Journal of Experimental Psychology: General
2006, Vol. 135, No. 2, 262–281


0096-3445/06/$12.00 DOI: 10.1037/0096-3445.135.2.262

# The Influence of Improper Sets of Information on Judgment: How Irrelevant Information Can Bias Judged Probability

Michael R. Dougherty and Amber Sprenger
University of Maryland

This article introduces 2 new sources of bias in probability judgment, discrimination failure and inhibition failure, which are conceptualized as arising from an interaction between error prone memory processes and a support theory like comparison process. Both sources of bias stem from the influence of irrelevant information on participants' probability judgments, but they postulate different mechanisms for how irrelevant information affects judgment. The authors used an adaptation of the proactive interference (PI) and release from PI paradigm to test the effect of irrelevant information on judgment. The results of 2 experiments support the discrimination failure account of the effect of PI on probability judgment. In addition, the authors show that 2 commonly used measures of judgment accuracy, absolute and relative accuracy, can be dissociated. The results have broad implications for theories of judgment.

*Keywords:* probability judgment, hypothesis generation, working memory, proactive interference, discrimination bias

Most theories of probability judgment assume that people compare a focal hypothesis with at least one alternative hypothesis (Dougherty, Gettys, & Ogden, 1999; Tversky & Koehler, 1994; Windschitl & Wells, 1998). The comparison process inherent in these models assumes that the strength of the focal hypothesis is compared to the strength of *relevant* competing hypotheses—hypotheses that are logical contenders to the focal. In many cases, the assessment of the competing hypotheses necessitates that the competing hypotheses be generated from long-term memory (Dougherty, Gettys, & Thomas, 1997; Dougherty & Hunter, 2003a; Gettys & Fisher, 1979). For example, a physician considering the likelihood that a patient has pneumonia presumably generates relevant competing alternatives to the pneumonia hypothesis prior to rendering a diagnosis (Elstein, Shulman, & Sprafka, 1978; Weber, Böckenholt, & Hilton, 1993).

Particularly germane to the assessment, and presumably the accuracy, of one's judgment is the degree to which one is influenced by *irrelevant* alternative hypotheses. Relevant alternative hypotheses are those 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 a disease hypothesis that has been eliminated on the basis of a blood test. However, if the patient's symptomology, other than the blood test result, resembles the disconfirmed diagnosis, associative memory processes alone may prevent the physician from being able to completely inhibit the disconfirmed diagnosis.

The present research examined the extent to which the generation of irrelevant alternative hypotheses affects the accuracy of probability judgments. Our research addressed two general questions. First, can judgments of probability be affected by irrelevant information? Prescriptively, when one estimates the probability of a particular event, one's judgment should incorporate only judgment-relevant information. However, the failure to identify (i.e., discriminate) irrelevant alternatives as irrelevant or the failure to inhibit hypotheses identified as irrelevant might lead to biased probability judgments. 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 consideration of alternatives has been of interest to a variety of subdisciplines within psychology. For example, alternative generation has been used as a mechanism for debiasing overconfidence (Koriat, Lichtenstein, & Fischhoff, 1980) and understanding the source of hindsight bias (Roese & Olson, 1996), and it is the basis of counterfactual reasoning (Kahneman & Tversky, 1982; for a review, see Koehler, 1991). In addition, the consideration of

Michael R. Dougherty and Amber Sprenger, Department of Psychology, University of Maryland.

Both authors contributed equally to this research, which was supported by the National Science Foundation under Grant SES-0134678. We thank Thomas Wallsten, David Huber, and Rick Thomas for their invaluable comments and suggestions. We thank Jennifer Joy and Matthew Sams for help with data collection and Anuj Shah for help with programming Experiment 1.

Correspondence concerning this article should be addressed to Michael Dougherty, Department of Psychology, University of Maryland, 1147 Biology/Psychology Building, College Park, MD 20742-4411. E-mail: mdougherty@psyc.umd.edu

alternatives has been shown to be important in a variety of applied contexts, including medical diagnosis (Elstein et al., 1978; Weber et al., 1993), accountants' determination of accounting errors (Libby, 1985), and mechanics' assessments of causes of automobile failure (Mehle, 1982). However, most research on alternative generation has focused on the impact of *relevant* alternatives on judgment processes. Our goal in this research was to develop and test a theoretical framework that describes how both *relevant* and *irrelevant* alternatives influence judgment.

## Theoretical Framework

One can conceptualize the impact of relevant and irrelevant information on probability judgment within the support theory framework. Tversky and Koehler (1994) proposed that judgments of probability are made by comparing the support for a focal hypothesis ($A$) with the support for a set of alternative hypotheses ($B$):

$$P(A, B) = \frac{s(A)}{s(A) + s(B)} \quad (1)$$

where $s(A)$ and $s(B)$ represent the support for $A$ and $B$, respectively, and $P(A, B)$ is the probability of hypothesis A versus hypothesis B occurring. Support theory assumes that $A$ and $B$ are drawn from sample space $S$ where $\{A, B\} \in S$, $\{b_1, b_2 \ldots b_N\} \in B$, and where the $N$ $b_i$s are mutually exclusive and exhaustive (Tversky & Koehler, 1994). However, one can think of the set of alternatives to $A$ as consisting of a relevant set $B$ and an irrelevant set $B'$, where $B' \in S'$, and $\{b'_1, b'_2 \ldots b'_N\} \in B'$. Within the support theory framework, one typically assumes that (a) participants are able to accurately discriminate between elements of $B$ and $B'$, and (b) elements of $B'$ have no effect on $P(A, B)$. Indeed, that judgment is based only on elements drawn from the proper set is a fundamental assumption, not only of support theory, but also of any normative or prescriptive theory of probability judgment. However, there are at least two ways in which elements drawn from an improper set $B'$ might influence $P(A, B)$.

Consider the case in which one is asked to judge the likelihood that the University of North Carolina will win the Atlantic Coast Conference (ACC) basketball tournament. In support theory terms: P(UNC win, UNC not win) = s(UNC)/ [s(UNC) + s(not UNC)]. The *relevant* set of alternative hypotheses includes all other teams in the ACC. The judgment, P(UNC win, UNC not win), should decrease as the number of alternative ACC teams considered increases. *Irrelevant* alternatives are those that are drawn from an inappropriate sample space (e.g., the University of Florida is a Southeastern Conference [SEC] team and is therefore irrelevant to the judgment of UNC) or those alternatives that are inconsistent with the available data (e.g., if one learns that Clemson, an ACC team, was not allowed to participate in the tournament because of rules violations, Clemson should be eliminated from consideration). Thus, teams from conferences outside the ACC should be deemed irrelevant and be excluded from the comparison process.

There are two possible effects that irrelevant alternatives can have on judged probability. First, if the judge generates an irrelevant alternative but fails to identify the alternative as irrelevant, then the irrelevant alternative will be included in the probability judgment. For example, if one generates the University of Florida (an SEC team) when judging the likelihood that UNC would win the ACC, and fails to identify Florida as a SEC team, it could be included in the comparison process. This type of failure to discriminate is assumed to occur if the decision maker (a) has no knowledge on which to base a discrimination between relevant and irrelevant alternatives, (b) lacks source monitoring cues to discriminate between relevant and irrelevant alternatives, or (c) is unable to initiate the consistency-checking process that is responsible for checking whether an alternative is relevant or irrelevant. Because the number of items one can compare while making judgments is limited (Dougherty & Hunter, 2003a, 2003b), the inclusion of an irrelevant alternative would take the place of other possible relevant alternatives. To the degree that the strength of the irrelevant alternative differs from the relative alternative it displaces, judgments will be biased by the inclusion of an irrelevant alternative in the judgment.

A second type of effect that might arise as a result of generating irrelevant alternatives is that the irrelevant alternatives might actually interfere with the generation of relevant alternatives. For example, normatively, if one generates the University of Florida but is able to identify it as a non-ACC team, it should have no effect on judged probability. However, the generation of University of Florida might actually interfere with one's ability to generate ACC teams or might occupy a slot in working memory (WM) that would otherwise be used for holding a relevant alternative. Thus, when making a judgment, if the irrelevant alternative is discriminated as irrelevant it is not included in the computation. However, generating and failing to inhibit an irrelevant alternative can lead to fewer relevant alternatives being included in the judgment and consequently can lead to increased judged probability. In terms of support theory, the support for B will decrease as the number of interfering items increases. Note that if an irrelevant alternative is generated and discrimination failure occurs, inhibition will *not* play a role because the irrelevant hypothesis was incorrectly considered relevant. Thus, accurate discrimination is a prerequisite for inhibition factors to play a role when irrelevant alternatives are generated.

The discrimination failure and inhibition failure accounts are depicted in Figure 1. These two accounts, coupled with assumptions about how many and which hypotheses are considered by the decision maker, enable us to formulate a set of competing hypotheses, which we detail below. Our first assumption concerns the role of WM in judgment:

> *Assumption 1:* WM constrains the total number of alternative hypotheses that the decision maker can entertain.

Dougherty and Hunter (2003a, 2003b) and Sprenger and Dougherty (2006) argued that the inability to consider all alternative hypotheses can lead to excessive probability judgments. Indeed, consistent with Assumption 1, several studies have shown that judgments are negatively correlated with individual differences in WM capacity and the number of hypotheses explicitly considered (Dougherty et al., 1997; Dougherty & Hunter, 2003a, 2003b; Sprenger & Dougherty, 2006).

Figure 1A illustrates Assumption 1. Recall that hypotheses can be partitioned into three sets: $A$, the to-be-judged item; $B$, the set

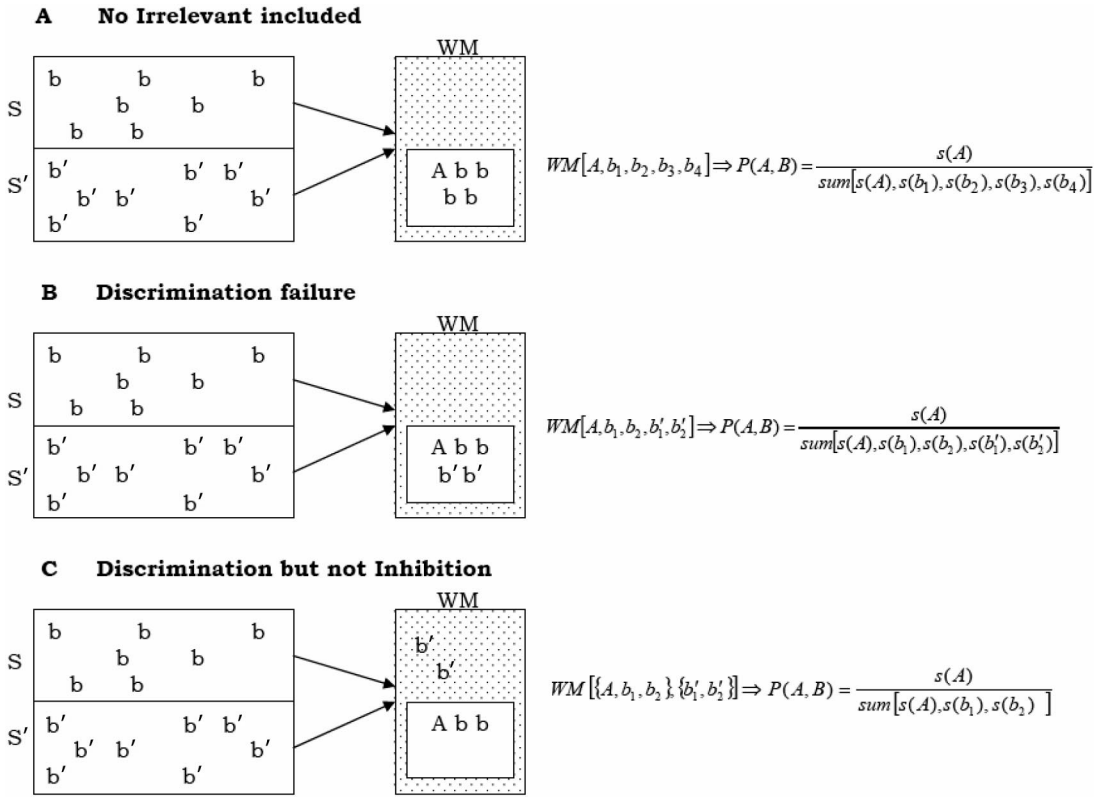


*Figure 1.* The box S represents the set of relevant alternative hypotheses one could generate when making the probability judgment, and box S′ represents a set of irrelevant hypotheses. The central box represents those items active in working memory (WM). The items within the inner box represent items one considers relevant, and items outside that box represent items that have been discriminated as irrelevant but that are still active in WM. Figure 1A illustrates the case in which people do not sample irrelevant alternatives. Figure 1B illustrates the case in which people generate irrelevant alternatives and fail to discriminate them as irrelevant. Then, those irrelevant alternatives are included in the denominator of the probability assessment. Figure 1C illustrates the case in which people generate irrelevant alternatives and correctly discriminate them as irrelevant.

