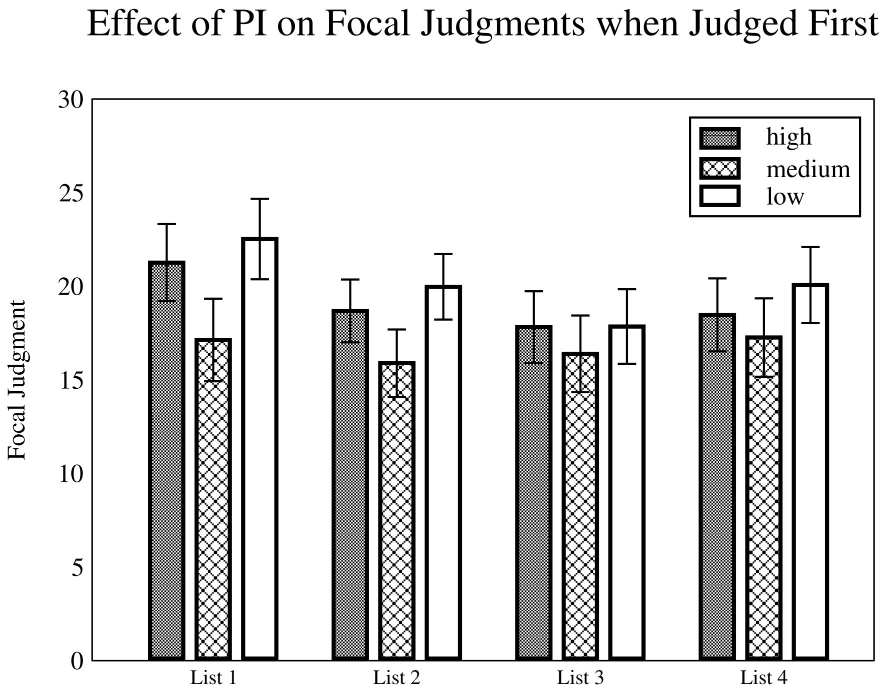

Figure 2.2: Change in first focal judgments as a function of PI.

be additive. Perhaps the focal items judged first better reflect the effect of PI on judgment, and focal items judged last better reflect the effect of learning on judgment. No main effect of WM span or interaction between WM span and List on focal judgments was found.

2.2.4 Relative Accuracy of Judgments

The relative accuracy of judgments was measured with Somers's Dxy (Somers, 1962). Somers's Dxy analyzes all possible predictor-criterion (objective proportion-subjective judgment) pairs and examines the degree to which when one predictor is greater than the other predictor the respective criterion is also greater than the other criterion. Gonzalez and Nelson (1996) discussed the implications of using various

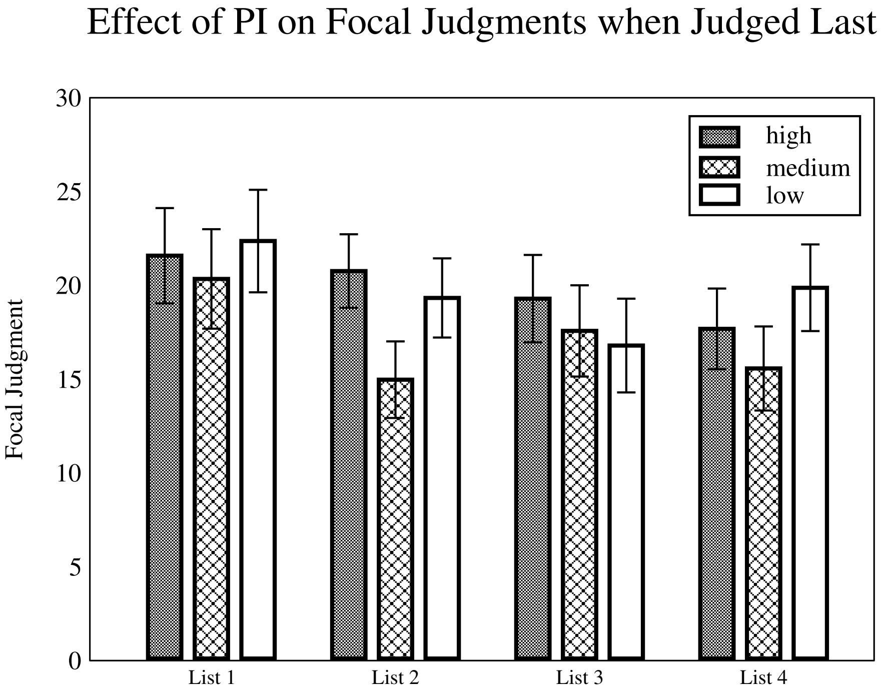

Figure 2.3: Change in last focal judgments as a function of PI.

measures of association and concluded that Somers's Dxy should be used for cases in which ties on the criterion variable are unambiguous, but ties on the predictor variable are ambiguous. Somers's Dxy was calculated for each participant for each list. The analyses reported here include the 4 irrelevant previous list judgments[1].

We hypothesized that if discrimination failure occurred, relative accuracy would decrease as PI increased across the first three lists, and then relative accuracy would increase again on the fourth, release from PI list. In contrast, if inhibition failure occurred, we hypothesized that relative accuracy would be unaffected by changes in PI across lists.

[1] Note that the analysis examining the effect of PI on relative accuracy were the same when the relative accuracy correlations included the 4 irrelevant previous list judgments and when those judgments were not included in the analysis

Figure 2.4 presents the mean of participants' Somers's D correlations for each of the four lists as a function of WM span tertial. A main effect of List on relative accuracy was found, $F(3, 309) = 52.78, p < 0.0001, \omega^2 = .27$. Further, trend analyses revealed that both a linear trend, $F(1, 103) = 10.01, p < 0.005$, and a quadratic trend, $F(1, 103) = 132.03, p < 0.0001$, fit the changes in relative accuracy suggesting that as PI increased, participants' subjective proportion judgments became less correlated with the objective frequencies. Then, at release from PI, correlations increased. This finding is consistent with the discrimination failure account. No interaction between List and WM span was found, $F(6, 309) = 0.74, p > 0.05$. A main effect of WM was found, $F(2, 103) = 3.29, p = 0.041, \omega^2 = .01$. High spans had higher correlations and thus greater relative accuracy than low spans, but this did not interact with the amount of interference present. Therefore although high spans had better correlations overall than did low spans, they were equally affected by discrimination failure in lists two and three when interference was greatest. Although differences in relative accuracy due to WM span were not predicted, perhaps high spans were more consistent in assessing support for the set of alternative hypotheses. Differences between high and low spans in relative accuracy do not appear to be due to differences in discrimination ability, as WM span did not interact with the amount of PI to account for changes in relative accuracy.

2.2.5 Irrelevant Items

On each list participants judged four items that actually occurred on the previous list, and not on the current list. Non-zero judgments for those items reflect discrimi-

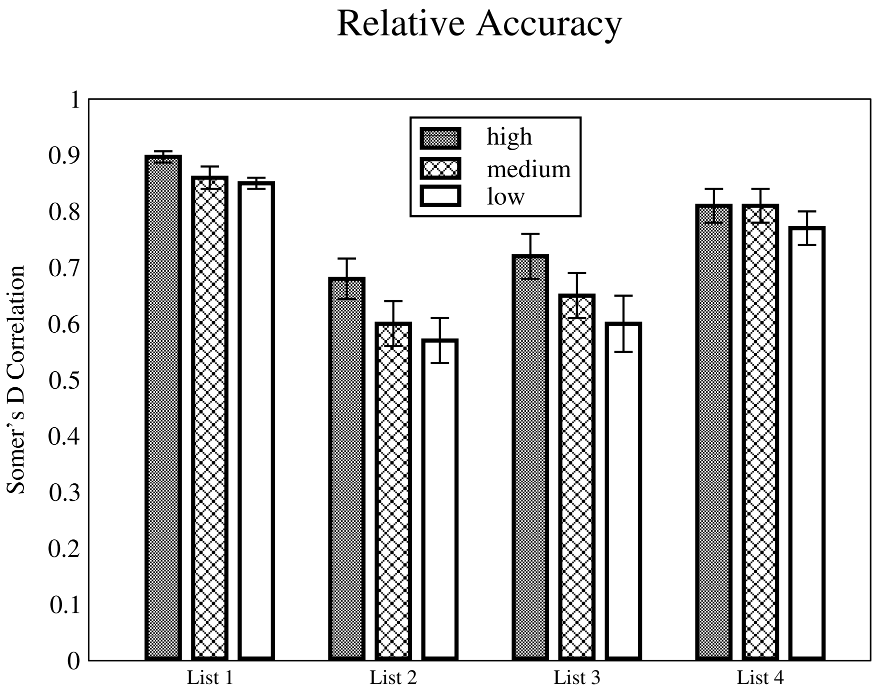

Figure 2.4: Change in relative accuracy as a function of PI.

nation failure in that they reflect that participants failed to identify that those items occurred on the previous, irrelevant list rather than on the current, relevant list. If discrimination failure occurred, we hypothesized that the number of irrelevant items given non-zero judgments would be greatest on lists 2 and 3 when PI was greatest, and that the number of irrelevant items given nonzero judgments would be least on lists 1 and 4 when PI was less present. In contrast, if inhibition failure occurred, we hypothesized that few irrelevant items would be given non-zero judgments, and that no differences across list would be found.

The left part of Table 2.3 presents the mean number of previous list items given non-zero judgments (out of four possible) for each of the four lists as a function of WM span tertial. A main effect of List on the mean number of irrelevant items given non-zero judgments was found, such that the mean number of irrelevant items given non-zero judgments on list 1 and on list 4 when PI was less than the mean number of irrelevant items given non-zero judgments on list 2 and on list 3, $F(3, 309) = 68.38, p < 0.0001, \omega^2 = .32$. This suggests that on lists 2 and 3 when PI was greatest, participants had difficulty distinguishing relevant from irrelevant items. On lists 1 and 4 when PI was not present, participants were able to easily distinguish relevant from irrelevant items. No main effect of WM span or interaction between WM span and List were found for the number of irrelevant items given nonzero judgments.

This finding adds further direct support that participants experienced discrimination failure. If participants were able to completely discriminate relevant from irrelevant alternatives, they would give all items from previous lists judgments of 0, indicating that no proportion of those items had been bought on the current shopping trip. It could be argued that because the majority of items that participants judged were relevant items (8 items out of 12), participants had a response bias to give nonzero judgments to all items, since that was the correct response the majority of the time. However, the finding that fewer irrelevant items were given nonzero judgments on lists 1 and 4 indicates that the results were due to discrimination failure, and not only due to a response bias.

2.2.6 Reaction Time

Irrespective of whether inhibition failure or discrimination failure occurred, we hypothesized that reaction time (RT) rate would be fastest when PI was least present on lists one and four and slowest when PI was greatest on lists two and three.

Table 2.3: Mean Number of Irrelevant Items Given Non-Zero Judgments as a Function of List

	Experiment 1			Experiment 2		Experiment 3	
	WM Span			List Frequency		List Frequency	
	High	Middle	Low	Ascending	Descending	Ascending	Descending
List 1	0.07 (.07)	0.17 (.11)	0.07 (.07)	0.17 (.07)	0.11 (.08)	0.35 (.16)	0.43 (.16)
List 2	1.07 (.34)	1.92 (.43)	2.00 (.39)	1.50 (.27)	1.11 (.23)	0.07 (.11)	0.14 (.11)
List 3	0.87 (.36)	1.08 (.42)	1.20 (.39)	1.17 (.23)	0.92 (.20)	0.18 (.14)	1.14 (.14)
List 4	0.07 (.07)	0.17 (.11)	0.13 (.13)	0.19 (.15)	0.31 (.14)	0.07 (.11)	0.14 (.11)

Reaction times were transformed to rates (1/RT) to reduce the skew-ness of the distribution. Then, two rate measures were calculated: the average rate to make all judgments was measured by averaging all 12 judgment RT rates for each list, and the average RT rate to make irrelevant judgments was measured by averaging the four irrelevant (previous list) judgment RT rates for each list. Separate ANOVAs were conducted for each of these two rate measures.

