Human Judgment under Sample Space Ignorance

Michael Smithson
Division of Psychology
Australian National University
Michael.Smithson@anu.edu.au

Thomas Bartos
Division of Psychology
Australian National University

Kazuhisa Takemura
Institute of Policy and Planning Sciences,
University of Tsukuba
takemura@shako.sk.tsukuba.ac.jp

Abstract

This paper surveys results of a research program investigating human judgments of imprecise probabilities under sample space ignorance. The framework used for comparisons with human judgments is primarily due to Walley [9, 10]. The five studies reported here investigate four of Walley's prescriptions for judgment under sample space ignorance, as well as assessing the impact of the number of observations and types of events on subjective lower and upper probability estimates. The paper concludes with a synopsis of future directions for empirical research on subjective imprecise probability judgments.

Keywords. Human judgment, uncertainty, ambiguity, vagueness, probability.

1 Introduction

As Hogarth and Kunreuther [3] observed, most studies and theories of decision making under uncertainty investigate only very restricted forms of uncertainty. Although there is a growing body of research on imprecise probabilities (cf. Camerer & Weber [1]), all such research assumes that the sample space has been partitioned uniquely. In the extreme case where there is no information on how to partition the sample space, we will use the phrase "sample space ignorance." Sample space ignorance implies that the decision maker has no prior information about the nature of the possible outcomes on relevant dimensions. The main focus of this paper is on how people make subjective probability judgments under sample space ignorance.

The primary stimulus for the research reported here has been the framework developed by Walley [9, 10]. He proposes four principles for judgment under sample space ignorance:

1. <u>Vacuous Prior Probabilities</u> [9]. Under complete ignorance, the upper probability of any conceivable event should be 1 and the lower probability should be 0.
2. <u>Embedding Principle</u> [9]. The initial (prior) upper and lower probabilities assigned to an event A should not depend on the sample space in which A is embedded.
3. <u>Symmetry Principle</u> [10]. In the absence of any prior information, the same prior upper and lower probabilities should be assigned to all elements of the sample space.
4. <u>Representation Invariance Principle</u> [10]. The posterior upper and lower probabilities assigned to an observable event A should not depend on the sample space in which A and the previous observations are embedded.

Walley [10] provides an imprecise Dirichlet model (IDM) for discrete events. Suppose we have no idea what events could occur, but we are interested in event A. Then given N trials and n occurrences of A, Walley proposes upper and lower probabilities:
$P^+(A|n) = (n+s)/(N+s)$, and
$P_-(A|n) = n/(N+s)$,
where $s > 0$ is a subjective parameter, indicating how cautious the decision maker is. Where possible, we will just use P_- for the lower probability and P^+ for the upper probability. Regardless of the value given to s, all four principles above are satisfied. Also:

- When $N = 0$, $P_- = 0$ and $P^+ = 1$.
- When $n = 0$, $P_- = 0$ and $P^+ = s/(N+s)$. The same holds for the union of two heretofore unobserved events, A and B, or indeed "any new event".
- When $n = N$, $P_- = N/(N+s)$ and $P^+ = 1$.

The five studies reported in this paper investigate various aspects of Walley's prescriptions, focusing initially on the four principles outlined above and comparing people's subjective estimates with the IDM and related imprecise probability formalisms. Before proceeding to the studies themselves, we briefly outline some issues regarding the elicitation and/or evaluation of subjective imprecise probabilities. These issues are largely unresolved, and so the merits of the elicitation methods adopted in our studies are debatable. Nonetheless, these remarks and our findings may encourage further debate and experimentation on these matters.

First, many previous studies of imprecise probability judgments have used comparative evaluations rather than direct estimations. Moreover, most such comparisons have asked subjects to evaluate probability intervals against a pointwise alternative (see [1] for a review). Only in the last decade or so have comparisons between intervals become a major focus (cf. Curley and Yates [2]). Almost all of these studies have used these comparisons for investigating "ambiguity aversion", i.e., whether people prefer more precise estimates to supposedly equivalent but less precise estimates.

More recently, Yaniv and Foster [11] have asked subjects to choose between two interval estimates (not of probabilities but other quantities such as dates or distances) when given the true value of the parameter being estimated. Their object was an accuracy-informativeness tradeoff, but there is no reason that this could not be extended to an accuracy-decisiveness tradeoff even when the true probability of an event is unknown.