of relevant alternatives to $A$; and $B'$ the set of irrelevant alternatives to $A$. In Figure 1A, the box S represents the set of relevant hypotheses and the box S′ represents the set of irrelevant hypotheses. Items are sampled from the relevant or irrelevant set and placed into WM. The unshaded portion of WM represents those hypotheses that are deemed relevant and included in the judgment comparison. First, suppose that the capacity of WM is limited to five elements and that no irrelevant, $b'$, alternatives were generated. This is illustrated Figure 1A and gives rise to Equation 2:

$$WM[A, b_1, b_2, b_3, b_4] \Rightarrow P(A, B) = \frac{s(A)}{sum[s(A), s(b_1), s(b_2), s(b_3), s(b_4)]} \quad (2)$$

As is clear from Equation 2, $P(A, B)$ is a function of the overall support for $A$ versus $B$, where the support for $B$ is a function of the number of relevant alternatives (weighted by their support values) included in the computation. Equation 2 is an instantiation of support theory where it is assumed that elements of $B$ are generated from long-term memory and that the number of elements generated and maintained in WM is limited by WM capacity.

Figures 1B and 1C demonstrate cases in which two elements from the irrelevant set (denoted $b'$) are generated and stored in working memory. As shown in Figure 1B, if the set of irrelevant alternatives is non-null, and the judge *fails to discriminate* elements of set $B'$ from elements of set $B$, then elements of $B'$ will enter into both WM and the computation of $P(A, B)$. Thus, the discrimination failure hypothesis leads to Equation 3:

$$WM[A, b_1, b_2, b_1', b_2'] \Rightarrow P(A, B) = \frac{s(A)}{sum[s(A), s(b_1), s(b_2), s(b_1'), s(b_2')]} \quad (3)$$

Equation 3 predicts what we term the *discrimination* bias.[1] $P(A, B)$ will be biased by the irrelevant alternatives when $s(b_1') + s(b_2') \neq s(b_3) + s(b_4)$, where $b_3$ and $b_4$ are alternatives from set $B$ that were

[1] Note that we use the term *discrimination failure* to refer to the process of failing to discriminate irrelevant alternatives and the term *discrimination bias* to refer to the biased judgments that result from failing to discriminate irrelevant alternatives.

not included in the judgment because $b'_1$ and $b'_2$ displaced them. Generally, we assume that if participants consider irrelevant alternatives, the irrelevant alternatives will tend to have higher strength values in memory than the relevant alternatives that are not considered. This leads to our second main assumption:

> *Assumption 2:* The probability that irrelevant hypotheses included in the comparison are stronger than the relevant hypotheses they replace is greater than the probability of the converse: $p[s(b'_1) + s(b'_2) > s(b_3) + s(b_4)] > q[s(b'_1) + s(b'_2) \leq s(b_3) + s(b_4)]$, where $p + q = 1.0$, and $p > q$.

This assumption is based on the finding that events with higher frequency of occurrence are preferentially recalled because they are more strongly represented in memory (cf. repetition and word frequency effects in long-term memory retrieval, Roediger & Challis, 1992; Hintzman, 1970; for a review, see Horton & Mills, 1984). Therefore, among a group of relevant and irrelevant hypotheses, strong irrelevant hypotheses are more likely to enter WM than weak ones, and the entry of these irrelevant hypotheses will tend to be at the expense of weak relevant hypotheses, not strong ones.

Inasmuch as irrelevant hypotheses are retrieved instead of relevant hypotheses, we would expect them to have higher memory strengths on average. Given Assumption 2, $P(A, B)$ generally should decrease when participants fail to discriminate between relevant and irrelevant alternatives. This is because the denominator of Equation 3 will tend to be higher when irrelevant hypotheses are generated.[2] Judgments are often found to be excessively high because of the failure to consider all hypotheses. Thus the decrease in the magnitude of judgment might be viewed as an *improvement in the absolute accuracy* of participants' judgments. Because the relationship expressed in Assumption 2 is stochastic, one can view this process as adding *error* to the assessment of probability. Adding error to the assessment of probability will decrease the correlation between subjective and objective probabilities, because it will perturb the rank order of the to-be-judged items. The decrease in the correlation between subjective and objective probabilities can be viewed as a *decrease in the relative accuracy* of participants' judgments. Thus, our discrimination failure hypothesis predicts a dissociation between relative and absolute accuracy.

Figure 1C illustrates the second way in which the set of irrelevant alternatives can affect judgment. In this case, we assume that irrelevant alternatives are discriminated from relevant alternatives (as illustrated by the $b'$ alternatives being outside the box of items to be included in the judgment). However, because the irrelevant alternatives are not inhibited, they are assumed to take up resources in WM that would otherwise be used to generate and maintain additional alternatives from the relevant set. If irrelevant alternatives are discriminated but not inhibited, then the judged probability of $A$ should increase as a function of the number of irrelevant alternatives considered increases. This is demonstrated in Equation 4 where $b'_1$ and $b'_2$ both occupy slots in WM but are not included in the computation of $P(A, B)$.

$$WM[\{A, b_1, b_2\}, \{b'_1, b'_2\}] \Rrightarrow P(A, B) = \frac{s(A)}{sum[s(A), s(b_1), s(b_2)]} \quad (4)$$

Equation 4 predicts what we call the *dysinhibition* bias[3]: Judged probability should increase when participants are able to discriminate between relevant and irrelevant alternatives but are unable to inhibit the irrelevant items from taking up WM resources.

Although the failure to inhibit irrelevant alternatives could also affect the rank-order correlations between the objective probabilities and the subjective probabilities (because there could be variation in which and how many relevant alternatives are not included from judgment to judgment), the amount of error due to the removal of relevant alternatives in the inhibition failure hypothesis should be less than the amount of error in judgment that occurs with discrimination failure. In both the discrimination failure and inhibition failure accounts, error is introduced by virtue of allowing irrelevant elements to supplant relevant hypotheses in WM. However, in the discrimination failure case, error is also introduced by virtue of the irrelevant hypotheses entering into the comparison process. Thus, one would expect the discrimination failure account to produce a greater decrease in the correlational accuracy than inhibition failure, because in the discrimination failure account error is introduced at both the sampling and comparison stages whereas in the inhibition failure account error is only introduced at the sampling process.

## Overview of Experiments

Our approach to studying the effect of irrelevant information on judgments of probability was to use an adaptation of the proactive interference (PI) and release from PI (RPI) paradigm. The typical PI/RPI paradigm involves two phases. In the first phase, participants learn and recall several lists of different words that are related by category membership. During this phase, PI builds up with each successive list of words and recall performance decreases (Postman & Keppel, 1977; Underwood, 1945, 1957; Wickens, 1970; Wickens, Born, & Allen, 1963). The second phase involves the release from PI. In this phase, participants learn and recall a list of words from a new category. Recall of words learned in the release phase increases back to the level of recall before PI was introduced (Wickens et al., 1963). The finding that PI subsides when a distinct category is used suggests 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 the reduction in retrieval in the buildup phase is due to interference from similar, but irrelevant, prior lists.

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 pro-

[2] Note that it may be possible to set up experimental situations in which weak irrelevant hypotheses displace stronger relevant hypotheses, a possibility allowed for under our Assumption 2. In this case, discrimination failure would lead to judgments that increase, rather than decrease, in magnitude. Experiment 2 addressed this possibility.

[3] Note that we use the term *inhibition failure* to refer to the process of failing to inhibit irrelevant alternatives and the term *dysinhibition bias* to refer to the biased judgments that result from failing to inhibit irrelevant alternatives.

vides a well-specified context for testing whether irrelevant information affects judgment, because the separation between relevant and irrelevant information is easily conceptualized. In our adaptation of the PI paradigm to study probability judgments, we were interested in examining how information learned on prior lists might influence one's judgments of information learned on a later list. In our experiments, items learned on the prior list are irrelevant to judgments of items learned on later lists. However, given that PI builds up across successive lists of information, we were interested in what effect irrelevant information has on participants' probability judgments when PI is present. More specifically, would the results better support the inhibition failure account or the discrimination failure account?

## Effect of Irrelevant Information on Judgments of Probability

In our modification to the PI/RPI task as implemented in Experiment 1, participants imagined making several successive trips to a grocery store. On the first three trips, participants imagined purchasing various amounts (between 2 and 16) of produce items, such as bananas, broccoli, and apples. Items were experienced one exemplar at a time, such that if a participant bought 8 of a particular item, they would actually experience 8 separate exemplars of that item over the course of the grocery trip. A specific kind of item (e.g., bananas) was purchased on one and only one trip. The first three trips to the store were considered the buildup of PI phase, because produce items from previous trips could potentially interfere with the retrieval of produce items bought on the current shopping trip. For each list, items from the current list were considered relevant items and items from previous lists were considered irrelevant items. 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 of the four trips, participants judged the proportion of each kind of item in their shopping bag from the most recent trip. Because participants were presented only with the focal hypothesis during the judgment phase, the judgment task required participants to generate or retrieve the alternatives from long-term memory. We assumed that participants would retrieve more irrelevant (B′) alternatives as PI increased.

Equations 2, 3, and 4, coupled with Assumptions 1 and 2, suggest three competing hypotheses concerning the effect of PI on judgment:

> *Hypothesis 1:* If irrelevant hypotheses are identified as irrelevant and inhibited so that one could continue to generate further hypotheses for inclusion in the comparison, then we would expect absolute and relative accuracy to be unaffected by the buildup of, and release from, PI.

> *Hypothesis 2:* If irrelevant hypotheses are not identified as irrelevant, one could include those hypotheses in the probability computation. Based on Assumption 2 stated earlier, increases in PI should lead to a decrease in the magnitude of participants' judgments, as well as a decrease in the relative accuracy of their judgments. Both magnitude and relative accuracy should increase when PI is released.

> *Hypothesis 3:* If irrelevant hypotheses are identified as irrelevant but are not successfully inhibited, the irrelevant alternative might occupy resources in WM that otherwise would be used to maintain relevant alternatives. However, if participants know the item is irrelevant, we assume that it is not included in the computation of probability. This leads to the prediction that as PI increases the magnitude of participants' judgments should increase, with little effect of PI on relative accuracy.

The three hypotheses correspond, respectively, to the three processes illustrated in Figure 1: (a) no irrelevant hypotheses included, (b) discrimination failure, and (c) discrimination but not inhibition. To further validate the predictions of each of these hypotheses, we conducted a set of simulations to test the effect of PI given the specific experimental designs used in Experiments 1 and 2. The simulations are detailed in the Appendix and illustrate the predictions of the discrimination and inhibition failure accounts.