Figure 2.5 presents mean RT rates for each of the four lists. For average RT rates, a main effect of List was found such that average RT rates were slowest when PI was greatest (on lists 2 and 3) and fastest on list 1 and 4 when PI was least, $F(3, 309) = 67.55, p < 0.0001, \omega^2 = .32$. Figure 2.6 presents mean irrelevant judgment RT rates for each of the four lists. For average irrelevant judgment RT rates, a significant main effect of List was found, $F(3, 309) = 111.25, p < 0.0001, \omega^2 = .44$. Again, participants were slower to judge lists 2 and 3 which had the most interference than to judge lists 1 and 4. No interaction between List and WM or main effects of WM were found for the average RT rates or for the average irrelevant judgment RT rates. The finding that participants were slower to make proportion judgments on lists in which PI was greatest suggests that more processing was necessary for those lists than when no PI was present. That the rate to make judgments was slower as PI increased could suggest that discrimination between current list information and prior list information became increasingly difficult with the build-up of PI.

Experiment 1 revealed three main findings consistent with the discrimination failure account. The discrimination failure account predicts that as PI increased, participants would fail to discriminate relevant from irrelevant alternatives and con-

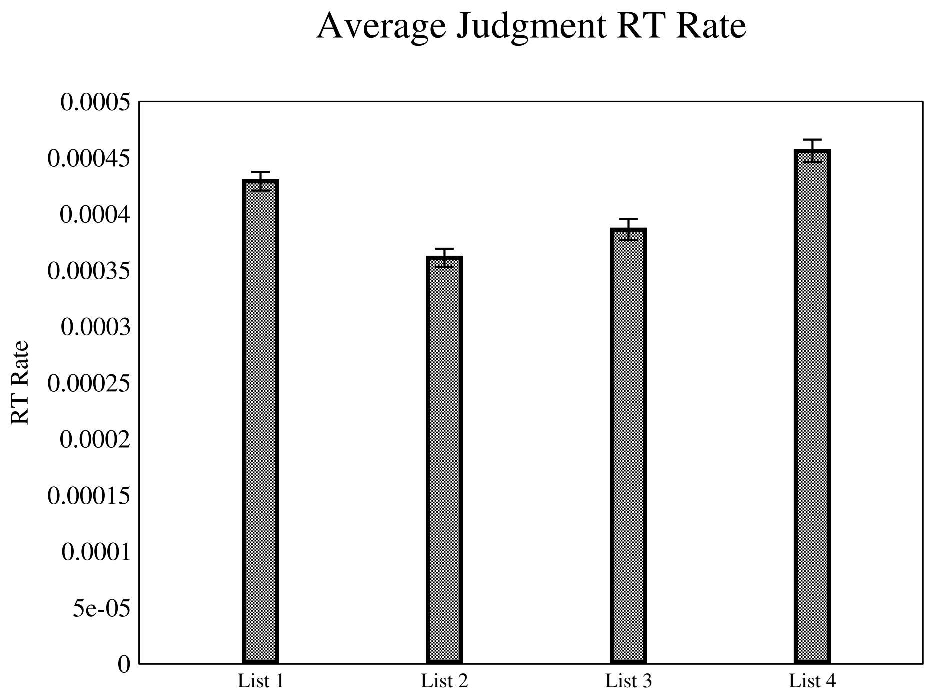

Figure 2.5: Change in average judgment RT rate as a function of PI.

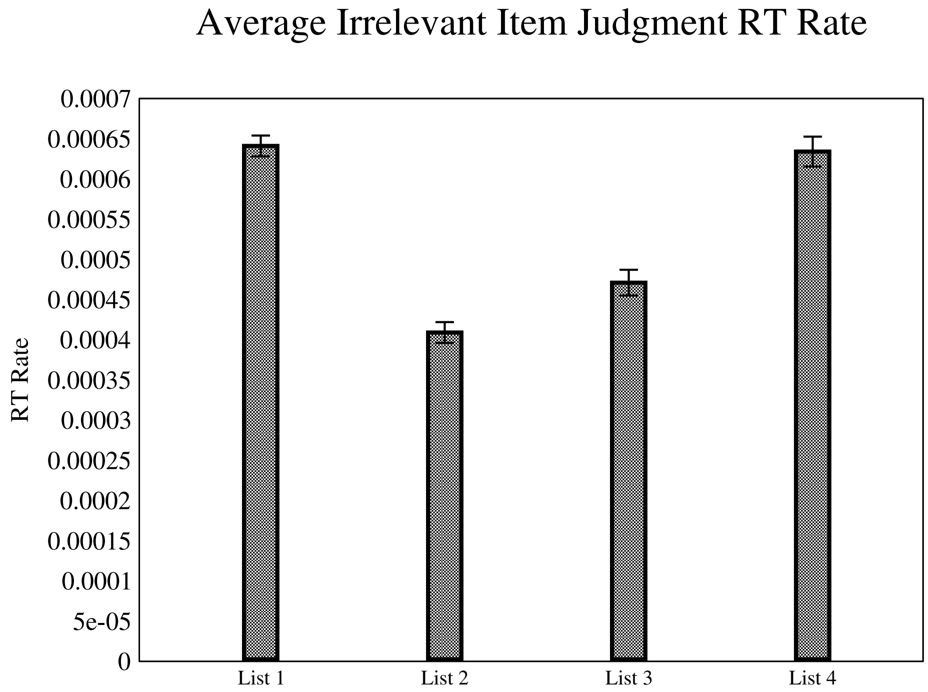

Figure 2.6: Change in average irrelevant judgment RT rate as a function of PI.

sequently that irrelevant alternatives would supplant relevant alternatives in the comparison process. Further, the discrimination failure account predicts that the non-discriminated, irrelevant alternatives would be stronger than the relevant alternatives they would supplant, causing judgments to decrease as PI increased. The first support for the discrimination failure account we found was that as PI increased, participants' judgments decreased. Importantly, on the release from PI trial, judgments increased, though they did not increase back to the level that was observed for the first list. One possible explanation for why judgments did not increase to a greater extent on the release from PI list is that changes in subadditivity were also in part due to learning effects. Perhaps participants learned that their judgments were too high and not additive, and attempted to make lower judgments on each trial. Or, perhaps participants learned the number of alternatives on each list was eight, and better partitioned their judgments across more alternatives than on the original list. Further evidence that learning may have affected judgments is that focal judgments judged last on each list showed a linear trend, suggesting that they decreased across each list. In Experiment 2, we manipulated the number of alternatives on each list in an attempt to reduce participants' learning the number of alternatives on each list. The discrimination failure account predicts that relative accuracy would decrease as PI increased, because as PI increased the amount of variance in the assessment of the set of alternative hypotheses increased. Variation in the assessment of the set of alternative hypotheses could increase due to the stochastic nature of discrimination failure in whether irrelevant alternatives are generated, in whether irrelevant alternatives are discriminated, and in whether non-discriminated irrelevant items are stronger than the relevant items they supplant in the judgment process. This increased variability in assessing support for the set of alternative hypotheses would cause judgments to vary more and thus to be less correlated with their objective proportions. The second main finding consistent with the discrimination failure account was that relative accuracy decreased as PI increased, but increased back to the level of list one on the release from PI list. The third main finding consistent with the discrimination failure account was that as PI increased participants gave irrelevant, previous list items nonzero judgments, suggesting that they failed to discriminate those items as irrelevant. Also consistent with the discrimination failure account, WM span did not interact with list for subadditivity, relative accuracy, or the number of nonzero judgments given to irrelevant items. This suggests that inhibition failure did not play a role in changes in judgments as PI increased, because WM is related to the ability to inhibit irrelevant information. Therefore, if inhibition failure was a factor, WM span should have interacted with list to affect judgment performance. In contrast, the discrimination failure account does not predict a relationship between WM span and judgments as a function of list.

The failure to identify irrelevant information as irrelevant could be related to source-monitoring ability, the ability to identify the origins of memories (Johnson, Hashtroudi, & Lindsay, 1993). To monitor the source of a memory, people use those characteristics that specify the conditions under which a memory was learned such as the time, social context, and spatial location in which the memory occurred, and the cognitive strategy that was used to encode the information. Perhaps under conditions of PI, one attempts to use source cues to discriminate between relevant and irrelevant alternatives. Johnson et al. (1993) noted memory traces of recent events entail more perceptual detail than do past events. In the PI paradigm, such context information might be used to discriminate among memories from current relevant list items from previous irrelevant list items. If perceptual source information is not well encoded or is too similar to the source of irrelevant information, one is more likely to retrieve irrelevant memory traces. In making judgments under conditions of PI, one must monitor those memories that originated from sources that are relevant to the judgment space (current list items) versus those which originated from sources that are irrelevant to the judgment space (previous list items), an inherent source monitoring task.

Experiment 2 attempted to replicate the findings of Experiment 1, and also attempted to extend the findings in two ways. First, participants make probability judgments rather than proportion judgments. Second, in Experiment 2 we manipulated the total frequency with which alternatives were presented on each list in order to manipulate the amount of PI. In the list frequency descending condition the total frequency with which the eight alternatives on the first list were presented was higher than the total frequency with which the eight alternatives on the second list were presented; and the total frequency with which the eight alternatives on the second list were presented was higher than the total frequency with which the eight alternatives on the third list were presented. We hypothesized presenting alternatives more frequently on the first list would increase their memory strength, making those items more likely to be retrieved. Thus, we expected that participants would experience much PI buildup in this condition because previous list alternatives were stronger in memory strength than current list alternatives and thus were more likely to be retrieved than when the frequency with which items are presented remains constant across lists. In the list frequency ascending condition, the total frequency with which the eight alternatives on the first list were presented was lower than the total frequency with which the eight alternatives on the second list were presented; and the total frequency with which the eight alternatives on the second list were presented was lower than the total frequency with which the eight alternatives on the third list were presented. We expected that participants would experience little PI buildup in this condition because previous list items were weaker in memory strength than current list alternatives and thus less likely to be retrieved than when the frequency with which items are presented remains constant across lists. We were interested in whether the amount of PI would relate to the degree of probability judgment bias. In Experiment 2 we also manipulated the number of alternatives on each list to eliminate learning effects.