It is not clear from the literature whether estimation tasks would yield patterns consistent with comparisons. Under what conditions would people display ambiguity aversion, for instance, when they have produced interval probability estimates themselves rather than being presented with them? More importantly, reliance on comparisons has impoverished our understanding of issues that cannot be studied via comparisons, such as people's preferences for representing uncertainty or ignorance, and how they respond to their own representations.

Regardless of whether comparisons or estimates are involved, several fundamental issues have not been clarified in the empirical literature on imprecise probabilities, of which two deserve mention here. One is stochastic dominance. When should one interval probability be regarded as dominating another? Most of the comparative evaluation studies have assumed that the dominating interval is the one with the higher midpoint. Thus, most studies of this kind have presented subjects with choices between two intervals that share a midpoint but have different widths, under the belief that those intervals are 'equivalent' gambles. This is at odds with many imprecise probability models, and is contraindicated by evidence in one of the studies reported here (Study 5, although there is insufficient space to elaborate on this aspect of the study in this paper).

The second issue concerns the reference-class of events being estimated. Reichenbach [4] was among the first to systematically eludicate the problem, and Kyburg's comments on Walley's [11] paper raised it anew. For example, we may ask subjects to estimate the probability of event A occurring in the next draw from a population of unknown composition, or we may ask them to estimate the proportion of that population consisting of A's. In Walley's IDM, the lower and upper expected values for θ_A, the probability of getting event A under random sampling, are identical to $P_-(A|n)$ and $P^+(A|n)$ given n occurrences of A in N draws ([10]: 17). What is not known is whether people in fact give different estimates in response to these two tasks. We report some tentative but pertinent findings on this issue in this paper.

Finally, there is the problem of how best to elicit lower and upper probability estimates. Little is known about how people intuitively produce lower and upper estimates, or what the gap between such estimates means to them. There is a small, scattered empirical literature involving subjective lower, upper, and 'best' estimates of various quantities, but it is not oriented towards subjective probability and lacks a psychometric foundation. In the studies reported here, we have simply used direct estimates rather than undertaking systematic psychometric scaling approaches.

2 Sample Space Ignorance Aversion

To start with, we wished to ascertain whether people prefer partial to complete sample-space ignorance, and whether sample space ignorance has behavioral consequences for decision makers. This issue does not pertain directly to any of Walley's four principles, but instead to the question of whether people behave as if they distinguish between a situation in which there is no sample-space information and one in which at least some such information has been provided. Accordingly, Study 1 presented 46 volunteer undergraduate participants with two tasks (with the order counterbalanced), as described below.

Task 1

<u>Urn A</u> contains 1000 marbles whose colours may be either Red or various other colours, <u>but you don't know how many are Red or what any of the other colours are</u>.
You will receive $100 if the next marble drawn is <u>Red</u>, otherwise you pay $5.

<u>Urn B</u> contains 1000 marbles <u>whose colours are unknown</u>.
You will receive $100 if the next marble drawn is <u>Red</u>, otherwise you pay $5.

<u>Which urn would you choose for your bet? Why?</u>

Task 2

Urn A contains 1000 marbles whose colours may be either Green or various other colours, but you don't know how many are Green or what any of the other colours are.
You will receive $100 if the next marble drawn is Black, otherwise you pay $5.

Urn B contains 1000 marbles whose colours are unknown.
You will receive $100 if the next marble drawn is Black, otherwise you pay $5.

Which urn would you choose for your bet? Why?

The hypothesis was that people will prefer Urn A in task 1 (i.e., partial ignorance) but Urn B in task 2 (complete ignorance), thereby switching their preference from partial to complete ignorance. Table 1 shows that 71.7% of the participants' choices were as hypothesized. In Task 1 most of them chose the urn with partial information about the sample space, whereas in Task 2 most chose the urn with no information. This demonstrates that sample space ignorance has behavioral consequences. However, there is also a mild tendency for people to favor partial ignorance over complete ignorance.

		Task 1 Urn A	Task 1 Urn B	Total
Task 2	Urn A	9 19.6%	2 4.3%	11 23.9%
	Urn B	33 71.7%	2 4.3%	35 76.1%
	Total	42 91.3%	4 8.7%	46

Table 1: Choice Patterns for Tasks 1 and 2

3 Vacuous Priors and Symmetry Principles.

Studies 2-5 investigated when people are willing to assign vacuous priors under complete ignorance, and two other related rules.

1. The Vacuous Priors and Symmetry Principles were investigated by comparing lower and upper probability assignments for simple versus compound events.
2. Given a sequence of N events, people's upper and lower assignments for unobserved events (n = 0) were examined to see whether $P_- = 0$.
3. Given a sequence of N events, people's upper and lower assignments for observed events ($0 < n < N$) were examined to see whether $P^+ < 1$ and $P_- > 0$.