As an exploratory factor, we examined whether WM capacity would interact with the effect of PI on judgments. Previous research has demonstrated that one of the primary functions of WM is the inhibition of goal-irrelevant information (Lustig, May, & Hasher, 2001; May, Hasher, & Kane, 1999), and a variety of studies have shown that individual differences in WM capacity predict participants' performance on tasks that require inhibition (e.g., Carretti, Cornoldi, & De Beni, 2004; Engle, Conway, Tuholski, & Shisler, 1995; Rosen & Engle, 1997), including retrieval in the PI paradigm (Kane & Engle, 2000). Dougherty and Hunter (2003a, 2003b; see also Sprenger & Dougherty, 2006) revealed that individual differences in WM capacity were negatively correlated with the magnitude of participants' probability judgments. Implied in their treatment of the role of WM was that the negative correlation between judgment and WM span was due to high-span participants' ability to include more relevant alternatives in the comparison process (e.g., Equation 2). Dougherty and Hunter (2003a) speculated that irrelevant alternatives might affect judgment by interfering with the retrieval of relevant alternatives; however, 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 the differential susceptibility of high- and low-span participants to interference. Therefore, in this study, we measured WM capacity as an exploratory factor to see whether WM span related to judgment magnitude when PI was present. Within the context of our two accounts, only the inhibition failure hypothesis anticipates a negative correlation between judgment magnitude and WM span.

## Experiment 1

Experiment 1 was designed to examine the relationship between PI and proportion judgments. The method for Experiment 1 was based on the multiple trial, free-recall PI paradigm. Participants learned words, which were 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 (build up of PI) and then learned and judged a list of words from a new category (release from PI). As an exploratory variable, WM span was measured using Turner and Engle's (1989) Operation Span (O-span) task.

## Method

### Participants

The 112 University of Maryland students who participated received monetary compensation for their participation.

### Materials

Experiment 1 was conducted on the computer using a Java program for the PI task. A tape recorder was used to record verbal memory recall. For the PI learning task, 48 words from the categories of 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.

### Design and Procedure

The design was a 4 (list: 1–4) × 5 (item frequency: 0, 2, 6, 12, 16) within-subjects design. Further, WM span was measured for exploratory purposes.

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 four parts: participants first engaged in a learning phase, then engaged in a rehearsal prevention task to eliminate recency effects, verbally recalled the items from the current trial, and finally made proportion judgments of 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. 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 (i.e., trip to the grocery store), 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 presentation frequency was determined randomly for each participant. For the first three trips to the store (PI buildup), 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 to reduce possible recency effects. Participants counted backward by threes for 20 s.

*PI task—memory recall.* We included a recall task as a manipulation check to ensure that we could replicate standard PI effects on recall within our modified task. Participants heard a series of 12 beeps (each 1,350 ms 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 four other fruits and vegetables not appearing in that section or in any future sections. We asked participants to judge items from previous shopping trips for two 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 nonzero 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 to determine if participants' proportion judgments changed as they made other judgments. All other items were judged in a random order.

*O-span task.* After completing the PI task, participants completed the O-span task as a measure of WM span (Turner & Engle, 1989). A detailed description of the O-span task is presented in Turner and Engle (1989). Participants were classified into WM capacity tertials based on their O-span score. The tertial classification levels were determined from a separate experiment of over 150 participants (Sprenger & Dougherty, 2006).

## Results and Discussion

The data from six participants were excluded from the analyses for the following reasons: scoring less than 85% correct on the O-span math problems (3 participants); computer failure (1 participant); giving judgments of only zero (1 participant); and having judgment sums greater than 2.5 standard deviations above the mean (1 participant).[4]

### Manipulation Check: Recall

We examined recall for the four lists to ensure that our modification to the PI paradigm would yield the standard PI effects.

[4] The participant's judgment sums for each list were 450 for List 1, 450 for List 2, 440 for List 3, and 430 for List 4. The group mean judgment sums for each list were 142.60 ($SE$ = 7.54) for List 1, 122.84 ($SE$ = 6.24) for List 2, 112.15 ($SE$ = 5.53) for List 3, and 120.66 ($SE$ = 5.86) for List 4.

Consequently, we hypothesized that recall would decrease from List 1 through 3, then increase on List 4 when PI was released. Mean recall for the four lists are presented in Table 1. There was a significant main effect of list on the number of words recalled such that as PI increased, recall decreased, $F(3, 309) = 15.03$, $p < .0001$, $\hat{\omega}^2 = .09$. Further, trend analyses revealed that a quadratic trend fit the changes in recall well, $F(1, 103) = 48.29$, $p < .0001$. 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, $F(2, 103) = 8.39$, $p = .0004$, $\hat{\omega}^2 = .03$, there was no interaction between list and WM span. Overall, high spans recalled more words than did low or middle spans.

We also examined the number of intrusions of previous list items participants reported during recall. The left part of Table 2 presents the mean number of recall intrusions for each of the four lists. 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, $F(3, 309) = 29.85$, $p < .0001$, $\hat{\omega}^2 = .17$. No main effect of WM and no interaction between WM and list on the number of intrusions were found. Overall, the number of intrusions reported was low. Intrusions reflect a failure to discriminate between lists. The finding that participants reported intrusions of previous list items suggests that they experienced discrimination failure. However, 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 memory. Thus, people can quickly detect whether they have seen items before, but they require more time to determine from where exactly they saw items. Time pressure may have impaired participants' ability to monitor whether their recall output was a current list item or an intrusion, and thus the mean number of intrusions reported in this experiment may be biased upward.

### *Judgment Magnitude*

Hypotheses 1, 2, and 3 presented previously suggest the three potential effects of PI on judgment. Hypothesis 1 represents the null, wherein PI would have no effect on judgment if people could completely discriminate and/or inhibit alternatives from prior lists. Hypotheses 2 and 3 represent the discrimination and inhibition failure accounts. If the discrimination failure hypothesis is correct, we expect judgment magnitude to show a quadratic trend, wherein judgments should *decrease* from List 1 to List 3, then increase on List 4 at release from PI. In contrast, if the inhibition failure hypothesis is correct, we expect judgment magnitude to *increase* from List 1 to List 3 as PI increases and decrease on List 4 at release from PI.

Table 1
*Mean Recall As a Function of List*

| | | Experiment 2 | |
|---|---|---|---|
| | | List frequency | |
| List | Experiment 1 | Ascending | Descending |
| 1 | 6.25 (.19) | 6.57 (.18) | 6.82 (.26) |
| 2 | 5.84 (.14) | 6.49 (.20) | 6.06 (.27) |
| 3 | 5.56 (.17) | 6.34 (.21) | 5.60 (.28) |
| 4 | 6.58 (.12) | 7.17 (.14) | 6.90 (.19) |

*Note.* Standard errors of the mean are presented in parentheses.

Table 2
*Mean Number of Verbal Intrusions During Recall As a Function of List*

| | | Experiment 2 | |
|---|---|---|---|
| | | List frequency | |
| List | Experiment 1 | Ascending | Descending |
| 1 | 0.00 (.00) | 0.00 (.00) | 0.03 (.02) |
| 2 | 0.48 (.08) | 0.28 (.11) | 0.17 (.11) |
| 3 | 0.90 (.15) | 0.36 (.11) | 0.25 (.11) |
| 4 | 0.02 (.02) | 0.00 (.00) | 0.03 (.02) |

*Note.* Standard errors of the mean are presented in parentheses.

Judgment magnitude can be analyzed on two levels. First, we can examine the degree to which participants' judgments sum to more than 100%. Analyzing the data this way gives us a method for examining the degree to which absolute accuracy deviates from additivity. Sums greater than 100% demonstrate subadditivity. Second, we can examine the individual probability judgments for the frequency-12 items (henceforth called the *focal item*). Recall that for each list, participants judged one focal item first and one last. Thus, we can examine whether participants learned as a function of making judgments within a list. This is important because participants' later judgments may be informed by the set size of the items being judged, which in turn could mitigate the effects of PI on the sum of the judgments.

*Judgment sums: subadditivity.* Judgment sums were computed by summing each participant's judgments of the eight relevant items for each list. Judgments of list irrelevant items were excluded from the judgment sum measure. A 3 (WM span: high, middle, low) × 4 (list: 1–4) mixed factorial analysis of variance (ANOVA) was conducted to determine whether judgment sums changed as a function of PI. Figure 2 presents mean judgment sums for each of the four lists. A main effect of list was found for judgment sums, $F(3, 309) = 18.63$, $p = .0001$, $\hat{\omega}^2 = .11$. Trend analyses revealed that both a linear trend, $F(1, 105) = 24.73$, $p < .0001$, and a quadratic trend, $F(1, 105) = 21.24$, $p < .0001$, fit the changes in judgment sums well, suggesting that as PI increased judgment sums decreased and that at release from PI judgment sums increased again. This pattern of results supports the discrimination failure hypothesis. There was no main effect of WM span on judgment sums, $F(2, 103) = 1.97$, $p > .05$, nor an interaction between list and WM span, $F(6, 309) = 1.48$, $p > .05$.

We had initially expected that the judgment sums would increase back to the level of the first list. However, the fact that we did not find this is not crucial. For example, one might suppose that participants learned the set sizes as they proceeded through the experiment and consequently learned to better distribute judg-

ments across the set of alternatives. The decrease in participants' judgment sums across lists was not due solely to learning, however, because at release from PI participants' judgment sums increased as indicated by the significant quadratic trend.

*Focal judgments.* Our hypotheses for the focal items paralleled those for sums of judgments. 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 all other judgments were made. Those two judgments provided an exploratory analysis of changes in judgment as participants made their other 10 judgments. Figure 3 presents the mean of participants' focal judgments judged first (white bars) and judged last (black bars) for each of the four lists. A marginally significant main effect of list was found for focal judgment judged first, $F(3, 309) = 2.26$, $p = .081$, and a significant main effect of list was found for focal judgments judged last, $F(3, 309) = 4.27$, $p = .006$, $\hat{\omega}^2 = .02$. For the focal judgments judged first, a marginally significant quadratic trend was found, $F(1, 103) = 3.75$, $p = .056$, and for the focal judgments judged last significant linear, $F(1, 103) = 8.46$, $p < .005$, and quadratic, $F(1, 103) = 4.08$, $p < .05$, 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 focal judgments increased again at release from PI. Note in Figure 3 that 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 summed to something greater than 100% and consequently attempted to adjust their judgments downward to 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.

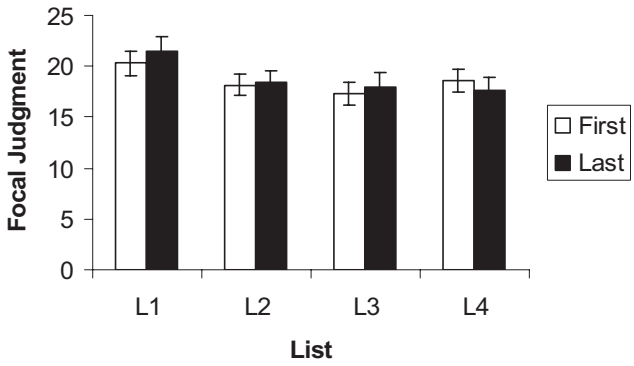


*Figure 3.* Mean proportion judgment for focal item that was judged first as a function of list. PI = proactive interference; L1 = List 1; L2 = List 2; L3 = List 3; L4 = List 4.

## *Relative Accuracy of Judgments*

The relative accuracy of judgments was computed using Somers's D (Somers, 1962).[5] Somers's D is a measure of ordinal association and therefore provides the degree to which two variables are related monotonically. Somers's D was calculated for each participant for each list. The analyses reported here include the four irrelevant previous list judgments in the computations of each Somer's D. However, it is important to point out that the results where we included the irrelevant judgments in the computations were the same as those in which we excluded the irrelevant items from the computations of Somer's D.