Chapter 3

Experiment 2

The goals of Experiment 2 were to replicate the effect of PI on judgment, and to examine how varying the amount of PI affects the degree of judgment bias. The amount of PI was manipulated by varying the List Frequency, where List Frequency was operationalized as the total number of times alternatives were presented on the list. In this experiment there were two conditions. In the list frequency ascending condition, the number of times alternatives were presented on the first list was low (36), and the number of times alternatives were presented on each subsequent list was larger (72 times on the second list and 108 times on the third list). For instance, on the first list one of the eight alternatives was presented at each of the following frequency levels: 1, 2, 3, 4, 5, 6, 7, or 8 times respectively; on the second list one of the eight alternatives was presented at each of the following frequency levels: 2, 4, 6, 8, 10, 12, 14, or 16 times respectively; and on the third list one of the eight alternatives was presented at each of the following frequency levels: 3, 6, 9, 12, 15, 18, 21, or 24 times respectively. Note that the ratio of each item in a list to other items in its list remained constant across lists. For instance the ratio of List Frequency for the weakest item in each list to all other items in the list was $\frac{1}{35}$ for the first list, $\frac{2}{70}$ for the second list, and $\frac{3}{105}$ for the third list, all equal ratios. We hypothesized that the effect of PI on judgments would be relatively small in this list frequency ascending condition, because in this condition the alternatives on the current list were always more frequent (and stronger in memory strength) than alternatives on previous lists. In the list frequency descending condition, alternatives on list 1 were presented more often (108 times) than were alternatives on list two (72 times), and alternatives on list 2 were presented more often than were alternatives on list 3 (36 times). In this condition, on the first list one of the eight alternatives was presented at each of the following frequency levels: 3, 6, 9, 12, 15, 18, 21, or 24 times respectively; on the second list one of the eight alternatives was presented at each of the following frequency levels: 2, 4, 6, 8, 10, 12, 14, or 16 times respectively; and on the third list one of the eight alternatives was presented at each of the following frequency levels: 1, 2, 3, 4, 5, 6, 7, or 8 times respectively. It was hypothesized that the effect of PI on judgments would be relatively large in this list frequency decreasing condition because in this condition, the alternatives on the current list were always less frequent (and weaker in memory strength) than were alternatives on previous lists. WM was not measured in this experiment. We varied the number of alternatives on each list in order to reduce the effect of learning. Similar to Experiment 1, we expected discrimination failure to occur because as PI increased, participants would be unable to discriminate relevant from irrelevant alternatives. In this experiment, discrimination bias should be greatest in the list descending condition and more or less absent in the list ascending condition.

3.1 Method

3.1.1 Participants

Seventy-seven University of Maryland undergraduate students participated in this experiment for course extra credit.

3.1.2 Materials

The materials were the same as those used in Experiment 1, except that WM span was not measured and thus those materials were not used, and recall in the PI task was typed into the computer rather than spoken, and therefore the tape recorder was not used. Also, instead of using only produce items for the buildup of PI phase of the experiment and only beverages for the release from PI phase, animal words or fruit words were counterbalanced across the buildup and the release from PI phases to ensure that the effects found were not due only to the materials used. Words used were from Van Overschelde, Rawson, and Dunlosky's (2004) category norms, an updated version of the Battig and Montague (1969) word norms.

3.1.3 Design and Procedure

The design was a 2 (List Frequency: ascending or descending) X 4 (List: 1-4) X 8 (Alternative Frequency) Mixed Factorial design with List and Alternative Frequency manipulated within subjects and List Frequency manipulated between subjects. Participants were randomly assigned to one of either the List Frequency ascending or the List Frequency descending condition. The procedure was the same as that of Experiment 1 except for the following changes. First, WM was not measured in this experiment. Second, the paradigm was altered slightly in that rather than imagining going to a grocery store and buying items, participants were instructed to imagine that they would be observing items sold at different market stands of an international farmer's market. For each PI trial, participants observed what items were sold by a given person at their market stand. Participants were instructed that items sold at one person's stand were never sold at any other stands. Third, the instructions to consider only current list items when making probability judgments were emphasized by using bold text and all capital letters. Fourth, participants made probability rather than proportion judgments. For each item on a given list and for four items from a previous list participants were asked, "Out of all of the kinds of items sold at the stand you just observed, what is the probability that the next item sold at this stand will be [x]?" where x represents the item to be judged. Fifth, rather than always having 8 alternatives on each list, we added either one, two, or three extra alternatives to the buildup of PI lists so that participants would not learn that there were always 8 alternatives on each list and use this information when making probability judgments. The number of presentations of extra item(s) always summed to 5. In other words, when one extra alternative was shown in a PI buildup list, it was presented 5 times; when two extra alternatives were shown, one was presented two times and the other was presented three times; and when three extra alternatives were shown, one was presented one time, and two were presented two times. The order of the number of extra alternatives added to each list was fully counterbalanced. Sixth, no time limits were imposed for the recall phase of the task, and participants typed recalled items into the computer instead of recalling them aloud as in Experiment 1. And finally, whereas in Experiment 1 the lists were all of equal frequency, in Experiment 2, List Frequency either increased (from 36 to 72 to 108 items per list) or decreased (from 108 to 72 to 36 items per list). The release from PI list always had 72 items, distributed 2, 2, 6, 6, 12, 12, 16, 16.

3.2 Results and Discussion

The data from five participants were not included in this study for the following reasons: One participant typed words instead of giving probability responses, three participants gave judgments of only 100 or 0, and one participant's subadditivity was greater than 2.5 standard deviations above the mean on the fourth list. The participant excluded due to being an outlier was in the descending group. The mean subadditivity for the descending group for list 4 was 265.58 (SD=152.59) and the participant's subadditivity for list 4 was 720.

3.2.1 Recall

We hypothesized that buildup of PI would lead to a decrease in recall, and then at release from PI recall would increase again. Further, it was hypothesized that PI would have a stronger effect on recall in the descending list frequency condition than in the ascending list frequency condition, because in the descending condition alternatives on the previous lists had been presented more frequently than alternatives on the current list. Consequently previous list alternatives were likely to be retrieved, and consequently the descending condition provided more PI than in the ascending condition. The middle section of Table 2.1 presents the mean number of words recalled (out of eight) for each of the four lists as a function of List Frequency condition. As predicted, a main effect of List was found, $F(3, 210) = 15.86, p < 0.0001, \omega^2 = .13$, and a significant interaction between List Frequency and List was found, $F(3, 210) = 3.16, p < 0.05, \omega^2 = .02$. Further, even when individual differences on the practice list recall were used as a covariate, a significant interaction between List and List Frequency was found, $F(3, 207) = 3.00, p < 0.05, \omega^2 = .02$. These results suggest that the List Frequency manipulation worked in that participants' recall was more affected by PI in the List Frequency descending condition than in the List Frequency ascending condition. Further, trend analyses revealed that a quadratic trend fit the changes in recall well, $F(1, 70) = 42.77, p < 0.0001$.Overall, recall decreased as PI increased and then increased again on the release from PI list.

We again examined the number of intrusions of previous list items participants reported during recall. The right section of Table 2.2 presents the mean number of recall intrusions for each of the four lists as a function of List Frequency condition. There was a significant main effect of List on the number of intrusions reported, such that more intrusions were reported on the second and third lists when PI was greatest than on the first and fourth lists when PI was least present, $F(3, 210) = 8.03, p < 0.0001, \omega^2 = .07$. No main effect of List Frequency and no interaction between List Frequency and List on the number of intrusions reported were found. Overall the number of intrusions reported was low (less than 1), but more intrusions were reported when PI was greatest. Intrusions reflect a failure to discriminate between lists. The finding that some participants reported intrusions of previous list items suggests that they experienced list discrimination failure. However, fewer intrusions were reported in this experiment, when participants were not under time pressure than in Experiment 1 when participants were under time pressure. This suggests that time pressure impaired participants' ability to monitor if their recall output was a current list item or an intrusion in Experiment 1.

3.2.2 Subadditivity

It was hypothesized that a main effect of List on subadditivity would be found such that as PI increased, the subadditivity of probability judgments would decrease as in Experiment 1. Secondly, it was hypothesized that a significant interaction between List and List Frequency would be found, such that subadditivity would decrease most in the descending condition, and the effects of PI would be minimal in the ascending condition.

Figure 3.1 presents the mean subadditivity for each of the four lists as a function of List Frequency condition. As predicted, a main effect of List on subadditivity was found, $F(3, 210) = 2.94, p < 0.05, \omega^2 = .02$, and a significant interaction between List and List Frequency was found, $F(3, 210) = 13.64, p < 0.0001, \omega^2 = .12$. A main effect of List Frequency was also found, F(1, 70)=7.99, p¡.001, =.02. Participants' subadditivity differed on the practice list before PI was introduced, $t(70) = 2.45, p < 0.05$. Thus, analyses were performed using the practice list as a covariate, to examine if the effects existed independent of individual differ-

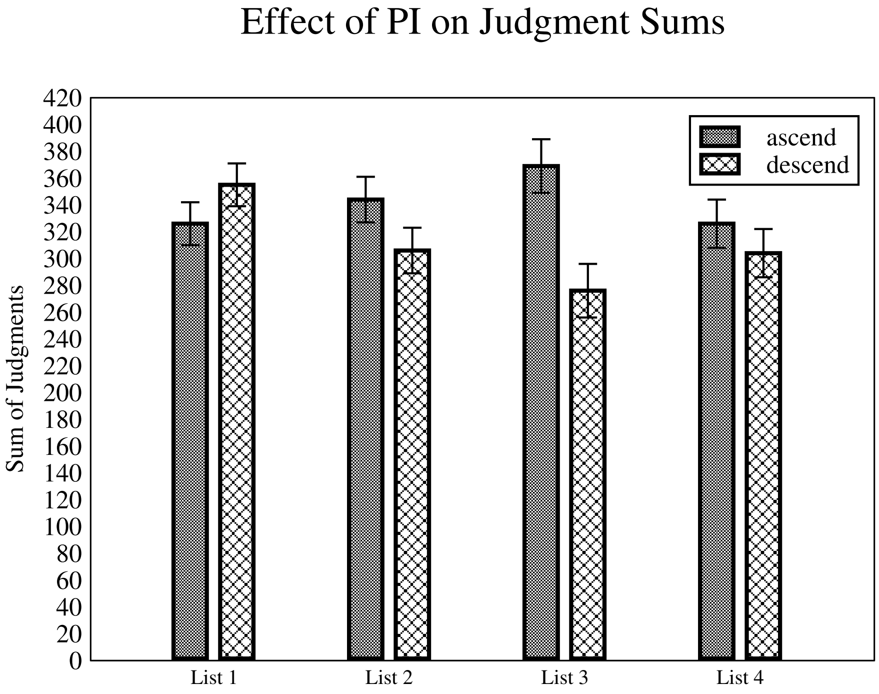

Figure 3.1: Change in judgment sums as a function of PI.

ences in subadditivity. When practice list subadditivity was used as a covariate, a significant interaction between List and List Frequency condition was still found, $F(3, 207) = 13.44, p < 0.0001, \omega^2 = .11$. However no significant main effects of List Frequency or of List were found.

The univariate analyses examining the main effect of List on subadditivity within the ascending condition was significant, $F(3, 105) = 3.53, p < 0.025$, as was the main effect of List on subadditivity in the descending condition, $F(3, 105) = 15.56, p < 0.0001$. A significant quadratic trend was found for the effect of List in the ascending condition, $F(1, 35) = 7.02, p = 0.012$. In the descending condition significant linear $F(1, 35) = 18.77, p < 0.0001$ and quadratic $F(1, 35) = 19.35, p < 0.0001$ trends were found. In the ascending condition, subadditivity significantly increased on lists 2 and 3 and then decreased again on list 4. In the descending condition, subadditivity significantly decreased on lists 2 and 3 and then increased again on list 4. These effects could have been due either to list frequency effects or due to PI. Subadditivity was greatest when list frequency was highest and was least when list frequency was lowest in the ascending and descending conditions. Thus, rather than the subadditivity effects occurring due to PI, it could be the case that subadditivity increased when the overall number of items on the list increased, and that in the descending condition, judgments decreased when the number of items on the list decreased. We address the possibility that changes in subadditivity were due to List Frequency in Experiment 3. In Experiment 3, List Frequency was manipulated but PI was not present. Thus, it was possible to examine the effects of List Frequency alone on subadditivity and compare those results to these results examining the effects of both List Frequency and PI on subadditivity.