3.1 Prior Probabilities

Two tasks in Study 2 tested the Vacuous Priors Principle and a strong version of the Symmetry Principle by eliciting prior probability estimates from participants. The strong version of the Symmetry Principle holds in Walley's IDM, and stipulates that in the absence of prior information the same prior upper and lower probabilities should be assigned to all conceivable events. In the two tasks concerned, participants were asked to consider a bag of marbles of unknown colors, and to estimate the lower and upper probability of drawing a red marble on the first draw (in one task) and the lower and upper probability of drawing a blue or yellow marble (in the other task). The order of the tasks was counterbalanced.

Table 2 shows that a majority of participants in both tasks assigned vacuous probabilities (59.6% for the Red task and 66.3% for the Blue-Yellow task). No one assigned a lower probability greater than .5 nor an upper probability less than .5. Interestingly, in both tasks any participant who gave $P_- = 0$ was almost certain to give $P^+ = 1$ (59 out of 60 in the Red task and 63 out of 64 in the Blue-Yellow task), but the converse was not as likely (59 out of 73 in the Red task and 63 out of 71 in the Blue-Yellow task). A plausible explanation is that mere mention of an event makes some people reluctant to assign a lower prior probability of 0 to it.

Red		P^+ .5	P^+ >.5 & <1	P^+ 1	Total
P_-	.5	6	2	2	10
	>0 & <.5	3	14	12	29
	0	1	0	59	60
	Total	10	16	73	99

Blue or Yellow		P^+ .5	P^+ >.5 & <1	P^+ 1	Total
P_-	.5	3	2	2	7
	>0 & <.5	7	11	6	24
	0	1	0	63	64
	Total	11	13	71	95

Table 2: Prior Probabilities for Two Tasks

The most direct test of whether the strong version of the Symmetry Principle is violated is to determine whether each participant gave the same lower and upper probabilities for both tasks, which was the case in 58 out of 95 participants (61%) whose responses on both tasks were valid. In fact, 65% of the participants gave the same lower probabilities whereas 78% of them gave the same upper probabilities.

Finally, adherence to the Vacuous Priors Principle and the strong Symmetry Principle were closely linked. Of the 62 participants who gave identical lower probabilities, 55 (89%) of them assigned 0; and of the 74 who gave identical upper probabilities, 65 (88%) of them assigned 1. Both of these percentages are substantially higher than the respective marginal percentages of people assigning a lower probability of 0 or an upper probability of 1.

3.2 Posterior Probabilities

In another part of Study 2, participants' subjective estimates were elicited after they had been provided with some observations. Participants were asked to consider a bag of marbles of unknown colors, and were shown eight marbles that had been drawn from the bag. Two were red. They were asked to provide a lower and an upper probability of drawing a red marble from the bag on the next turn, and lower and upper probabilities of drawing an orange marble on the next turn (none of the eight marbles were orange). They were then shown eight more marbles. Again, two were red and none were orange. They were asked for their estimates once again.

First, we examine whether participants were willing to assign a lower probability of 0 for a heretofore unobserved event (i.e., an orange marble). For N = 8 this assignment was made by 33 (71.7%) out of 46 participants, and for N = 16 it was made by 36 (78.3%). Turning now to whether participants observed the $P_- > 0$ rule for an observed event (i.e., a red marble), in Study 2 the rate of conformity with this rule was similar to the foregoing. For N = 8 an assignment of $P_- > 0$ was made by 30 (65.2%) participants, and for N = 16 this increased to 34 (73.9%).

Study 3 (by Takemura) partly replicated Study 2 (by Smithson), with a Japanese sample of undergraduate students (Study 2 participants were Australian undergraduates). Participants were given identical marble-draws (N = 8 and N = 16) to those in Study 2 and they were likewise asked to provide a lower and an upper probability of drawing a red marble on each occasion. However, they were also asked for lower and upper probabilities of drawing a marble of any new color (i.e., a color that has not yet been observed, rather than an orange marble). A greater percentage of this sample observed the $P_- > 0$ rule for an observed event than the Australian sample on both occasions. In fact, 116 out of 122 (95.2%) did so for N = 8, and 96.0% did so for N = 16.