Figure 4 presents the mean Somers's D correlations for each of the four lists. A main effect of list on relative accuracy was found, $F(3, 309) = 52.78$, $p < .0001$, $\hat{\omega}^2 = .27$. Further, trend analyses revealed that both a linear trend, $F(1, 103) = 10.01$, $p < .005$, and a quadratic trend, $F(1, 103) = 132.03$, $p < .0001$, 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. It is interesting to note that relative accuracy on List 3 was better than on List 2, even though PI was greater on List 3 than on List 2. However, this difference was not significant, $t(105) = 1.50$, $p > .10$. No interaction between list and WM span was found, $F(6, 309) = 0.74$, $p > .05$, but a main effect of WM was found, $F(2, 103) = 3.29$, $p = .041$, $\hat{\omega}^2 = .01$. High spans had higher correlations and thus greater relative accuracy than low spans, but this did not interact with list. Therefore although high spans had better correlations than did low spans overall, they were equally affected by PI.

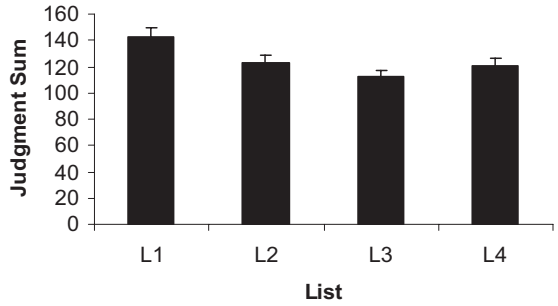


*Figure 2.* Subadditivity of list-relevant probability judgments as a function of list. As proactive interference (PI) increased on L2 and L3, judgment sums decreased. Then on the final release from PI list (L4), judgment sums increased again. L1 = List 1; L2 = List 2; L3 = List 3; L4 = List 4.

## *Irrelevant Items*

As a secondary set of analyses, we also examined participants' judgments of the irrelevant alternatives that occurred in each of the four judgment phases. On each list, participants judged four items that actually occurred on the previous list and not on the current list. Nonzero judgments for those items reflect discrimination failure in that they indicate 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

[5] Gonzalez and Nelson (1996) discussed the implications of using various measures of association and concluded that Somers's $D_{xy}$ should be used for cases in which ties on the criterion variable are unambiguous but ties on the predictor variable are ambiguous.

hypothesized that the number of irrelevant items given nonzero 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 least present. In contrast, if inhibition failure occurred, we hypothesized that few irrelevant items would be given nonzero judgments and that no differences across list would be found. That is, if participants generated irrelevant alternatives and correctly discriminated them as such, they should give those items judgments of zero, regardless of whether they could inhibit them.

The left part of Table 3 presents the mean number of previous list items given nonzero judgments (out of four possible) for each of the four lists. A main effect of list on the mean number of irrelevant items given nonzero judgments was found, such that the mean number of irrelevant items given nonzero judgments on List 1 and on List 4 when PI was less than the mean number of irrelevant items given nonzero judgments on List 2 and on List 3, $F(3, 309) = 68.38$, $p < .0001$, $\hat{\omega}^2 = .32$. 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 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 zero, 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 could not be entirely due to a response bias.[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 1 and 4 and slowest when PI was greatest on Lists 2 and 3. RTs were transformed to rates (1/RT) to reduce the skewness 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 5 (top) 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 were fastest on List 1 and 4 when PI was least, $F(3, 309) = 67.55$, $p < .0001$, $\hat{\omega}^2 = .32$. Figure 5 (bottom) 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, $F(3, 309) = 111.25$, $p < .0001$, $\hat{\omega}^2 = .44$. 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 was 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 buildup of PI.

Table 3
*Mean Number of Irrelevant Items Given Nonzero Judgments As a Function of List*

| | | Experiment 2 | |
|---|---|---|---|
| | | List frequency | |
| List | Experiment 1 | Ascending | Descending |
| 1 | 0.06 (.02) | 0.17 (.07) | 0.11 (.08) |
| 2 | 1.69 (.15) | 1.50 (.27) | 1.11 (.23) |
| 3 | 1.13 (.13) | 1.17 (.23) | 0.92 (.20) |
| 4 | 0.27 (.09) | 0.19 (.15) | 0.31 (.14) |

*Note.* Standard errors of the mean are presented in parentheses.

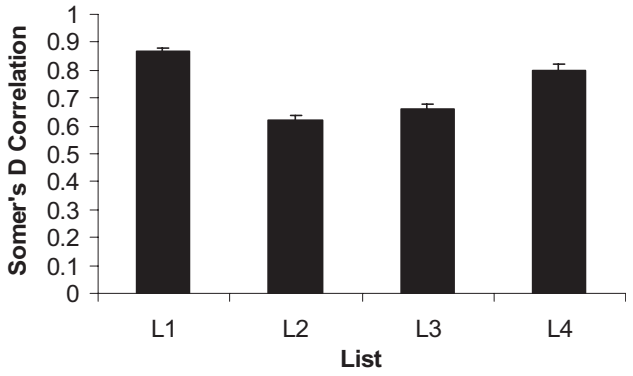


*Figure 4.* Relative accuracy of judgments (as measured by Sommer's $D_{xy}$ correlations between proportion judgments and objective probabilities) as a function of list. A main effect of list was found, such that as proactive interference (PI) increased on L2 and L3, relative accuracy decreased. Then on the release from PI list (L4), relative accuracy increased again. L1 = List 1; L2 = List 2; L3 = List 3; L4 = List 4.

### *Summary of Experiment 1*

Experiment 1 revealed three main findings consistent with the discrimination failure account. First, as predicted by the discrimination failure account, we found that as PI increased, participants'

[6] One might question whether participants understood the instructions that irrelevant items should be given judgments of zero. However, had participants failed to understand our instructions, we would have expected roughly the same number of irrelevant items judged greater than zero across the four lists. The finding that participants gave almost all irrelevant items judgments of zero on the first and fourth lists suggests that they did understand the instructions. This suggests that the higher number of irrelevant items judged greater than zero on Lists 2 and 3 was due to discrimination failure and was not due to a failure to understand the instructions.

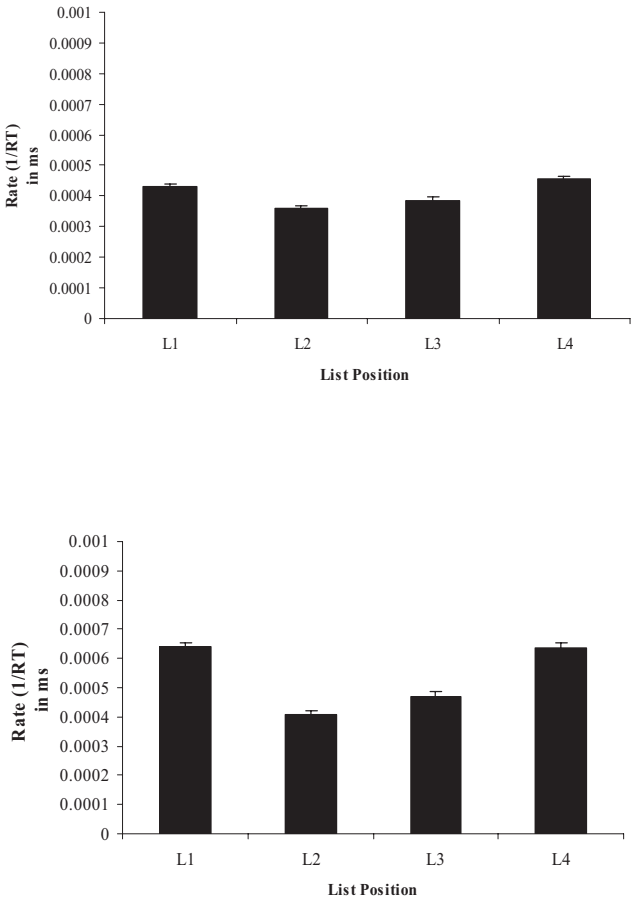


*Figure 5.* Top: Average reaction time (RT) rate to make each judgment as a function of list. Participants were faster at making judgments on L1 and L4 when little PI was present than on L2 and L3 when PI was greatest. Bottom: Average RT rate to make judgments of irrelevant items as a function of list. Participants were faster at making judgments on L1 and L4 when little PI was present than on L2 and L3 when PI was greatest. L1 = List 1; L2 = List 2; L3 = List 3; L4 = List 4, the final release from PI list.

judgments decreased. It is important to note that judgments increased on the release from PI trial.

Second, relative accuracy decreased as PI increased, but it increased back to the level of List 1 on the release from PI list. That relative accuracy increased on the release from PI list indicates that the reduction of relative accuracy was directly related to the buildup of PI and not to a general inability to retrieve information from successive lists or a general tendency for participants' judgments to become less accurate the more time they spent on the task. We argue that variability due to PI should have been eliminated on the release from PI trial.

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.

Note that the pattern of participants' judgments, as plotted in Figure 2, suggests that judgments are more accurate (i.e., less subadditive) under conditions of high proactive interference (Lists 2 and 3). Indeed, this conclusion would be warranted if one ignored the fact that relative accuracy is *poorer* under conditions of high proactive interference, as illustrated in Figure 4. This illustrates that supposed improvements in absolute accuracy can occur alongside decreases in relative accuracy. Moreover, as implied by our Assumption 2, relative accuracy is dependent on the stochastic nature of the discrimination process, whereas absolute accuracy is dependent on the strength of what participants fail to discriminate. This finding is important in that it suggests that relative accuracy and absolute accuracy can be independently manipulated. That is, one should be able increase or decrease judgment magnitude by manipulating the strength of the irrelevant items, without having a concomitant effect on relative accuracy.

## Experiment 2

The purpose of Experiment 2 was to replicate and to extend Experiment 1 in two important ways. First, participants made probability judgments rather than proportion judgments. We wanted to verify that our results were not specific to instructions to make proportion judgments and that PI can affect probability judgments as well. Second, in Experiment 2, we manipulated the total frequency with which alternatives were presented on each list to manipulate the strength of the prior list irrelevant items. As discussed above, judgments should decrease to a greater extent when one fails to discriminate strong items than when one fails to discriminate weak items.

We developed two conditions within the PI paradigm to manipulate the strength (i.e., absolute presentation frequency) of the prior list items, while holding *relative* frequency constant. 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. 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 hypothesized that in the list-descending condition when the prior list items were stronger than the current list, failing to discriminate a strong prior list item would lead to replacing a current list item with a stronger prior list item. From Equation 3, we expected that including a stronger alternative should produce a decrease in judgment magnitude. Thus, in the list frequency-descending condition, we expected judgments to decrease as PI increased. In contrast, in the list frequency-ascending condition, where the prior list items were weaker than the current list, failing to discriminate a weak prior list item would lead to replacing a current list item with a weaker item. Including a weaker alternative would lead to an increase in judgment magnitude. Thus, in the list frequency-ascending condition, we expected judgments to increase as PI increased. However,

given that discrimination failure can occur independently of the strength of the prior list items,[7] we hypothesized that relative accuracy would be affected by the buildup of PI to the same degree, irrespective of the strength of the prior list items. Thus, we did not anticipate an effect of list strength on relative accuracy. As mentioned previously, these hypotheses were validated using simulation methodology (see the Appendix). Because WM capacity showed little relationship to judgment in Experiment 1, we chose not to measure WM in Experiment 2.

## Method

### Participants

The 77 University of Maryland undergraduate students who participated in Experiment 2 received course extra credit.

### Materials

The materials were the same as those used in Experiment 1, except that 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 Experiment 2 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.