We manipulated the number of alternatives on each list in an attempt to eliminate changes in judgment due to learning the number of alternatives on each list, and still found that subadditivity in the descending condition on the release from PI list did not increase back to levels observed on List 1. On the other hand, subadditivity could also have been affected by List Frequency. If List Frequency affected subadditivity, the finding that subadditivity on the release from PI list was not closer to original subadditivity levels on list 1 could be explained, since the List Frequency on list 4 was less than that of list 1 in the descending condition.

3.2.3 Relative Accuracy

It was hypothesized that Somers's D correlations between participants' judgments and objective probabilities would decrease as PI increased, and that the effect of PI would interact with the List Frequency manipulation, in that correlations would decrease more as PI increased in the descending condition than in the ascending condition because PI was greater in the descending condition.

Figure 3.2 presents the mean of participants Somers's D correlations for each of the four lists as a function of List Frequency condition. As predicted, a main effect of List on relative accuracy was found, $F(3, 210) = 47.40, p < 0.0001, \omega^2 = 0.32$. Further, trend analyses revealed that both a linear trend, $F(1, 70) = 8.38, p < 0.005$ and a quadratic trend, $F(1, 70) = 14.70, p < 0.0001$ fit the changes in relative accuracy. Relative accuracy was highest on lists 1 and 4 when PI was less present and was lowest on lists 2 and 3 when PI was most present. However, no interactions between List and List Frequency were found, nor was a main effect of List Frequency on relative accuracy found. Further, even when individual differences in relative accuracy were controlled for by using the practice list Somers's D correlations as a covariate, a marginally significant main effect of List was still found, $F(3, 207) = 2.36, p = 0.072, \omega^2 = 0.02$. Thus, as predicted discrimination failure occurred as a result of PI causing participants' relative accuracy to decrease as PI increased. However, in contrast with our predictions, relative accuracy did not decrease more in the descending list frequency condition than in the ascending list frequency condition.

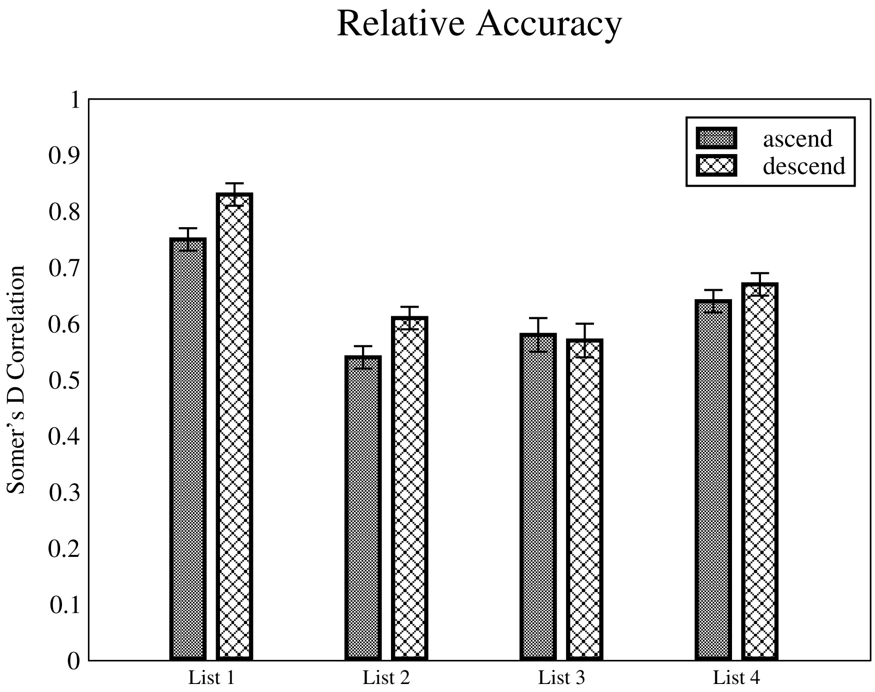

Figure 3.2: Change in relative accuracy as a function of PI.

3.2.4 Irrelevant Items

Further evidence for discrimination failure comes from an examination of participants' judgments of previous list items, irrelevant items to which judgments of zero are optimal. We hypothesized that as PI increased on the first three lists, the number of irrelevant items given nonzero judgments would increase, and then at release from PI the number of irrelevant items given nonzero judgments would decrease again. Further, we hypothesized that these effects would be stronger in the descending condition than in the ascending condition. The middle section of Table 2.3 presents the mean number of previous list items (out of 4 possible) given nonzero judgments for each of the four lists as a function of List Frequency condition. As in Experiment 1, a main effect of List was found, $F(3, 210) = 25.75, p < 0.0001, \omega^2 = .20$. Participants judged significantly more irrelevant items as relevant on lists two and three when PI was greatest than on lists one and four. No significant interaction between List and List Frequency were found, nor was a main effect of List Frequency found on the number of irrelevant items given nonzero judgments.

3.2.5 Reaction Time

We hypothesized that judgments would be slowest when PI was greatest on lists two and three, and that judgments would be fastest when PI was least on lists one and four. Reaction times for each judgment were transformed to rates $\frac{1}{RT}$ to reduce the skew-ness of their distribution.

Figure 3.3 presents the mean judgment RT rate for each of the four lists. As in Experiment 1, a main effect of List on participants' average judgment rate was found, $F(3, 210) = 48.78, p < 0.0001, \omega^2 = 0.33$. Judgments were slower when PI was present (on lists 2 and 3) than when PI was not present (on lists 1 and 4). Figure 3.4 presents the mean irrelevant judgment RT rate for each of the four lists. A significant main effect of List on the reaction time to judge irrelevant items (items which did not occur on the list in question) was also found, $F(3, 210) = 72.26, p < 0.0001, \omega^2 = 0.43$. Participants were slower at judging irrelevant items on lists 2 and 3 which had the most interference than to judge lists 1 and 4. Neither an interaction between List and List Frequency nor a main effect of List Frequency for the average judgment rate or for the average irrelevant judgment rate was found. The finding that participants took longer to make probability judgments on lists in which PI was greatest suggests that more processing was necessary for those lists.

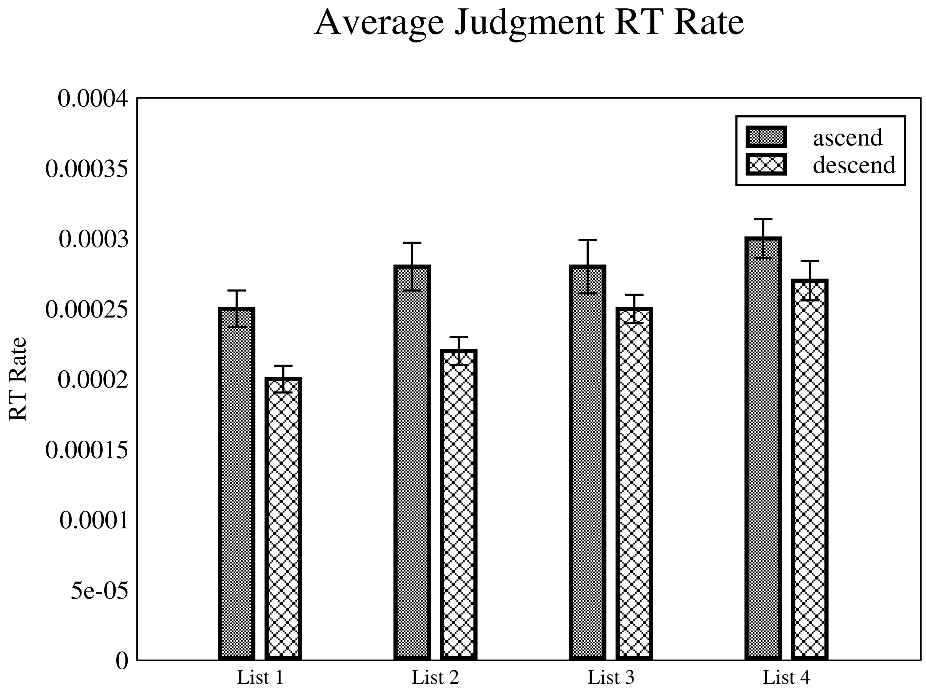

Figure 3.3: Change in average judgment RT rate as a function of PI.

Again, perhaps the finding that the rate was slower when PI was greatest reflects that discrimination between current list information and previous list information became increasingly difficult with the build-up of PI.

Experiment 2 replicated three main findings consistent with the discrimination failure account. First, in the descending condition, as PI increased, judgments decreased, consistent with the account that participants failed to discriminate relevant from irrelevant alternatives and that irrelevant alternatives supplanted relevant alternatives in the probability judgment. Importantly, on the release from PI list judgments increased, though they did not increase back to the level that was observed on the first list. The second finding consistent with the discrimination failure account was that relative accuracy decreased as PI increased in both of the List

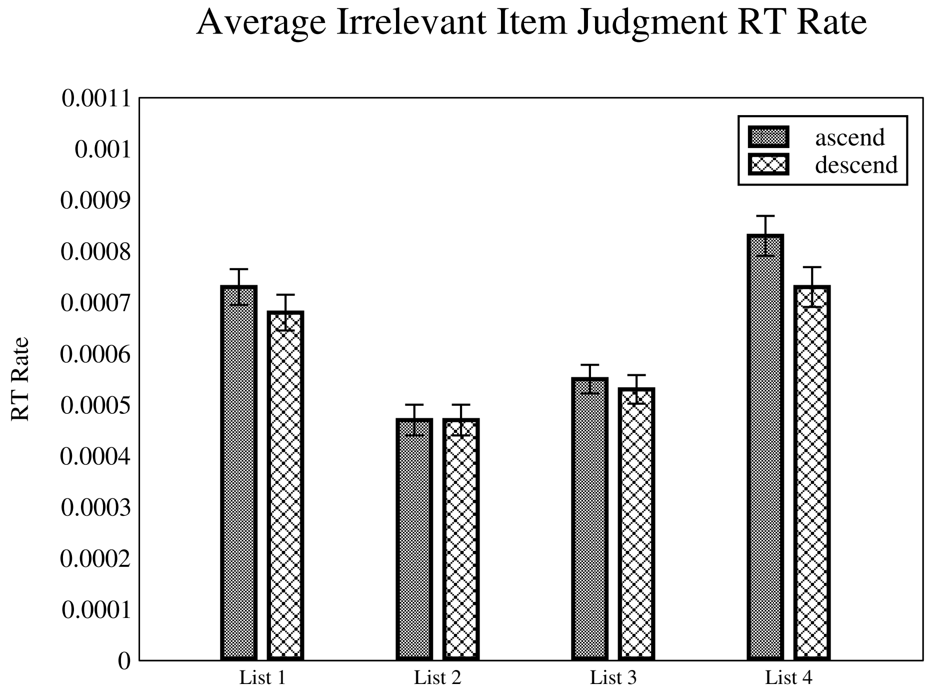

Figure 3.4: Change in average irrelevant judgment RT rate as a function of PI.

Frequency ascending and descending conditions, but increased on the release from PI list. This could be due to increased variance in the assessment of the support for alternative hypotheses as PI increased. The third main finding consistent with the discrimination failure account was that as PI increased, participants gave irrelevant, previous list items nonzero judgments, suggesting that they failed to identify irrelevant alternatives as irrelevant.