However, only a minority of participants were willing to assign a lower probability of 0 for any new color. This assignment was made by 20 (16.4%) for N = 8 and 26 (21.3%) for N = 16. A substantial number of participants chose nonzero lower probability estimates of 0.1 or less (27, or 22.1% for N = 8, and 42, or 34.3% for N = 16). These percentages were greater than those found in the Australian sample. One obvious possible explanation for these findings is that the Japanese participants simply preferred to use lower probabilities of greater than 0 to a greater extent than Australians. Another explanation, however, is that people are more reluctant to assign 0 as a lower probability for "any new color" than they are for a specifically named color (such as "orange").

This latter possibility was explored in Study 4 (by Smithson). This study had a similar pair of marble-draw scenarios to Studies 2 and 3, but this time N = 4 and then N = 16. However, there is also an important methodological difference between this study and the others in this paper, namely that participants were asked to provide lower and upper estimates of the percentage of marbles in the bag that were, say, red. We recognize that this is not the same as asking for lower and upper probabilities of getting a red marble on the next draw, and pertains to people's hypotheses about possible values of the population percentage of red marbles instead. We were interested in comparing how subjects responded to the two different estimation tasks, for reasons that space precludes from elucidating here. Although we realize that an airtight empirical investigation requires an additional study which uses lower-upper probability elicitation rather than percentage estimation, we nevertheless claim that the results of Study 4 are suggestive and usable for our purposes here.

Participants were asked to give estimates for both orange and any new color. The guiding hypothesis in this study, as suggested by results from Studies 2 and 3, was that conformity with the $P_- = 0$ rule for "any new color" may be lower than for a named (specific) color. Table 3 shows that this hypothesis was supported, with lower rates of conformity for the any new color task than for the orange task. The difference is close to 10% for both N = 4 and N = 16, as well as jointly.

Orange N = 16

	P_-	0	>0	Total
N = 4	0	45	2	47
		43.7%	1.9%	45.6%
	>0	18	38	56
		17.5%	36.9%	54.4%
	Total	63	40	103
		61.2%	38.8%	

Any new color N = 16

	P_-	0	>0	Total
N = 4	0	33	3	36
		32.0%	2.9%	35.0%
	>0	20	47	67
		19.4%	45.6%	65.0%
	Total	53	50	103
		51.5%	48.5%	

Table 3: Assignment Patterns for N = 4 vs. N = 16

Finally, Study 5 (by Bartos) used a game-like scenario to elicit estimates from a sample of Australian undergraduate participants over time. Participants were 'fishing' for a particular kind of microorganism from a pool. They were asked to estimate the lower and upper probability of finding the target organism on the next trial.

The P_- = 0 rule for a heretofore unobserved event (the target organism) was adhered to in about half of the trials as was the P^+ = 1 rule for an always observed event; and the $P_- > 0$ and $P^+ < 1$ rules for an (intermittently) observed event were adhered to in a very large majority of trials.

	Event Occurrence	Total Trials	Vacuous Estimate	Non-vacuous
P_-	Never seen	254	123 (48%)	131 (52%)
	Sometimes seen	847	103 (12%)	744 (88%)
P^+	Always seen*	38	19 (50%)	19 (50%)
	Sometimes seen	1063	14 (1%)	1049 (99%)

* Usually this meant that the Target had appeared on the first trial and the participant on the second trial therefore gave an Upper Estimate for an event which had never failed to occur.

Table 4: Lower and Upper Estimates in Study 5.

3.3 Discussion

The findings in Study 2 indicate that a sizable percentage of people seem to conform with the Vacuous Priors Principle and the strong version of the Symmetry Principle, and if a person gives 0 as a lower prior probability then they are very likely to give 1 as an upper bound. Moreover, those conforming with the Symmetry Principle are quite likely to also adhere to the Vacuous Priors Principle. Based as they are on a single study involving only one task, these are tentative conclusions entailing replication and extensions in future research.

Studies 2-5 found that the P_- = 0 rule for an unobserved event is adhered to by a substantial percentage of participants, but not always a majority of them. Many prefer to give a value of P_- that is slightly above 0. However, compliance with the $P_- > 0$ rule for observed events is quite high. Compliance with 0 as a lower probability for an unobserved event has a lower rate for smaller N and for an unspecified (rather than a named) event. Moreover, there is a moderately strong positive association between nominating a lower bound of 0 for larger N and smaller N, and likewise across tasks.

Finally, although Study 4 involved eliciting lower and upper estimates of population percentages rather than lower and upper probabilities of an event occurring on the next trial, the pertinent aspects of the results did not differ much from Study 2 which sampled subjects from the same population.