### Design and Procedure

The design was a 2 (list frequency: ascending or descending) × 4 (list: 1–4) × 8 (item frequency) mixed factorial design with list and item frequency manipulated within subjects and list frequency manipulated between subjects. Participants were randomly assigned to 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 Experiment 2. 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. Participants were instructed to base their judgments on what they learned in the observation phase. 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, we manipulated the number of alternatives on each list in an attempt to reduce the effect of learning the set size on judgment magnitude. Rather than always having eight 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 eight alternatives on each list and use this information when making probability judgments. The number of presentations of extra item(s) always summed to five. In other words, when one extra alternative was shown in a PI buildup list, it was presented five 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). Although the total list frequency was manipulated, the relative frequency was held constant. For example, in the ascending condition, List 1 consisted of eight items distributed as 1, 2, 3, 4, 5, 6, 7, 8. The second list was equal to List 1 multiplied by 2, with the eight items distributed as 2, 4, 6, 8, 10, 12, 14, 16. List 3 was equal to List 1 multiplied by 3, with frequencies distributed as 3, 6, 9, 12, 15, 18, 21, 24. For the descending condition, the order was reversed, with the third distribution in the ascending condition occurring first. The release from PI list (List 4) was the same for both the ascending and descending conditions and always had 72 items distributed as 2, 4, 6, 8, 10, 12, 14, 16. Note that the ratio of each item in a list to other items in its list remained constant across lists. For instance, in the ascending condition, the ratio of the presentation frequency of the weakest item in each list to the summed frequency of all other items in the list was 1/35 for the first list, 2/70 for the second list, and 3/105 for the third list, all equal ratios.

## Results and Discussion

The data from 5 participants were not included in this study for the following reasons. One participant typed words instead of giving probability responses, 3 participants gave judgments of only 100 or 0, and 1 participant's judgment sums were greater than 2.5 standard deviations above the mean on the fourth list.[8]

### Manipulation Check: Recall

We hypothesized that the buildup of PI would lead to a decrease in recall and then would increase again at release from PI recall. The right section of Table 1 presents the mean number of words recalled (out of eight) for each of the four lists as a function of list frequency condition. A main effect of list was found, $F(3, 210) = 15.86$, $p < .0001$, $\hat{\omega}^2 = .13$, and a significant interaction between list frequency and list was found, $F(3, 210) = 3.16$, $p < .05$, $\hat{\omega}^2 = .02$. 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, $F(3, 207) = 3.00$, $p < .05$, $\hat{\omega}^2 = .02$. These results suggest that participants' recall was more affected by PI in the list frequency-descending condition than in the list frequency-ascending condition. Perhaps this finding occurred because alternatives on the previous lists had been presented more frequently than alternatives on the current list in the descending

[7] We assume the probability of generating an alternative is a function of its memory strength. However, whether an alternative is discriminated as being irrelevant may be unrelated to memory strength. That is, the *p*(fail to discriminate | strong alternative generated) = *p*(fail to discriminate | weak alternative generated), but *p*(generate strong alternative) ≥ *p*(generate weak alternative).

[8] The participant excluded due to being an outlier was in the descending group. The mean judgment sums for the descending group for List 4 was 265.58 ($SD = 152.59$), and the participant's judgment sum for List 4 was 720.

condition. Consequently, previous list alternatives were likely to be retrieved and to compete with current list items for recall in the descending rather than in the ascending condition. Trend analyses revealed that a quadratic trend fit the changes in recall well, $F(1, 70) = 42.77$, $p < .0001$. Recall decreased as PI increased and 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 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, $F(3, 210) = 8.03$, $p < .0001$, $\hat{\omega}^2 = .07$. 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, 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. Fewer intrusions were reported in Experiment 2 when participants were not under time pressure than in Experiment 1 when participants were under time pressure, suggesting that time pressure impaired participants' ability to monitor if their recall output was a current list item or an intrusion in Experiment 1.

### *Judgment Magnitude*

It was hypothesized that a significant interaction between list and list frequency would be found, such that judgment sums would decrease in the descending condition and increase in the ascending condition (see the Appendix for a simulation of hypothesis). Figure 6 presents the mean judgment sums for each of the four lists as a function of list frequency condition. A main effect of list on judgment sums was found, $F(3, 210) = 2.94$, $p < .05$, $\hat{\omega}^2 = .02$, and a significant interaction between list and list frequency was found, $F(3, 210) = 13.64$, $p < .0001$, $\hat{\omega}^2 = .12$. A main effect of list frequency was also found, $F(1, 70) = 7.99$, $p < .001$, $\hat{\omega}^2 = .02$. Participants' judgment sums differed on the practice list before PI was introduced, $t(70) = 2.45$, $p < .05$. Thus, analyses were performed using the practice list as a covariate to examine whether the effects existed independent of individual differences in judgment sums. When practice list judgment sums were used as a covariate, a significant interaction between list and list frequency condition was still found, $F(3, 207) = 13.44$, $p < .0001$, $\hat{\omega}^2 = .11$. However no significant main effects of list frequency or of list were found.

The univariate analysis examining the main effect of list on judgment sums within the ascending condition was significant, $F(3, 105) = 3.53$, $p < .025$, as was the main effect of list on judgment sums in the descending condition, $F(3, 105) = 15.56$, $p < .0001$. A significant quadratic trend was found for the effect of list in the ascending condition, $F(1, 35) = 7.02$, $p = .012$. In the descending condition, significant linear, $F(1, 35) = 18.77$, $p < .0001$, and quadratic, $F(1, 35) = 19.35$, $p < .0001$, trends were found. As predicted, in the ascending condition, judgment sums significantly increased on Lists 2 and 3 and then decreased again on List 4. In the descending condition, judgment sums significantly decreased on Lists 2 and 3 and then increased again on List 4. These results suggest that discrimination failure occurred and that in the descending condition, when the prior list items were stronger than the current list, failing to discriminate strong prior list items replaced weaker current list items. Including a stronger alternative led to a decrease in judgment magnitude. In contrast, in the list frequency-ascending condition, the irrelevant alternatives that participants included in the comparison process tended to be less strong than the relevant alternatives they replaced. Including a weaker alternative led to an increase in judgment magnitude. These results provide strong evidence that discrimination failure occurred.

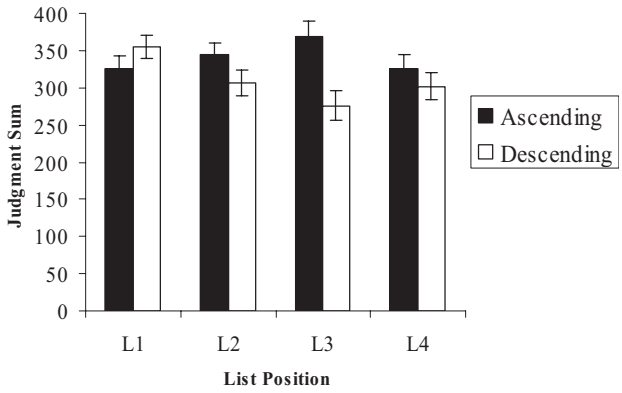


*Figure 6.* Subadditivity of list-relevant probability judgments as a function of list and list frequency. Judgment sums on the practice list were used as a covariate. L1 represents List 1, L2 represents List 2, L3 represents List 3, and L4 represents List 4, the final release from PI list. In the descending condition, as proactive interference (PI) increased on L2 and L3, judgment sums decreased. Then on the final release from PI list, judgment sums increased again. In the ascending condition, judgment sums increased on L2 and L3 and then decreased on L4.

### *Relative Accuracy*

It was hypothesized that Somers's D correlations between participants' judgments and the objective probabilities would decrease as PI increased, but list frequency would have no effect on relative accuracy if the probability of discrimination failure was independent of the strength of the items on the prior lists.

Figure 7 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, $F(3, 210) = 47.40$, $p < .0001$, $\hat{\omega}^2 = .32$. Further, trend analyses revealed that both a linear trend, $F(1, 70) = 8.38$, $p < .005$, and a quadratic trend, $F(1, 70) = 14.70$, $p < .0001$, fit the changes in relative accuracy. Relative accuracy was highest on Lists 1 and 4 when PI was least 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 using the practice list as a covariate, a marginally significant main effect of list was still found, $F(3, 207) = 2.36$, $p = .072$, $\hat{\omega}^2 = .02$.

Thus, as predicted, participants' relative accuracy decreased as PI increased, but the degree to which relative accuracy decreased was unaffected by list frequency.

### *Irrelevant Items*

Participants' judgments of previous list irrelevant items provide further evidence for discrimination failure. 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. The two rightmost columns of Table 3 present the mean number of previous list items (out of four possible) given nonzero judgments for each of the four lists, as a function of list frequency condition, for Experiment 2. As in Experiment 1, a main effect of list was found, $F(3, 210) = 25.75$, $p < .0001$, $\hat{\omega}^2 = .20$. Participants judged significantly more irrelevant items as relevant on Lists 2 and 3 when PI was greatest than on Lists 1 and 4. No significant interaction between list and list frequency was found, nor was a main effect of list frequency found on the number of irrelevant items given nonzero judgments. Taken together with the relative accuracy data given earlier, this suggests that list strength did not affect the probability of discrimination failure.

### *RT*

We hypothesized that judgments would be slowest when PI was greatest on Lists 2 and 3 and that judgments would be fastest when PI was least on Lists 1 and 4. RTs for each judgment were transformed to rates (1/RT) to reduce the skewness of the distribution. Figure 8 (top) 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, $F(3, 210) = 48.78$, $p < .0001$, $\hat{\omega}^2 = .33$. In contrast to Experiment 1, no quadratic trend was found.[9] Figure 8 (bottom) presents the mean irrelevant judgment RT rate for each of the four lists. A significant main effect of list on the RT to judge irrelevant items (items that did not

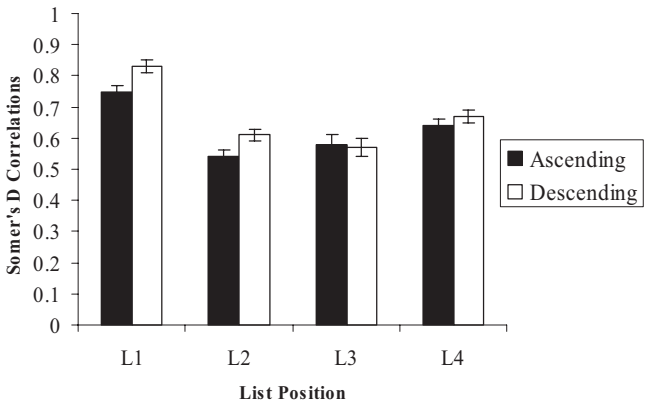


*Figure 7.* Relative accuracy of judgments (as measured by Sommer's $D_{xy}$ correlations between proportion judgments and objective probabilities) as a function of list and list frequency. Relative accuracy on the practice list was used as a covariate. A main effect of list was found, such that as proactive interference (PI) increased on L2 and L3, relative accuracy decreased. Then on the release from PI list, relative accuracy increased again. L1 = List 1; L2 = List 2; L3 = List 3; L4 = List 4, the final release from PI list.