However in contrast with our predictions, the amount of PI (as measured by List Frequency condition) did not affect the amount of change in relative accuracy, RT rate, or in the number of nonzero judgments given to irrelevant previous list items. For recall, stronger preceding lists in the descending condition appeared to have caused more interference than did weaker preceding lists in the ascending condition as recall reduction as a function of PI was worse in the descending than in the ascending condition. However relative accuracy, judgments of irrelevant items, and RT rates were unaffected by List Frequency condition. Further, we found that subadditivity increased as PI increased in the ascending condition, an unexpected finding. Perhaps the increased subadditivity in the ascending condition was due to effects of List Frequency, rather than due to PI. When the List Frequency was greatest, subadditivity was greatest, and when List Frequency was least, subadditivity was least. Perhaps participants focused more on the focal hypothesis than on each of the alternative hypotheses as predicted by support theory (Tversky & Koehler, 1994). Because the focal items are all stronger in the strongest (108 item) list than in weaker (72 or 36 item) lists, participants gave larger judgments to each item on the strongest list causing subadditivity to be greatest. However this hypothesis could not account for the findings that relative accuracy decreased as PI increased, or that participants incorrectly judged irrelevant items as being relevant. To separately examine the effects of List Frequency on probability judgment, Experiment 3 manipulated List Frequency, but did not manipulate PI. In this way, it was possible to examine the effect of List Frequency on probability judgment separately from the effects of PI on judgment.

Chapter 4

Experiment 3

The purpose of Experiment 3 was to examine the role of List Frequency when PI was not present. We predicted that if List Frequency alone caused the changes in subadditivity in Experiment 2, in the ascending condition judgments should increase as List Frequency increased, and in the descending condition judgments should decrease as List Frequency decreased. However, if the findings of Experiment 2 were due to PI, then Experiment 3 should find no changes in subadditivity as a function of List Frequency.

4.1 Method

4.1.1 Participants

Fifty-eight University of Maryland undergraduate students participated in this experiment for course extra credit.

4.1.2 Materials

Materials used in Experiment 3 were the same as those used in Experiment 2 except that for each list participants saw different categories of words. Eight words each from the categories spices, tools, fabrics, flowers, and beverages from Van Overschelde et al.'s (2004) category word norms were used.

4.1.3 Design and Procedure

The design was a 2 (List Frequency: ascending or descending) X 4 (List: 1-4) X 8 (Alternative Frequency) Mixed Factorial design with List and Alternative Frequency manipulated within subjects and List Frequency manipulated between subjects. Participants were randomly assigned to List Frequency condition. The procedure followed that of Experiment 2 exactly except for the following two changes. First, each market stand sold a different category of items, to eliminate PI. Second, no extra items were added to each list.

4.2 Results and Discussion

Data from two participants were excluded from the analysis. One participant's subadditivity was greater than 2 SD above the mean, and the other excluded participant reported that s/he had made frequency judgments rather than probability judgments. The participant excluded due to being an outlier was in the descending condition. The mean subadditivity for participants in the descending condition for each list was: 337.18 (SD=155.51) for list 1; 301.14 (SD= 167.87) for list 2; 265.29 (SD=156.93) for list 3; and 301.14 (SD=167.87) for list 4. The participant's subadditivity for each list was: 635 for list 1; 678 for list 2; 530 for list 3; and 600 for list 4.

4.2.1 Recall

It was hypothesized that recall would not be affected by List or by List Frequency since PI was not present. The right section of Table 2.1 presents the mean number of words recalled (out of eight possible) for each of the four lists as a function of List Frequency condition. In line with our predictions, no main effects of List Frequency or of List on recall were found. However, a significant interaction between List Frequency and List was found, $F(3, 162) = 3.11, p < 0.05, \omega^2 = .03$. Univariate analyses (Bonferonni adjusted) revealed an almost significant effect of List on recall occurred within the ascending condition $F(3, 81) = 3.08, p = 0.032$, but not within the descending condition. In the ascending condition, recall increased as List Frequency increased; and recall decreased again when List Frequency decreased (the opposite pattern of that found when PI is present). Perhaps the increased recall on lists that were stronger was due to having more learning trials for each item. However the finding that overall recall did not decrease as a function of List when no PI was present confirms that previous recall trends in Experiments 1 and 2 were due to effects of PI.

4.2.2 Subadditivity

It was hypothesized that if List Frequency alone caused differences in subadditivity in Experiment 2, then in Experiment 3 a significant interaction between List and List Frequency would be found such that subadditivity would increase as List Frequency increased in the ascending condition and would decrease as List Frequency decreased in the descending condition. In contrast, if PI caused the changes in subadditivity found in Experiment 2, then subadditivity should be unaffected by List Frequency or List in Experiment 3.

Figure 4.1 presents mean subadditivity for each of the four lists as a function of List Frequency condition. A main effect of List on subadditivity was found, $F(3, 162) = 11.29, p < 0.0001, \omega^2 = .12$ which was qualified by a significant interaction between List and List Frequency, $F(3, 162) = 3.46, p = 0.018, \omega^2 = .03$. Univariate analyses revealed a significant effect of List in the descending condition, $F(3, 81) = 16.13, p < 0.0001$, but no effect of List in the ascending condition. As List Frequency decreased in the descending condition, subadditivity decreased and then increased again in list 4 when List Frequency increased again. No main effect of List Frequency was found.

4.2.3 Relative Accuracy

We hypothesized that correlations between objective frequencies and subjective judgments would not change as a function of List or of List Frequency because no PI was present to affect relative accuracy. As predicted, no main effect of List Frequency condition was found, nor was an interaction between List and List Frequency found.

Figure 4.2 presents the mean of participants' Somers's D correlations for each of the four lists as a function of List Frequency condition. However, a significant main effect of List was found, $F(3, 162) = 3.05, p = 0.030, \omega^2 = .03$. However, the changes in relative accuracy as a function of List did not follow a quadratic

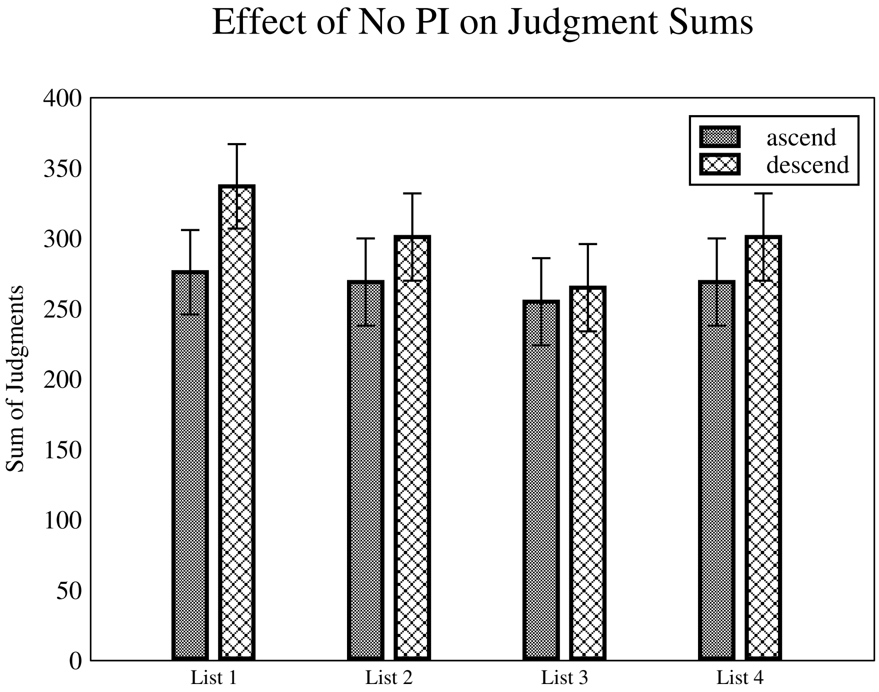

Figure 4.1: Change in judgment sums as a function of PI.

trend, as would be predicted if PI affected the correlation results. Further, using Bonferonni adjustments, relative accuracy on list 1 did not differ significantly from relative accuracy on list 2, 3, or 4. These results support that the changes in relative accuracy in Experiments 1 and 2 were due solely to PI, as when PI was not present, relative accuracy was unaffected by List in the manner that it was affected when PI was present.

4.2.4 Irrelevant Alternatives

It was hypothesized that because no PI was present in this experiment, no main effects or interaction between List and List Frequency on the number of irrelevant items given nonzero judgments would be found. The right section of Table 2.3

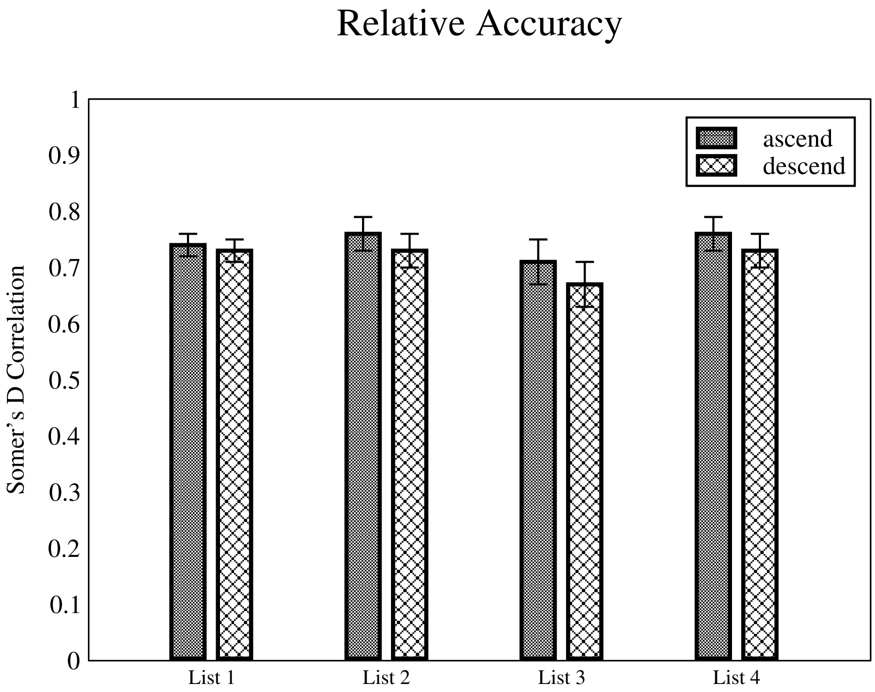

Figure 4.2: Change in relative accuracy as a function of PI.

presents the mean number of previous list, irrelevant items given nonzero judgments (out of four possible) for each of the four lists as a function of List Frequency condition. No main effect of List Frequency was found, however a main effect of List was found $F(3, 162) = 4.21, p < 0.01, \omega^2 = 0.04$ such that more participants gave nonzero judgments on list 1 than on following lists. Although this effect was not predicted, it does not follow the pattern found in Experiments 1 and 2 in which more nonzero judgments of irrelevant items were given in lists 2 and 3 when PI was greatest than on lists 1 and 4 when PI was least present. Overall the number of nonzero judgments in Experiment 3 was lower than in Experiments 1 and 2 in which PI was present. In fact, only 2 out of 56 participants gave nonzero judgments to irrelevant items on lists 2 and 3 in Experiment 3, in comparison with 44 out of 106 participants in Experiment 1, and 28 out of 72 participants in Experiment 2.

4.2.5 Reaction Time

It was hypothesized that RT rate would be unaffected by List or by List Frequency.

Figure 4.3 presents the mean judgment RT rate for each of the four lists. No effect of List Frequency on RT rate was found, nor was an interaction between List Frequency and List on RT rate found. However, a main effect of List on RT rate was found, $F(3, 162) = 4.28, p < 0.05, \omega^2 = .04$. In general, participants made judgments faster in lists 2 and 3 than in lists 1 and 4. This trend opposes the trend found in Experiments 1 and 2 that participant's judgments were faster in lists 1 and 4 than in lists 2 and 3 when PI was greatest. Figure 4.4 presents the mean irrelevant judgment RT rate for each of the four lists. A significant main effect of List on the irrelevant judgment RT rate was also found, $F(3, 162) = 11.38, p < 0.0001, \omega^2 = .12$. Participants made judgments faster in lists 2 and 3 than in lists 1 and 4.

Experiment 3 clarified some aspects of the findings of Experiment 2, however some questions still remain. On one hand, Experiment 3 revealed different patterns for relative accuracy and for the number of irrelevant items judged as relevant (given non-zero judgments) when PI was not present than were found in Experiments 1 and 2. These results are consistent with the account that the changes in relative accuracy and in the number of irrelevant items judged relevant in Experiments 1 and 2 were due to PI. When PI was present, participants' relative accuracy decreased, their judgment RT rate slowed, and they gave nonzero judgments to irrelevant previous list items. In contrast, when PI was not present, relative accuracy did not

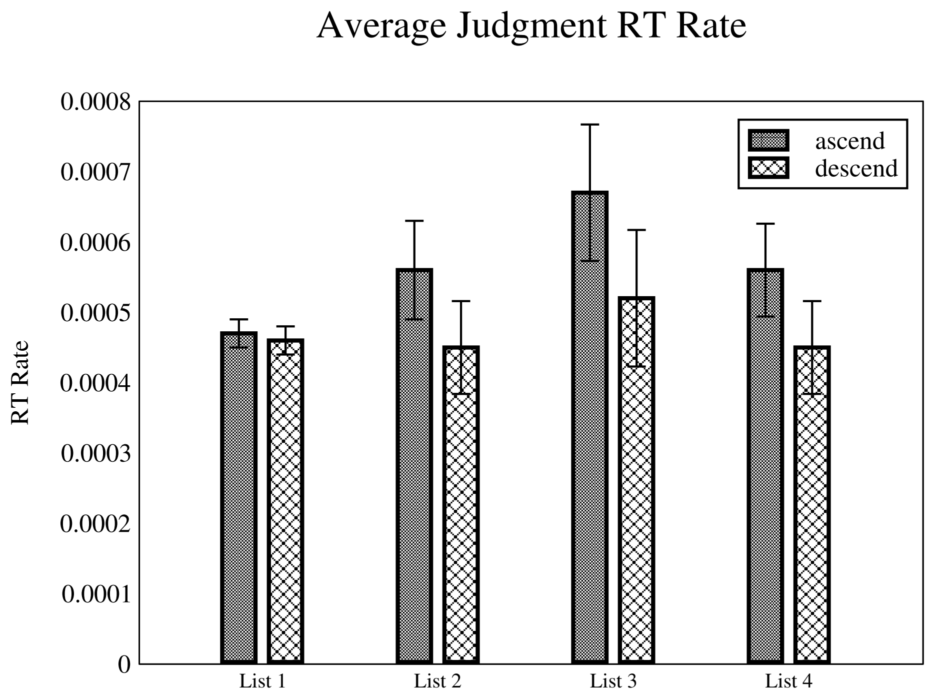

Figure 4.3: Change in average judgment RT rate as a function of PI.

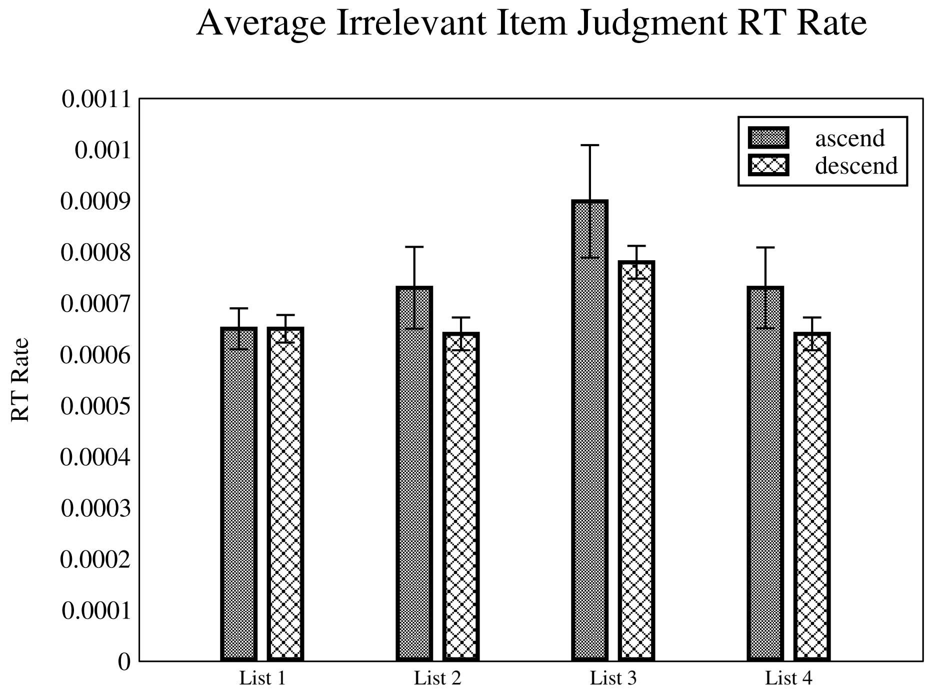

Figure 4.4: Change in average irrelevant judgment RT rate as a function of PI.

decrease for lists 2 and 3, the number of irrelevant items judged relevant did not increase for lists 2 and 3, and RT rate did not slow for lists 2 and 3. On the other hand, the subadditivity results in Experiment 3 did not fully distinguish whether discrimination failure or List Frequency caused changes in subadditivity in Experiment 2. In the ascending condition, the results supported the discrimination failure hypothesis in that subadditivity was unaffected by List Frequency when PI was not present. However, in the descending condition, the List Frequency hypothesis was supported in that as List Frequency decreased, subadditivity decreased and then when the List Frequency increased in list 4, subadditivity increased again.

Chapter 5

General Discussion

Our goal in this paper was to examine the effect of generating irrelevant alternative hypotheses on judgments of probability. We proposed two theoretical accounts of the effect of generating irrelevant alternatives on judgment: the inhibition failure account and the discrimination failure account. The discrimination failure account predicted that as PI increased, participants would generate irrelevant alternative hypotheses and will fail to recognize that the alternatives were irrelevant. Participants would then include those irrelevant alternatives in their assessment of support for the set of alternative hypotheses. Further, we hypothesized that in most cases, but not always, the non-discriminated, irrelevant alternatives would be stronger than the relevant alternatives they supplanted. Thus, this account predicted that subadditivity would decrease as PI increased. The discrimination failure account also predicted that as PI increased, the assessment of support for the set of alternative hypotheses would entail more error variance, because the discrimination failure process is stochastic in nature. For instance, whether or not irrelevant alternatives are generated is stochastic, whether or not those irrelevant alternatives that are generated are correctly discriminated is stochastic, and whether or not alternative hypotheses that are generated and not discriminated are stronger than, equal to, or weaker than the relevant alternatives they replace is a stochastic. In combination, the discrimination failure account predicted that as PI increased variability in assessing the support for the set of alternatives would increase. Consequently, as PI increased judgments would vary more and the degree to which subjective judgments would correlate with objective probabilities (relative accuracy) would decrease. Finally, the discrimination failure account predicted that as PI increased, and the number of irrelevant alternatives judged relevant would increase. In contrast, the inhibition failure account posited that when participants retrieved irrelevant alternative hypotheses, they would correctly identify them as relevant, but would fail to inhibit them from WM. Consequently, fewer attentional resources would be available to generate further relevant alternative hypotheses. Generating fewer relevant alternative hypotheses, in turn, would lead to increased subadditivity, no changes in relative accuracy, and few irrelevant alternatives given nonzero judgments.

Across three experiments, the results supported the discrimination failure account. As PI increased, subadditivity decreased in Experiment 1 and in the descending List Frequency condition of Experiment 2. However, subadditivity also appears to have been affected by other factors such as learning effects and List Frequency, which will be further discussed below. In Experiments 1 and 2 we found that relative accuracy decreased as PI increased, and increased again when PI was removed at the release from PI list. Further, in Experiment 3 when PI was not present, relative accuracy on lists 2 and 3 was not lower than relative accuracy on lists 1 and 4, supporting our claim that changes in relative accuracy in Experiments 1 and 2 was due to discrimination failure. The discrimination failure account predicted that people would fail to recognize some previous list items as irrelevant, and consequently when asked to make judgments about previous list items, people would give some irrelevant items nonzero judgments indicating that they did not recognize the items were irrelevant. In fact, in Experiments 1 and 2 we found that participants correctly identified irrelevant items as irrelevant on lists 1 and 4 when PI was not present, but that they failed to identify irrelevant items as irrelevant on lists 2 and 3 when PI was present. Further, in Experiment 3, we found that when PI was not present, only 2 out of 56 participants failed to identify irrelevant items as irrelevant, suggesting that discrimination failure did not occur when PI was not present, as we predicted. Finally, the finding in Experiment 1 that WM span did not moderate the effects of PI on judgment suggests that inhibition failure did not occur. If inhibition failure was a factor, we would predict that high spans would be less affected than would low span participants, because high spans have been found to better suppress irrelevant information than low spans (Rosen & Engle, 1997; Kane & Engle, 2000; Lustig, May, & Hasher, 2001; May, Hasher, & Kane, 1999; Carretti, Cornoldi, & De Beni, 2004; Engle, Conway, Tuholski, & Shisler, 1995).

It is likely that multiple factors affected subadditivity in the three experiments, and only when all of these factors are considered together can the subadditivity effects be explained. First, participants learned about the task (i.e., the number of alternative hypotheses on each list) as they experienced each trial, and this learning caused subadditivity to decrease with each list. Fox and Rottenstreich (2003) argued that when participants make probability judgment, at times they partition the probability space by the number of alternatives they consider, as though they are ignorant of the prior probabilities. Perhaps in the present study, participants learned about the number of items on each list with each trial, and this learning caused their judgments to decrease as they better partitioned their judgments across more alternatives. It is not argued that they only used partitioning to make judgments, but rather that with learning their judgments decreased because they became more aware of the number of alternatives in each list. Perhaps they used the partitioned judgment as an anchor and then adjusted the judgment based on the relative support for the focal to the alternative hypotheses. This kind of learning was not a factor in Experiment 2 in which we added 1, 2, or 3 extra items to lists 1-3. Second, discrimination failure caused subadditivity to decrease as PI increased and then to increase with the release from PI list. Third, List Frequency caused subadditivity to be largest when the total number of presentations in a list was largest, and caused judgments to be smallest when the total number of presentations in a list was smallest. Thus, in ascending conditions List Frequency caused judgments to increase for the first 3 lists and then decrease with the final list, and in the descending conditions List Frequency effects caused judgments to decrease for the first 3 lists and then to increase with the final list.