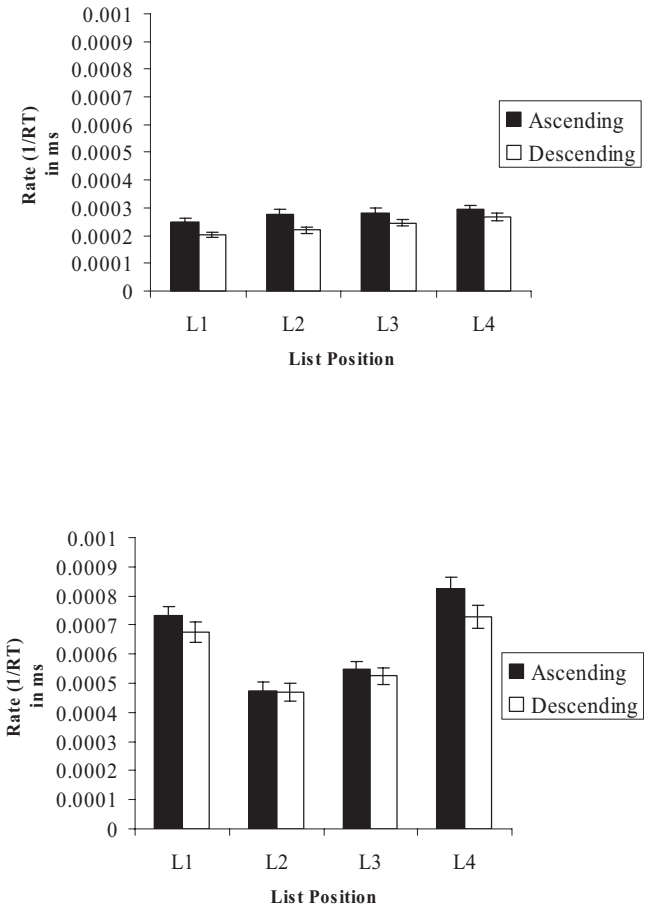


*Figure 8.* Top: Average reaction time (RT) rate to make each judgment as a function of list and list frequency. Participants were faster at making judgments on L1 and L4 when little proactive interference (PI) was present than on L2 and L3 when PI was greatest. Bottom: Average RT rate to make judgments of irrelevant items as a function of list. Participants were at making judgments on L1 and L4 when little PI was present than on L2 and L3 when PI was greatest. L1 = List 1; L2 = List 2; L3 = List 3; L4 = List 4, the final release from PI list.

[9] Note that in Experiment 1, the strength of the quadratic trend for the overall RTs was less than the strength of the quadratic trend for RTs for the irrelevant alternatives. Thus, our failure to find a significant effect for the overall RTs may merely reflect the possibility that the effect size is expected to be smaller for the overall RTs compared with the RTs for the irrelevant alternatives. This may well be because the overall RTs are an average of the relevant and irrelevant RTs. To the extent to which the effect of PI on the overall RTs is due to only the irrelevant alternatives, averaging over both relevant and irrelevant alternatives would reduce the effect size. Thus, the lack of a statistically significant quadratic trend for the overall RTs merely suggests that the effect size is smaller when RTs are pooled across the relevant and irrelevant alternatives.

occur on the list in question) was also found, $F(3, 210) = 72.26$, $p < .0001$, $\hat{\omega}^2 = .43$. Further, a significant quadratic trend was found, $F(1, 70) = 144.85$, $p < .0001$. Participants were slower at judging irrelevant items on Lists 2 and 3, which had the most interference, than on 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. 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 buildup of PI.

### *Summary of Experiment 2*

Experiment 2 revealed three main findings consistent with the discrimination failure account. First, in the descending condition, as PI increased, judgments decreased, which is consistent with the account that participants failed to discriminate relevant from irrelevant alternatives and that strong irrelevant alternatives supplanted weaker relevant alternatives in the probability judgment. In the ascending condition, as PI increased, judgments increased, which is consistent with the account that weak irrelevant alternatives supplanted stronger relevant alternatives in the probability judgment. The second finding consistent with the discrimination failure account was that relative accuracy decreased as PI increased in both of the list frequency-ascending and -descending conditions, but it increased on the release from PI list. This is consistent with the idea that discrimination failure increases the error inherent in participants' judgments. 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.

As in Experiment 1, we illustrated that relative and absolute accuracy can be dissociated. Specifically, Experiment 2 showed that absolute accuracy is sensitive to the strength of the prior list items but that relative accuracy was completely unaffected by list strength. Indeed, the patterns of results shown in Figures 5 and 6 illustrate that prior list strength can increase or decrease judgment magnitude, whereas PI always decreases relative accuracy.

One interesting finding is that judgments in Experiment 2 were considerably higher than in Experiment 1. The only appreciable difference between these two experiments that would affect the magnitude, aside from PI, was that participants made proportion judgments in Experiment 1 whereas they made probability judgments in Experiment 2. That participants' judgments were excessively high in Experiment 2 is in keeping with past research. For example, Dougherty and Hunter (2003a; see also Dougherty & Hunter, 2003b) found that the average of the sum of judgments for a set of eight mutually exclusive and exhaustive hypotheses was about 260%. Although speculative, one possible explanation for the differences between proportion and probability judgments in Experiments 1 and 2 is that proportion judgment might make the requirement of additivity more explicit. Aside from the differences in the judgment magnitude, however, the two types of judgments (particularly the RTs and relative accuracy measures) appear to be affected similarly by PI. This suggests that the two types of judgments arise from the same basic cognitive processes, with perhaps the only difference being that the magnitude of probability judgments is an additive function of proportion judgments.

## General Discussion

The purpose of this research was to examine the effect of irrelevant information on probability judgment. To this end, we introduced two new theoretical accounts, the discrimination failure account and inhibition failure account, which demonstrate how information from outside the relevant sample space can bias probability judgments. As an empirical demonstration of the discrimination failure account, our experiments used an adaptation of the PI paradigm and showed that information from outside the relevant sample space (in our case, prior lists in the PI task) can have systematic effects on both the magnitude and relative accuracy of participants' judgments. Specifically, we showed that the failure to discriminate between sets of hypotheses led to an increase or decrease in the magnitude of participants' probability judgments, depending on the strength of the irrelevant alternatives. We also showed that the failure to discriminate among sets of information consistently led to a decrease in the relative accuracy of participants' judgments. Clearly, discrimination failure and inhibition failure are not mutually exclusive, and it may be possible for both to operate in conjunction. For example, it is possible for participants to generate a set of irrelevant alternatives, discriminate some but not others, and fail to inhibit a subset of the ones that were identified as irrelevant. Although this clearly is possible, our results, particularly the results of Experiment 2, suggest that discrimination failure was the primary factor underlying the effect of PI on judgment.

Note that our accounts of judgment predict, and our data confirm, that information from outside of the relevant sample space can have systematic effects on participants' probability judgments. This finding is a violation of one of the core principles of probability theory and is inconsistent with nearly all descriptive, prescriptive, and normative models of judgment. Indeed, most models of probability assume the events entering into the computation of a probability are drawn from a common and well-defined event space. Certainly this is true for normative models of probability judgment, including Bayesian statistics. It is also true of descriptive models of judgment. For instance, support theory (Tversky & Koehler, 1994) assumes $\{A, B\} \in S$, and that the computation of $P(A, B)$ is based on only the set of elements within $S$. Consequently, support theory does not anticipate that elements from an irrelevant sample space ($S'$) can influence judgment. Moreover, even if support theory were modified to handle the influence of irrelevant information on judgment, it would be difficult for the model to handle both discrimination and dysinhibition biases without specifying mechanisms for how irrelevant information is treated by the decision maker. To account for the effects of discrimination failure and inhibition failure on judgment, support theory needs to be specified within the context of a cognitive architecture that incorporates discrimination and inhibition processes (Thomas, Dougherty, & Sprenger, 2005).

What our theoretical framework suggests is that the subjective event space created by the participant can be ill-defined and inconsistent. The subjective event space is ill-defined whenever the participant fails to discriminate between elements drawn from the appro-

priate and inappropriate event spaces. Although one can develop paradigms, as we have here, where the objective event space is known, the assumption that participants' subjective event space maps onto the objective event space is tenable at best. In fact, the discrimination failure account suggests that the subjective event space over which probability judgments are made is constructed by the participant and can include elements drawn from irrelevant event spaces. Moreover, as suggested by the inhibition failure model, it is possible for participants to accurately discriminate between events drawn from the relevant and irrelevant event spaces and therefore have a subjective event space that maps onto the objective event space, and yet still be influenced by elements from outside the relevant event space. The distinction between biases arising from discrimination failure (i.e., the discrimination bias) and biases arising from inhibition failure (i.e., dysinhibition bias) is important, because it indicates that descriptive models of judgment need to postulate mechanisms for dealing with events from outside the relevant event space.

In this article, we proposed that dysinhibition bias would result from an inability to inhibit irrelevant alternatives from the focus of attention. Previous research has shown that WM capacity is related to the ability to inhibit irrelevant information (Lustig, May, & Hasher, 2001; May, Hasher, & Kane, 1999). To the extent this is the case, we would have expected our measure of WM span to correlate with judgment magnitude had our judgment task involved inhibition. Thus, the finding that WM capacity was not related to the sum of judgments in Experiment 1 suggests that the effect of PI on judgment was not due to inhibition failure. In contrast, the discrimination failure account does not predict a relationship between WM span and judgments as a function of list. Therefore, our null finding with respect to WM span is consistent with the discrimination failure account and is inconsistent with the inhibition failure account.

The research presented in this article provides support for the discrimination failure account. However, one question of interest is whether the inhibition failure hypothesis offers a plausible account of other research findings. We hypothesize that inhibition processes become important any time the decision maker can accurately discriminate between the relevant and irrelevant lists. In the present experiments, the only information on which to make the discrimination judgment was on the time of the learning (whether the item was learned on the most recent study interval or not). One way to make the lists more separable might be to introduce additional information on which a discrimination judgment can be made. For example, one might be able to develop a paradigm that better enabled participants to discriminate among the categories (i.e., lists), but where the categories were sufficiently similar to one another that they relevant and irrelevant categories competed with each other to gain access to the focus of attention. Dysinhibition bias would be introduced to the extent to which the decision maker was unable to prevent the irrelevant alternatives from gaining access to the focus of attention.

Although direct empirical tests of the inhibition hypothesis are presently being conducted, a few prior studies are consistent with the inhibition failure hypothesis. Dougherty and Hunter (2003a, 2003b) argued that subadditivity arises due to participants' inability to generate and compare all possible alternatives at one time, which presumably is due to WM constraints. They demonstrated that WM span was negatively correlated with the subadditivity of participants' judgments. However, it is possible that the storing capacity of WM is not the only factor reducing one's ability to generate and compare all possible alternatives when making probability judgments. It is also possible that one's failure to inhibit alternative hypotheses that are irrelevant keeps one from considering other possible relevant hypotheses. In fact, that WM span was unrelated to the subadditivity of judgments on the practice list of Experiment 1 in the present study suggests that WM relates to judgment only in conditions in which inhibition is necessary. In the practice list of Experiment 1, there was only one distribution of alternatives for participants to consider, and thus generation of irrelevant items was unlikely. In contrast, in Dougherty and Hunter's (2003a, 2003b) studies, participants learned four distributions of items (breakfast, snack, dinner, and dessert items), and it was therefore possible that alternatives from the three nonrelevant distributions could be generated when judging an item from one distribution. Participants in Dougherty and Hunter's study presumably could easily discriminate items they generated (i.e., lunch items should be easily discriminated from dessert items), but perhaps they could not inhibit those irrelevant items. If inhibition failure occurred, and inhibition is related to WM capacity, the relationship between WM and the subadditivity of judgments could have been due to inhibition failure rather than to the raw capacity to maintain items in WM. Note that it is likely that WM capacity relates both to one's capacity to compare multiple alternatives at one time as well as to one's ability to inhibit irrelevant alternatives from the comparison process.

A second set of studies that can be conceptualized with the inhibition failure framework is the work by Windschitl and Chambers (2004) on the dud alternative effect. Windschitl and Chambers examined the effect of adding highly unlikely alternatives (duds) to the set of alternatives to consider when making a judgment. They found that adding dud alternatives to the set of alternative hypotheses led to increased judgments of the focal hypothesis rather than to decreased judgments, as predicted by support theory. They proposed that the dud alternative effect was due to participants engaging in a comparison process by which they sequentially compare the focal hypothesis alternative hypotheses and that the number of times the focal is stronger than the alternative drives their judgment. Thus, including dud alternatives increases the number of times the focal compares favorably to alternative hypotheses and increases support for the focal hypothesis. However, the dud alternative effect can also be conceptualized within the context of the inhibition failure account. Perhaps the presence of the dud alternative prevents the participant from adequately assessing the support for the nondud alternative. In essence, one might propose that the assessment of the support for the dud alternative interferes with the assessment of the support for the nondud alternative.