Although the effects of learning, discrimination failure, and List Frequency are impossible to tease apart in our three experiments, we will provide a cursory explanation of how they might explain the subadditivity results of the three experiments. In Experiment 1, discrimination failure and learning lead to decreases in subadditivity across the first 3 lists, and in the final list RPI (and thus the elimination of discrimination failure) led to increases in subadditivity, but the increases were not back to list 1 levels because learning mitigated the RPI effect. Note that in

Experiment 2, learning may not have played as large a role in affecting subadditivity as in Experiments 1 and 3 because the number of alternatives presented in lists 1 through 3 varied, and thus participants could not learn how many alternatives were on each list. The number of alternatives on one list was not the same as the number of alternatives on the next list. In the descending List Frequency condition of Experiment 2, discrimination failure due to PI led to decreases in subadditivity across the first 3 lists, and then List Frequency and RPI led to increases in subadditivity for the fourth list, but List Frequency kept subadditivity from increasing to pre-PI levels. In the ascending condition, however, discrimination failure led to decreases in subadditivity, but the decreases were overcome by the List Frequency effect which drove subadditivity to increase across the first three lists. Perhaps the List Frequency effect overcame the discrimination failure effect because PI was not as strong as in the descending condition. Then at release from PI, List Frequency led to decreases in subadditivity, and release from PI led to increases in subadditivity. In actuality, subadditivity increased, but not back to pre-PI levels. In the ascending list of Experiment 3, learning led to decreases in subadditivity across all 4 lists, but List Frequency led to increases in subadditivity across the first 3 lists and then to decreases in subadditivity on the final list. In combination, the two factors were in opposition, and may account for the finding that subadditivity did not vary across lists. In contrast, in the descending List Frequency condition, both List Frequency and learning tended to cause subadditivity to decrease across the first three lists, and then in the final list, List Frequency effects caused judgments to increase, but learning effects caused judgments to decrease. In combination, these effects caused subadditivity to increase, but not back to the level of list 1.

In addition to our main findings, a few unexpected, but curious results were obtained. For example, we found that judgment magnitude and relative accuracy could be affected independently. In the ascending condition of Experiment 2, judgment magnitude (subadditivity) increased on lists 2 and 3 and then decreased on list 4 and relative accuracy decreased on lists 2 and 3 and then increased on list 4. On the other hand, in Experiment 1 and in the descending condition of Experiment 2, judgment magnitude decreased on lists 2 and 3 and then increased on list 4, and relative accuracy decreased on lists 2 and 3 and then increased on list 4. This is interesting because one might assume that relative accuracy and judgment magnitude (absolute accuracy) would both improve in good conditions, and would both deteriorate in poor conditions. We found that in some cases (the ascending condition of Experiment 2) this line of reasoning holds, but in other cases (the descending condition of Experiment 2 and in Experiment 1) this line of reasoning fails. In the descending condition of Experiment 2 and in Experiment 1, relative accuracy decreased on lists 2 and 3 while absolute accuracy (subadditivity) appeared to increase. Thus participants may appear more accurate in the absolute sense, but examination of their relative accuracy reveals that performance is decreasing.

One may question why we found no relationship between WM and PI in Experiment 1, whereas previous studies found a significant relationship between WM span and subadditivity of probability judgment. It is possible that the relationship Dougherty and Hunter (2003) found between WM span and subadditivity of probability judgments was due to inhibition failure, rather than to differences in the number of items participants generated and compared in their judgment. Participants in Dougherty and Hunter's study observed items that four different people ate each day for breakfast, lunch, snack, or dinner respectively. Later, they judged the likelihood that on the next day, a person would eat an item from a given meal. When making judgments, irrelevant alternative hypotheses from other meals could have been retrieved and caused interference when making judgments, but the irrelevant items would have been easily discriminated as irrelevant. It is easy to recognize that one has generated a lunch item when one should be considering only breakfast foods. In such a situation, dysinhibition bias could still occur if the irrelevant item was not inhibited. Thus, Dougherty and Hunter's finding that low spans' subadditivity was higher than high spans' subadditivity could have occurred due to inhibition failure. High spans were better able to inhibit irrelevant items than low spans, and therefore had more resource available to generate and compare alternative hypotheses in their judgment than did low spans. Considering more alternative hypotheses caused judgments to be lower because the relative support for the focal hypothesis to the set of alternative hypotheses was lower. Consequently, one reason that high and low span's judgments differed could be due to differential ability to inhibit irrelevant items that were retrieved.

5.1 Relating Findings to PI Explanations

Two main explanations of forgetting due to PI exist: list differentiation (also known as temporal discrimination theory) in which the ability to discriminate items that appeared on the most recent list from those that appeared on earlier lists decreases as PI increases and causes forgetting (Underwood, 1945); and response competition in which responses learned within the same situation compete with each other for recall at the time of test if irrelevant hypotheses are not suppressed. The results of the studies reported here support the list differentiation account of PI. Previous studies manipulating PI supported the list discrimination hypothesis, as well. Dallett and Wilcox (1968) manipulated the context (the location and surrounding visual stimuli) in which participants learned lists of words and found that PI was reduced, but not eliminated by changing the context between lists. This finding suggests that one aspect that causes PI is that participants cannot discriminate among the various lists. Changing contexts between lists improves the ability of participants to discriminate between the two lists, as more cues are available to discriminate between items from different lists. However, the finding that changing contexts did not completely eliminate PI suggests that factors other than discrimination ability also affect one's susceptibility to PI. Reutner and Graybeal (1978) found that when participants were shown words from a given taxonomic category for three buildup of PI trials and then were shown pictures from the same taxonomic category for fourth list, release from PI occurred such that participants recall increased almost back to original pre-PI recall. However, greater release from PI was found when participants saw color pictures of items from one taxonomic category for the buildup of PI lists and then saw color pictures of items from a different category for the release from PI list. Thus, when lists are identifiable by additional context cues, PI is reduced due to the increased discrimination ability, however list discrimination does not completely account for PI effects as changing cues do not completely eliminate forgetting due to PI.

Glenberg and Swanson (1986) developed search-set theory to explain how temporal context can be used to cue recall. Glenberg and Swanson argued that recall based on temporal cues requires the creation of a search set that specifies the time at which an event occurred. At recall, participants sample from the search set. Further, they argued that the retention interval determines the size of the search set. With short retention intervals, many specific, local temporal contextual components are present and are uniquely associated with the most recent list items. Those specific, local components are used to construct a (small) search set that consists of mainly the relevant list items. After a long retention interval, however, the specific, local components of the temporally changing context are no longer present, and consequently more stable, general temporal contextual components that were present during list presentation (and that are still present after a long retention interval) are used to establish a search set. However, these contextual features are associated with other events in addition to the original items, and consequently the search set created by the cues is larger. Further, they assumed that the more items included in a search set, the less likely any individual item in the set is recalled. Wixted and Rohrer (1993) found that recall latency associated with buildup of PI increased due to participants searching larger sets of items for recall, supporting Glenberg and Swanson's theory. Note that in our experiments, because participants were learning each word in the list multiple times, the time to present items was longer than in normal PI experiments. In our study list discrimination failure occurred because of the longer times between each list's recall. With the longer retention intervals, temporal cues led to retrieval of larger sets of items, including irrelevant items, because more general cues were used to determine search sets. Because no other cues, such as semantic, visual, or auditory cues, were available to discriminate between lists, and the temporal cues were less useful with each additional list, discrimination failed and affected performance. Then at release from PI, semantic cues were available to distinguish between lists, causing performance to increase again.

Although our findings support the list differentiation account of PI, we argue that list differentiation is not the only factor that causes forgetting due to PI. Dillon and Thomas (1975) found that when participants reported intrusions, they were able to identify which items were the intrusions, suggesting that participants are able to discriminate relevant from irrelevant list items. However, we argue that any reported intrusions, even a small number of intrusions that later can be identified as such, suggest that list differentiation has failed. Some PI studies have found results that suggest that response competition and failure to suppress competing information plays a role in forgetting due to PI. Postman and Hasher (1972) argued that output interference (which is similar to response competition) accounts for forgetting due to PI with free recall. They argued that recall of items from different lists depends on different sets of retrieval cues, and thus if the recall of relevant list items is interrupted by recall of a previous list item, the previous list item activates retrieval cues that are better for retrieving previous list recall than for current list recall. Postman and Hasher found that forgetting due to PI was greater when participants were informed that they would later be tested on a previous list, which, they argued, increased competition from alternative response sets. Perhaps under normal circumstances, participants attempt to inhibit previous list items, but when instructions are presented that participants will need to recall previous list items later, participants do not inhibit those items. In consequence, the previous list items compete with current list items more at the time of recall than if the instructions were not given. Kane and Engle (2000) found that participants high in WM span, who have been shown in other studies to have more capacity to inhibit task-irrelevant information, were less susceptible to the effects of PI than were participants low in WM span. This finding suggests that response competition arises and some participants are better than others at inhibiting previous list items to keep them from competing with current list items for recall. Kane and Engle argued that previous literature has found no differences in the ability of high and low spans to encode or retrieve context information, and consequently the finding that high and low spans differ in their susceptibility to PI support the response competition theory and not the list discrimination theory. Bjork, LaBerge, and LeGrand (1968) found that instructions to forget previous list items reduced forgetting due to interference, suggesting that inhibition of irrelevant information plays a role in ones susceptibility to PI. (Note, though, that Coccia and Wickens (1976) failed to replicate Bjork et al.'s findings when they used Peterson and Peterson methodology rather than cued-recall methodology to test the effect of instructions to ignore previous lists.)

Based on previous findings in combination with our findings, it appears that two factors account for forgetting due to PI: list discrimination failure and response competition (which is related to inhibition ability). In fact, Postman and Hasher

(1972) argued that the two mechanisms are not independent of each other. Although we argue that both list differentiation and response competition account for forgetting due to PI, interestingly we found no support for the response competition account of PI in our study. If previous list items were retrieved and made retrieving relevant list items more difficult because of competition from the previous list item, when judging probability this effect should cause fewer alternative hypotheses to be generated and included in the judgment comparison. Including fewer alternative hypotheses causes judgments to increase (Dougherty & Hunter, 2003; Tversky & Koehler, 1994). Further, a relationship between WM and subadditivity should then be found, as high spans are better able to inhibit irrelevant items that are retrieved in order to continue retrieving relevant items, which has been shown to increase their overall recall (Rosen & Engle, 1997). We found no such relation. Thus, our findings of discrimination failure may suggest that if items cannot be discriminated, they cannot be inhibited, and in our studies discrimination failure played a larger role than did response competition and failure to inhibit competing items.

5.2 Future Directions

In the future, a study determining how PI affects probability judgments when people can discriminate between relevant and irrelevant lists is recommended. In the present study, PI caused participants' judgments to decrease due to discrimination failure. Furthermore, in Experiment 1 no differences between high and low spans' judgments were found. In Experiment 1, inhibition was not a relevant process, because it presupposes discrimination. In the future, an experiment could be conducted which increases discrimination ability by increasing the amount of context information associated with each list. For each list, a picture could represent the context. In the PI paradigm, discriminating relevant from irrelevant information can be likened to discriminating sources in source monitoring paradigms where list is the source variable. When monitoring the source of a memory, context information is used to discriminate among memories from various sources (Johnson, Hashtroudi, & Lindsay, 1993). Thus in this experiment, including different contexts (pictures) for each list could increase participants' ability to discriminate current relevant lists from previous irrelevant lists. Recall that in Dallett and Wilcox's (1968) study, they manipulated the context in which participants learned lists, and found that changing contexts between lists reduced, but did not eliminate, the amount of forgetting due to PI. Thus, it appears that PI is affected by discrimination ability, but is not only affected by discrimination ability. When lists are identifiable by additional context cues, some interference among lists is still present causing PI affects above and beyond those presents when the lists cannot be discriminated as well.