Our examination of irrelevant information in probability judgment has parallels to research in the choice literature investigating violations of the normative principle of *independence of irrelevant alternatives* (Debreu, 1960). The independence principle states that participants' choice between two options (*A* and *B*) should be independent of *C*. However, a number of empirical findings, collectively called *decoy effects*, have revealed violations of the independence principle, where introducing *C* to the choice set affects one's preferences between *A* and *B*, even when *A* and *B* are both preferred

to *C* (Huber, Payne, & Puto, 1982; Simonson, 1989; Sjoberg,1977; Tversky, 1972). As an example, an individual might choose dessert A over dessert B but switch his or her preference to dessert B when dessert C is introduced (Ariely & Wallsten, 1995). Note, however, that dessert C is drawn from the same sample space as A and B, that is $\{A, B, C\} \in S$, and therefore is a legitimate option. Although the choice between *A* and *B* should be independent of *C*, *C* is not completely irrelevant because it is a member of the choice set.[10] An interesting question concerns whether people are equally influenced by *C* when *C* is drawn from an irrelevant sample space and is not a legitimate member of the choice set (e.g., $\{A, B\} \in S$ and $\{C\} \in S'$). Our theoretical framework suggests the effect of *C* is due to people's inability to exclude it from influencing the comparison between *A* and *B*. Thus, the degree to which participants can effectively inhibit *C* should be related to the probability that participants would switch preferences when *C* is introduced (for a discussion of the role of inhibitory processes in choice, see Roe, Busemeyer, & Townsend, 2001; Usher & McClelland, 2004).

### *Relative Versus Absolute Measures of Accuracy*

One important finding in this article is the observed dissociation between absolute (judgment magnitude) and relative accuracy. Note that absolute accuracy relates to the additivity of participants' judgments, whereas relative accuracy relates to how good participants are at discriminating high probability hypotheses from lower probability hypotheses. Thus, both measures reflect different ways of defining judgment accuracy. A good judge should, presumably, have judgments that are both additive and well correlated with the objective probabilities.

Of importance for theorists concerned with judgment accuracy is whether these two measures are based on the same underlying processes. To the extent that relative and absolute accuracy are based on the same processes, variables that lead participants to give judgments that are closer to additivity should also lead to improvements in the correlation between the objective and subjective probability judgments. However, as we showed in both Experiments 1 and 2, the magnitude of participants' judgments, that is, the degree to which participants' judgments were subadditive, could be affected independently of the relative accuracy of their judgments. Indeed, in Experiment 1, we showed that participants became less subadditive, simultaneously with producing judgments that were more poorly correlated with the objective probabilities. Experiment 2 extended this finding to show that the strength of the items on the prior lists can lead to increases or decreases in participants' judgments, but that relative accuracy always decreased as a function of the buildup of PI. This finding is important because it shows that these two measures of accuracy are based on different processes, and it suggests that conclusions based on measures of absolute accuracy may not be consistent with conclusions based on relative accuracy. Thus, as pointed out by Treadwell and Nelson (1996), how one defines *accuracy* is pivotal to conclusions regarding whether judgments are accurate or not.

Although prior research has shown that measures of relative and absolute accuracy can sometimes yield different conclusions (Meeter & Nelson, 2003), to the best of our knowledge our experiments are the first to illustrate this dissociation within a single task and in a within-subjects design. For example, Treadwell and Nelson (1996) found a dissociation between absolute and relative accuracy in a study comparing single item probability judgments to aggregate frequency judgments. They measured both the ordinal gamma correlation between participants' judgments and the correctness of their responses (relative accuracy) as well as the magnitude of participants' judgments (absolute accuracy). They found that although the relative accuracy of single item probability judgments was significantly higher than that of aggregate frequency judgments, the absolute accuracy (magnitude) of single item probability judgments was significantly lower than that of aggregate frequency judgments. Treadwell and Nelson's findings, therefore, illustrate a *task* dissociation, in which different types of judgment tasks (single item confidence vs. aggregate frequency judgments) yield different conclusions regarding judgment accuracy. In contrast, our experiments demonstrated a *process* dissociation, in which a single manipulation (the buildup of PI) can simultaneously improve absolute accuracy (participants became less subadditive) and decrease relative accuracy (lower correlations between objective and subjective judgments).

### *Implications for PI*

Our findings are relevant not only for theories of probability judgment but also for theories of PI. Two types of competing theories of PI exist, those that explain PI as resulting from a failure to discriminate temporally (Baddeley, 1990; Wixted & Rohrer, 1993; Underwood, 1945) and those that explain PI as resulting from a failure to suppress interfering information (Anderson & Neely, 1996; Kane & Engle, 2000; Postman & Hasher, 1972). Temporal discrimination theories argue that the buildup of PI reflects a growing impairment in the ability to distinguish items that appear on the most recent list from those that appeared on previous lists. In contrast, suppression theories argue that overcoming PI requires active suppression of competitors at retrieval (Anderson & Neely, 1996) and that persons who are better able to inhibit competition from prior list items are in turn less susceptible to the effects of PI (Kane & Engle, 2000). Note that our discrimination failure hypothesis relates to the temporal discrimination failure account of PI, whereas our inhibition failure hypothesis relates to the suppression failure account of PI. Thus, our findings that participants failed to discriminate relevant from irrelevant alternatives when making probability judgments support the temporal discrimination accounts of PI.

Although our studies support temporal discrimination theory, we are hesitant to draw conclusions regarding these two theories beyond the specific version of the PI task used in our experiments. There were a number of differences between our experiments and the standard PI paradigm that may limit the generalizability of our results to theories of PI. For example, in our experiments, the only cues available to help participants discriminate between lists were temporal cues. Further, the length of time to present items was longer in our experiments than in normal PI experiments, because

[10] Our use of the term *irrelevant alternative* is slightly different from how its used within the normative principle of *independence of irrelevant alternatives,* which states that the choice between two alternatives should be independent of a third alternative (Debreu, 1960). We define irrelevance more strictly in terms of whether the third alternative belongs to the choice set.

items were each presented multiple times rather than only once. And, as argued by Glenberg and Swanson (1986), longer time between study and test reduces the effectiveness of temporal cues to discriminate between lists. Although our specific experimental findings probably cannot shed much light on theories of PI in standard recall tasks, our methodology of adapting the PI paradigm for probability judgment might provide a testing ground for these theories. The main advantage of our adaptation of the PI paradigm is that, within the context of Equations 2 and 3, inhibition failure and discrimination failure yield distinct (and opposite) predictions. In contrast, although there are subtle differences in predictions between the inhibition and discrimination failure explanations of PI within recall tasks, both theories yield similar predictions regarding recall accuracy.

As a final note, it is worth pointing out that our use of the PI paradigm to explore processes in probability judgment demonstrates the utility of conceptualizing judgment processes within the broader literature of memory and cognition. Indeed, our use of memory theoretic constructs and methodologies developed in the memory literature to study memory phenomena have proven useful for developing and testing models of judgment. As alluded to earlier, theories and methodologies within the judgment literature arguably also would be useful for testing models of memory.

## References


Anderson, M. C., & Neely, J. H. (1996). Interference and inhibition in memory retrieval. In E. L. Bjork & R. A. Bjork (Eds.), *Memory* (pp. 237–313). San Diego, CA: Academic Press.

Ariely, D., & Wallsten, T. S. (1995). Seeking subjective dominance in multidimensional space: An exploration of the asymmetric dominance effect. *Organizational Behavior & Human Decision Processes, 63,* 223–232.

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.

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, 57*(A), 1059–1084.

Debreu, G. (1960). Review of R. D. Luce: Individual choice behavior. *American Economic Review, 50,* 186–188.

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. (2003a). Hypothesis generation, probability judgment, and individual differences in working memory capacity. *Acta Psychologica, 113,* 263–282.

Dougherty, M. R. P., & Hunter, J. E. (2003b). Probability judgment and subadditivity: The role of working memory capacity and constraining retrieval. *Memory & Cognition, 31,* 968–982.

Elstein, A. S., Shulman, L. S., & Sprafka, S. A. (1978). *Medical problem solving: An analysis of clinical reasoning.* Cambridge, MA: Harvard University Press.

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

Gettys, C. F., & Fisher, S. D. (1979). Hypothesis plausibility and hypothesis generation. *Organizational Behavior & Human Performance, 24,* 93–110.

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.

Hintzman, D. L. (1970). Effects of repetition and exposure duration on memory. *Journal of Experimental Psychology, 83,* 435–444.

Horton, D. L., & Mills, C. B. (1984). Human learning and memory. *Annual Review of Psychology, 35,* 361–394.

Huber, J., Payne, J. W., & Puto, C. (1982). Adding asymmetrically dominated alternatives: Violations of regularity and the similarity hypothesis. *Journal of Consumer Research, 9,* 90–98.

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.

Kahneman, D., & Tversky, A. (1982). The simulation heuristic. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 201–208). New York: Cambridge University Press.

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, and Cognition, 26,* 336–358.

Koehler, D. K. (1991). Explanation, imagination, and confidence in judgment. *Psychological Bulletin, 110,* 449–519.

Koriat, A., Lichtenstein, S., & Fischhoff, B. (1980). Reasons for confidence. *Journal of Experimental Psychology: Human Learning & Memory, 6,* 107–118.

Libby, R. (1985). Availability and the generation of hypothesis in analytical review. *Journal of Accounting Research, 23,* 648–667.

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.

Meeter, M., & Nelson, T. O. (2003). Multiple study trials and judgments of learning. *Acta Psychologica, 113,* 123–132.

Mehle, T. (1982). Hypothesis generation in an automobile malfunction inference task. *Acta Psychologica, 52,* 87–106.

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.

Roe, R. M., Busemeyer, J. R., & Townsend, J. T. (2001). Multialternative decision field theory: A dynamic connectionist model of decision-making. *Psychological Review, 108,* 370–392.

Roediger, H. L. III, & Challis, B. H. (1992). Effects of exact repetition and conceptual repetition on free recall and primed word-fragment completion. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18,* 3–14.

Roese, N. J., & Olson, J. M. (1996). Counterfactuals, causal attributions, and the hindsight bias: A conceptual integration. *Journal of Experimental Social Psychology, 32,* 197–227.

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

Simonson, I. (1989). Choice based on reasons: The case of attraction and compromise effects. *Journal of Consumer Research, 16,* 158–174.

Sjoberg, L. (1977). Choice, frames and similarity. *Scandinavian Journal of Psychology, 18,* 103–115.

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

Sprenger, A., & Dougherty, M. R. (2006). Differences between probability and frequency judgments: The role of individual differences in working memory capacity. *Organizational Behavior and Human Decision Processes, 99,* 202–211.

Thomas, R. P., Dougherty, M. R., & Sprenger, A. M. (2005). *Diagnostic hypothesis generation and human judgment.* Manuscript submitted for publication.

Treadwell, J. R., & Nelson, T. O. (1996). Availability of information and the aggregation of confidence in prior decisions. *Organizational Behavior and Human Decision Processes, 68,* 13–27.

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

Tversky, A. (1972). Elimination by aspects. *Psychological Review, 79,* 281–299.

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.

Usher, M., & McClelland, J. L. (2004). Loss aversion and inhibition in dynamical models of multialternative choice. *Psychological Review, 111,* 757–769.

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.

Weber, E. U., Böckenholt, U., & Hilton, D. J. (1993). Determinants of diagnostic hypothesis generation: Effects of information, base rates, and experience. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19,* 1151–1164.

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., & Chambers, J. R. (2004). The dud-alternative effect in likelihood judgment. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30,* 198–215.

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

## Appendix

We conducted a set of simulations to validate our hypotheses that inhibition failure and discrimination failure have differential effects on both judgment magnitude and relative accuracy. Our simulations were based on the experimental designs used in Experiments 1 and 2 and by incorporating Assumptions 1 and 2. The methodology used in the simulations is given below.

### Method

Two sets of simulations, one to simulate the discrimination failure account and one to simulate the inhibition failure account, were conducted for Experiments 1 and 2. Each simulation used 200 simulated participants.

Recall that Assumption 1 pertains to the number of hypotheses that can be held in working memory (WM). For all simulations, we set a WM capacity parameter to five items. Thus, for each simulated participant, four alternatives plus the focal were used in the computation of judged probability.

Assumption 2 states that the probability that irrelevant hypotheses included in the comparison are stronger than the relevant hypotheses they replace is greater than the probability of the converse. We instantiated this assumption by assuming that the probability of an item being sampled from memory was proportional to its relative frequency (i.e., how often it occurred on the study list). Thus, an item that occurred eight times on the list was twice as likely to be sampled than an item that occurred four times. In addition to setting the sampling of individual items as being proportional to its relative frequency, we also included a parameter $r$ that reflects the probability that an item is sampled from the relevant list. For all simulations detailed below, the probability that an item was sample from the relevant list was $r = .70$, and the probability that an item was sampled from an irrelevant (prior list) was $1 - r = .30$. Although we report the simulations for $r = .70$, the simulations were replicated using a variety of values of $r$. The results for $r = .70$ are representative of other values.

For our simulations, we calculated the judged probability of the focal by implementing Equation 3 to model the discrimination account and Equation 4 to model the inhibition failure account. For simplicity, we assumed that the judged probability was based on the true frequencies.

#### *Discrimination Failure*

To simulate the discrimination failure account, we assumed that there was a probability $h$ that participants would identify a focal hypothesis sampled from the relevant set as actually coming from the relevant set and some probability $c$ that participants would identify a hypothesis that had been sampled from the irrelevant set as actually being irrelevant. Within the context of our experiment, these two probabilities correspond to the proportion of items that participants gave a nonzero judgment, when in fact the item had a nonzero probability (i.e., $h = p$[nonzero judgment | relevant item] = the hit rate), and the proportion of items that participants gave judgments of zero when in fact the item had a zero probability (i.e., $c = p$[zero judgment | irrelevant item] = the correct rejection rate). Empirically, we observed $p$[nonzero judgment | relevant item] = .95, and $p$[zero judgment | irrelevant item] = .67, in Experiment 1. Consequently, for our simulations, we set $h = .95$ and $c = .67$. These same values were used in our simulations of Experiment 2. Thus, $h$ and $c$ were not allowed to vary across experiments.

For the discrimination failure account, we assumed that irrelevant items that were correctly identified as irrelevant and relevant items incorrectly identified as irrelevant were given judgments of zero. For all other cases, the judged probability of the focal hypothesis was given by the frequency of the focal hypothesis divided by the sum of the frequencies of all five items in WM.

#### *Inhibition Failure*

The inhibition failure simulations were similar to the discrimination failure simulations except for the following differences. First, we assumed that discrimination was perfect (i.e., $h = c = 1.0$). Thus, relevant focal hypotheses were always given nonzero judgments and irrelevant focal hypotheses were always given judgments of zero. Discrimination was also

(*Appendix continues*)

assumed to be perfect for determining whether the alternatives were from the relevant or irrelevant sets. Thus, if irrelevant alternative hypotheses were generated, it was assumed that they were correctly identified as irrelevant and not included in the judgment. In the inhibition failure case, irrelevant alternative hypotheses were inhibited with probability $i$, where $i = .7$. Alternatives that were generated and inhibited were removed from WM and replaced with a different hypothesis (which also went through the discrimination and inhibition processes). Irrelevant alternatives that were generated but not inhibited were maintained in WM (and hence acted as placeholders), but they were not included in the computation of judged probability. Judged probability was based only on the relevant items in WM as specified in Equation 4.

## Results

### *Simulation of Experiment 1*

*Judgment sums.* Table A1 presents the predicted sum of judgments for the discrimination failure and inhibition failure cases as a function of list. Note first that the magnitude of the judgments is greater than 100. This occurs in the simulations because only four alternative hypotheses are included in each judgment, leading each judgment to be overestimated. For the inhibition failure case, judgment sums are predicted to increase as PI increases across Lists 2 and 3 and then to decrease again on List 4 when PI is released. This occurs because when irrelevant alternatives are sampled and not inhibited fewer relevant alternative hypotheses are included in the judgment comparison. Consequently, the focal hypothesis is given a higher judgment than if inhibition had occurred. In contrast, in the discrimination failure case, judgment sums are predicted to decrease as PI increases across Lists 2 and 3 and then to increase again on List 4. This occurs for two reasons. First, 5% of the time participants fail to identify relevant hypotheses as relevant and incorrectly give them judgments of zero. Second, when participants sample alternative hypotheses from the irrelevant distribution, they are not identified as irrelevant and are consequently included in the judgment comparison. Because the probability of sampling each alternative is proportional to the relative frequency each alternative occurred on its list, stronger items are more likely to be sampled than weaker items. Thus the alternative hypotheses in the discrimination failure case tend to be stronger than if discrimination failure had not occurred. Consequently, each judgment tends to be lower than if discrimination failure had not occurred.

*Relative accuracy.* Table A2 presents predicted relative accuracy (Somer's D correlations) for the discrimination failure and inhibition failure cases as a function of list. For the inhibition failure case, relative accuracy is predicted to decrease slightly as PI increases across Lists 2 and 3 and then to increase again on List 4 when PI is released. This effect is small and occurs because the nature of sampling and inhibiting irrelevant alternatives is stochastic. One can view this process as adding error to the assessment of probability. Adding error to the assessment of probability decreases correlations between subjective and objective probabilities, because it will perturb the rank order of the to-be-judged items. For the discrimination failure case, relative accuracy is predicted to decrease as PI increases across Lists 2 and 3 and to increase again on List 4 when PI is released. This occurs for several reasons. First, on average, 5% of relevant items are incorrectly identified as irrelevant and given judgments of zero, and 33% of irrelevant items are incorrectly identified as relevant and given nonzero judgments. These discrimination failures at the level of the focal hypothesis directly alter the rank order of subjective judgments and objective proportions and lead to decreases in relative accuracy. Finally, as in the inhibition failure case, the nature of sampling irrelevant alternatives is stochastic, which adds error to the judgment and decreases relative accuracy. An important prediction shown in Table A2 is that relative accuracy is predicted to decrease much more in the discrimination failure case than in the inhibition failure case. This occurs because there are more opportunities for error in the discrimination failure case than for the inhibition failure case.

*Mean number of irrelevant items given nonzero judgments.* In the predictions for the inhibition failure case, both relevant and irrelevant focal items are always correctly discriminated, and therefore the mean number of irrelevant items given nonzero judgments is zero for all lists. In contrast, because the simulations were set to correctly discriminate irrelevant alternatives as irrelevant (i.e., give them judgments of zero) 67% of the time, on average 33% of the time irrelevant alternatives were given nonzero judgments on List 2 ($M = 1.54$, $SE = 0.07$) and on List 3 ($M = 1.47$, $SE = 0.07$).

### *Simulation of Experiment 2*

*Judgment sums.* Table A3 presents the predicted judgment sums obtained in simulation 2. For the inhibition failure case, the predictions are almost identical to those for Experiment 1. Judgment sums are predicted to increase as PI increases. For the discrimination failure case, the predictions are more interesting. In contrast to Experiment 1's predictions for discrimination failure, in the ascending condition case, judgment sums are predicted to increase as PI increases. This occurs because in the ascending condition (compared to Experiment 1 and the descending condition of Experiment 2) irrelevant items that are sampled from a prior list have a higher probability of being weaker than the items they replace (i.e., the ascending condition List 1 items are weaker than List 2 items). In contrast, in the descending condition for the discrimination failure, judgment sums are predicted to decrease as PI increases, even more so than in Experiment 1. This is because items from the previous lists are stronger than the items they replace (i.e., in the descending condition List 1 items are stronger than List 2 items).

Table A1

Predicted Sum of Judgments As a Function of List and Process Type for Experiment 1

| List | Inhibition failure | Discrimination failure |
|---|---|---|
| 1 | 134.78 (0.44) | 134.78 (0.44) |
| 2 | 151.87 (0.82) | 126.57 (0.89) |
| 3 | 150.93 (0.91) | 127.02 (0.90) |
| 4 | 134.89 (0.39) | 134.89 (0.39) |

*Note.* Standard errors of the mean are presented in parentheses.

Table A2

Predicted Relative Accuracy (Somer's D Correlations) As a function of List and Process Type for Experiment 1

| List | Inhibition failure | Discrimination failure |
|---|---|---|
| 1 | 0.96 (0.002) | 0.96 (0.002) |
| 2 | 0.92 (0.004) | 0.68 (0.02) |
| 3 | 0.92 (0.004) | 0.68 (0.01) |
| 4 | 0.96 (0.002) | 0.96 (0.002) |

*Note.* Standard errors of the mean are presented in parentheses.

Table A3

Predicted Sum of Judgments As a Function of List and Process Type for Experiment 2

| | Ascending condition | | Descending condition | |
|---|---|---|---|---|
| List | Inhibition failure | Discrimination failure | Inhibition failure | Discrimination failure |
| 1 | 134.78 (0.44) | 134.78 (0.44) | 134.78 (0.44) | 134.78 (0.44) |
| 2 | 153.00 (1.14) | 145.15 (01.18) | 150.05 (0.85) | 114.13 (0.88) |
| 3 | 150.90 (0.97) | 145.22 (1.10) | 150.76 (0.93) | 97.57 (1.00) |
| 4 | 134.89 (0.39) | 134.89 (0.39) | 134.89 (0.39) | 134.89 (0.39) |

*Note.* Standard errors of the mean are presented in parentheses.

Table A4

Predicted Relative Accuracy (Somer's D Correlations) As a function of List and Process Type for Experiment 2

| | Ascending condition | | Descending condition | |
|---|---|---|---|---|
| List | Inhibition failure | Discrimination failure | Inhibition failure | Discrimination failure |
| 1 | 0.96 (0.002) | 0.96 (0.002) | 0.96 (0.002) | 0.96 (0.002) |
| 2 | 0.92 (0.004) | 0.67 (0.01) | 0.92 (0.004) | 0.67 (0.01) |
| 3 | 0.93 (0.004) | 0.73 (0.01) | 0.93 (0.003) | 0.54 (0.02) |
| 4 | 0.96 (0.002) | 0.96 (0.002) | 0.96 (0.002) | 0.96 (0.002) |

*Note.* Standard errors of the mean are presented in parentheses.

*Relative accuracy.* Relative accuracy predictions for Experiment 2 are almost identical to those of Experiment 1 for both the ascending and descending conditions (see table A4).

*Mean number of irrelevant items given nonzero judgments.* The predictions for the mean number of irrelevant items given nonzero judgments for Experiment 2 are almost identical to those of Experiment 1: List 2 ascending, $M = 1.48$, $SE = 0.07$; List 2 descending, $M = 1.49$, $SE = 0.07$; List 3 ascending, $M = 1.64$, $SE = 0.06$; List 3 descending, $M = 1.50$, $SE = 0.08$).

For the inhibition failure case, we simulated a variety of values of $i$ to ensure that the results were not specific to one particular value. Of note, the inhibition failure account always predicts that judgment sums will increase as PI increases when $i < 1.0$: As inhibition ability improves, the effect of proactive interference decreases.

Received November 7, 2005
Revision received January 13, 2006
Accepted January 13, 2006 ■