It is hypothesized that dysinhibition bias would occur in the proposed experiment in which additional context cues are present, because participants would retrieve irrelevant information and be able to identify irrelevant information as irrelevant due to the added context information, but might be unable to inhibit the irrelevant information. Then, the irrelevant information would act as a placeholder in WM, diminishing the resources available to retrieve and compare further alternative hypotheses. It is hypothesized that high spans would be less affected by PI than would low spans in this experiment. High spans would be better able to suppress irrelevant items that are retrieved, and therefore would have more WM resources available to retrieve and process further relevant items. Contrarily low spans would perseverate on irrelevant retrieved items, due to failure to inhibit the irrelevant items. The perseveration would utilize WM resources that would otherwise be used to retrieve and compare other alternative hypotheses.

We also propose conducting a study examining which alternative hypotheses are active when participants make judgments under PI conditions. To examine which alternatives are active, a future study could use a lexical decision task to see which alternative items are primed directly after making probability judgments. The lexical decision task would require participants to monitor words on a screen and respond with one key if a word was presented and another key if a non-word was presented. In order to see if an improper response word (i.e. an alternative from a previous list) was active after making judgments, participants would be required to respond to the lexical decision task to one of four kinds of words: words from the previous list, words from the current lists, task-unrelated words, and non-words. The reaction time for the lexical decision would be measured. If only relevant items were primed, the responses to the lexical decision task should be fastest for the relevant items and slower for irrelevant items. For this experiment, differences between high and low spans would also be measured.

A future study should further examine the effects of List Frequency on the subadditivity of probability judgment. It appears from this study that List Frequency has an affect on subadditivity, but it is not clear whether the effect is due to between list effects or due to within list effects. The between list account of List Frequency effects would argue that participants' subadditivity increased or decreased relative to a previous list which is used as an anchor for list 2 judgments. If more items were on a previous list than a current list, participants' subadditivity decreased. If fewer items were on a previous list than on a current list, participants' subadditivity increased. In contrast, List Frequency could affect subadditivity due to a within list factor. For instance, if participants focus on the focal item and discount the alternative hypotheses, stronger items on a given list would lead to more subadditivity because people overweight the focal items. To test these two accounts of List Frequency effects on subadditivity, a future experiment could have participants learn and judge two successive lists of items. The first list would have either 108 items or 36 items, and the second list would always have 72 items. According to the between list account of List Frequency effects, participants should have equal subadditivity on the first list because nothing comes before it to anchor their judgments. Then on list two, subadditivity in the 108-72 condition would decrease because the previous list was stronger than the second list and subadditivity in the 36-72 condition would increase because the previous list was weaker than the second list. Contrarily, the within list account of List Frequency effects would argue that participants' judgments on the first list would differ such that participants who see 108 items should give higher judgments than do participants in the 36 condition, but participants judgments on the second list should be the same because the list has the same number of items for both lists.

When judging the likelihood of an event, one must first retrieve all relevant alternative hypotheses from long-term memory. This study has examined whether similar but irrelevant alternative hypotheses could be retrieved and interfere with one's ability to make proportion or probability judgments. Two main hypotheses of the effects of irrelevant alternatives on probability judgments were considered. First, it was possible that people could activate irrelevant alternatives, but be unable to identify irrelevant alternatives as irrelevant, causing them to be included in the probability judgment comparison and leading to discrimination bias. Second, it was possible that people could activate irrelevant alternatives and be able to identify them as irrelevant, but be unable to inhibit them. In this case, irrelevant alternatives would occupy WM space, limiting ones ability to generate and compare further relevant alternatives, leading to dysinhibition bias. Previous research (Kane & Engle, 2000) has found that individual differences in WM related to individual differences in ones ability to discriminate subsets of memories, because individual differences in WM relate to differences in susceptibility to PI. People with greater WM capacities are better able to selectively retrieve relevant information and to inhibit irrelevant information. Thus, it was also hypothesized that if inhibition failure occurred, high spans' probability judgments would be less affected by strong irrelevant hypotheses as would low spans' probability judgments.

The results revealed 1) that proportion and probability judgments can be affected by irrelevant past information; 2) that retrieving irrelevant items can cause discrimination failure, and 3) that when discrimination failure occurs, WM span does not relate to participants' judgments. Further, the study revealed that List Frequency can affect probability judgments. Stronger lists led to more subadditivity, and weaker lists led to less subadditivity. Our results support the list differentiation account of PI, but we acknowledge that response competition and failure to inhibit response competition also cause forgetting due to PI. However, our results suggest that perhaps list differentiation precedes response competition, in that without the ability to identify irrelevant information irrelevant information cannot compete with current list retrieval. Rather, if only list discrimination failure occurs, recall due to PI decreases because a larger set of items is searched than would be the case if no PI was present.

In sum, we have found evidence that at times, people may in fact fail to completely distinguish between relevant and irrelevant alternatives when making probability judgment. Thus, it may be possible that when people make likelihood judgments, such as the likelihood that a given basketball team will win its division championship, irrelevant alternatives (non-division basketball teams) may be generated and included in the likelihood estimate. Discrimination failure could occur in any of the following non-exclusive and non-exhaustive instances. Discrimination failure could occur when one lacks experience, such as when either a non-basketball fan or a person from a different division estimates the likelihood that a given team will win its division championship. Discrimination bias could occur when relevant and irrelevant alternatives are highly similar and confusable. Discrimination failure could occur when one must discriminate alternatives temporally. In any of these instances, if one fails to identify irrelevant alternatives that are generated, judgment bias may result.

References

Anderson, M. C., & Neely, J. H. (1996). Memory. In E. L. Bjork & R. A. Bjork (Eds.), (pp. 237–313). New York: Academic Press.

Baddeley, A. (1990). *Human memory: Theory and practice.* Needham Heights, MA: Allyn & Bacon.

Battig, W. F., & Montague, W. E. (1969). Category norms of verbal items in 56 categories: A replication and extension of the connecticut category norms. *Journal of Experimental Psychology*, *80*, 1–46.

Bjork, R. A., LaBerge, D., & LeGrand, R. (1968). The modification of short-term memory through instructions to forget. *Psychonomic Science*, *10*, 55–56.

Carretti, B., Cornoldi, C., & De Beni, R. (2004). What happens to information to be suppressed in working memory tasks? short and long term effects. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology*, *57A*, 1059–1084.

Coccia, M. R., & Wickens, D. D. (1976). The effect of instructions to forget on proactive inhibition. *Bulletin of the Psychonomic Society*, *7*, 479–480.

Dallett, K., & Wilcox, S. G. (1968). Contextual stimuli and proactive inhibition. *Journal of Experimental Psychology*, *78*, 475–480.

Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning & Verbal Behavior*, *19*, 450–466.

Dillon, R. F., & Thomas, H. (1975). The role of response confusion in proactive interference. *Journal of Verbal Learning and Verbal Behavior*, *14*, 603–615.

Donaldson, W., & Murdock, B. B. (1968). Criterion change in continuous recognition memory. *Journal of Experimental Psychology*, *76*, 325–330.

Dougherty, M. R. P., Gettys, C. F., & Ogden, E. E. (1999). Minerva-dm: A memory processes model for judgments of likelihood. *Psychological Review*, *106*, 180–209.

Dougherty, M. R. P., Gettys, C. F., & Thomas, R. P. (1997). The role of mental simulation in judgments of likelihood. *Organizational Behavior & Human Decision Processes*, *70*, 135–148.

Dougherty, M. R. P., & Hunter, J. E. (2003). Hypothesis generation, probability judgment, and individual differences in working memory capacity. *Acta Psychologica*, *113*, 263–282.

Engle, R. W., Conway, A. R. A., Tuholski, S. W., & Shisler, R. J. (1995). A resource account of inhibition. *Psychological Science*, *6*, 122–125.

Fox, C. R., & Rottenstreich, Y. (2003). Partition priming in judgment under uncertainty. *Psychological Science*, *14*, 195–200.

Glenberg, A. M., & Swanson, N. G. (1986). A temporal distinctiveness theory of recency and modality effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *12*, 3–15.

Gonzalez, R., & Nelson, T. O. (1996). Measuring ordinal association in situations that contain tied scores. *Psychological Bulletin*, *119*, 159–165.

Jarvis, B. (2004). *Media lab and direct rt software.* (http://www.empirisoft.com)

Johnson, M. K., Hashtroudi, S., & Lindsay, D. S. (1993). Source monitoring. *Psychological Bulletin*, *114*, 3–28.

Johnson, M. K., Kounios, J., & Reeder, J. A. (1994). Time-course studies of reality monitoring and recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*, 1409–1419.

Kane, M. J., & Engle, R. W. (2000). Working-memory capacity, proactive interference, and divided attention: Limits on long-term memory retrieval. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *26*, 336–358.

Lustig, C., May, C. P., & Hasher, L. (2001). Working memory span and the role of proactive interference. *Journal of Experimental Psychology: General*, *130*, 199–207.

May, C. P., Hasher, L., & Kane, M. J. (1999). The role of interference in memory span. *Memory & Cognition*, *27*, 759–767.

Postman, L., & Hasher, L. (1972). Conditions of proactive inhibition in free recall. *Journal of Experimental Psychology*, *92*, 276–284.

Postman, L., & Keppel, G. (1977). Conditions of cumulative proactive inhibition. *Journal of Experimental Psychology: General*, *106*, 376–403.

Reutner, D. B., & Graybeal, C. (1978). Multiple encoding of words and pictures in short/term memory within and between taxonomic categories. *The Journal of General Psychology*, *99*, 213–221.

Rosen, V. M., & Engle, R. W. (1997). The role of working memory capacity in retrieval. *Journal of Experimental Psychology: General*, *126*, 211–227.

Somers, R. H. (1962). A new asymmetric measure of association for ordinal variables. *American Sociological Review*, *27*, 799–811.

Sprenger, A., & Dougherty, M. R. (2005). Differences between probability and frequency judgments: The role of individual differences in working memory capacity. *Under revision.*

Turner, M. L., & Engle, R. W. (1989). Is working memory capacity task dependent? *Journal of Memory & Language*, *28*, 127–154.

Tversky, A., & Koehler, D. J. (1994). Support theory: A nonextensional representation of subjective probability. *Psychological Review*, *101*, 547–567.

Underwood, B. J. (1945). The effect of successive interpolations on retroactive and proactive inhibition. *Psychological Monographs*, *59*, 1–33.

Underwood, B. J. (1957). Interference and forgetting. *Psychological Review*, *64*, 49–60.

Van Overschelde, J. P., Rawson, K. A., & Dunlosky, J. (2004). Category norms: An updated and expanded version of the battig and montague (1969) norms. *Journal of Memory and Language*, *50*, 289–335.

Wickens, D. D. (1970). Encoding categories of words: An empirical approach to meaning. *Psychological Review*, *77*, 1–15.

Wickens, D. D., Born, D. G., & Allen, C. K. (1963). Proactive inhibition and item similarity in short-term memory. *Journal of Verbal Learning and Verbal Behavior*, *2*, 440–445.

Windschitl, P. D., & Wells, G. L. (1998). The alternative-outcomes effect. *Journal of Personality and Social Psychology*, *75*(6), 1411–1423.

Wixted, J. T., & Rohrer, D. (1993). Proactive interference and the dynamics of free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*, 1024–1039.