4 Representation Invariance Principle

Studies 2-4 investigated one test of the Representation Invariance Principle by randomly assigning participants to a homogeneous sample of marbles (blue and red only) or to a heterogeneous sample (greater variety of colors). Subjects did not know beforehand what colors the marbles would have, and our hypothesis was that a heterogeneous sample might make them more inclined to assign a greater probability of a heretofore unobserved color than a sample with only two colors occurring. The primary psychological interest in this manipulation stems from the possibility that prior information could influence subjective probabilities of novel phenomena. However, in all three studies we found no differences in subjects' lower or upper probability assignments for either observed or unobserved events.

On the other hand, in Study 4 participants' upper and lower percentage estimates for a specific unobserved event (<u>orange</u> colored marble) versus an unspecified event (<u>any new color</u>) suggested that the Representation Invariance Principle might be violated by at least some of them. Both the mean lower and upper probabilities for orange were significantly lower than those for any new color. One interpretation of this finding is that any new color is regarded as a more inclusive event than orange and therefore more likely to occur.

A repeated-measures ANOVA on the <u>lower probabilities</u> gave main effects for N and the type of event, with a small interaction effect. The main-effect $F(1,102)$ = 26.004 for the type of event, with $p < .0005$. The same kind of result was obtained for <u>upper probabilities</u> (but without the interaction effect). The main-effect $F(1,102)$ = 10.332 for the type of event, with $p = .002$. Both of these are sizeable effects, with $\eta^2 = .203$ for the first and $\eta^2 = .092$ for the second.

Clearly we have only just begun to explore the Representation Invariance Principle, and these tests are merely two among a myriad of possibilities. Nonetheless, those tests occupy what we argue is one of the most interesting subcategories, in which the posterior probabilities refer to novel (or previously unobserved) events. This type of estimation is important because it pertains to human abilities to cope with the novel or unexpected. In the next section, we focus on a comparison between the precision of subjective estimates for these probabilities and probabilities of previously observed events.

5 Unobserved Event Probabilities

Studies 2 - 5 provided opportunities to compare imprecision for observed events (<u>red</u> marble) and unobserved events. A motivation for doing this was to ascertain whether people tend to underestimate the likelihood of an unobserved event, compared to their estimates for observed ones. This hypothesis was suggested by a related phenomenon called the "Catch-All Underestimation Bias" (CAUB, cf. Russo & Kozlow [5]), which is a tendency for people to provide lower estimates for grouped alternatives than the sum of the estimates they provide for each ungrouped alternative.

The CAUB is one of the few commentaries provided to date by psychological research on how people estimate the likelihood of novel events. One of the primary reasons for the paucity of research in this area, of course, is the lack of normative consensus on this topic. After all, what are "reasonable" odds that the sun won't rise tomorrow morning?

Walley's IDM , on the other hand, provides us with one important guideline which is that one's precision in estimating lower and upper probabilities should not be affected by the sample space in which the previous observations are embedded. Any systematic inclination for people to be less or more precise in estimating probabilities of unobserved events than in observed events will have interesting implications for risk assessment and fault-tree analysis, among other topics. Here, we investigate whether any such bias occurs and whether it is linked with the number of observations (N).

5.1 Findings from Four Studies

Study 2 (Smithson) compared imprecision in assignments for observed events (<u>red</u> marble) with imprecision for unobserved events (<u>orange</u>). The effects of N=8 vs. N=16 and type of event on imprecision were tested in a 2x2 repeated measures ANOVA. The results indicated significant main effects for N ($F(1,45) = 4.14$, $p = .048$, $\eta^2 = .084$) and for event type ($F(1,45) = 8.08$, $p = .007$, $\eta^2 = .152$) but no interaction effect ($F(1,45) = 2.670$, $p = .109$).

Study 3 (Takemura) compared imprecision in assignments for observed events (<u>red</u> marble) with imprecision for unobserved events (<u>any new color</u>). The effects of N=8 vs. N=16 and type of event on imprecision were tested in a 2x2 repeated measures ANOVA. The results indicated significant main effects for N ($F(1,118) = 19.521$, $p < .0005$, $\eta^2 = .142$) and for event type ($F(1, 118) = 19.723$, $p < .0005$, $\eta^2 = .143$) and an interaction effect ($F(1, 118) = 4.198$, $p = .028$, $\eta^2 = .040$). It should be noted that the effect-size for the interaction effect in this study is similar to that obtained in Study 2 ($\eta^2 = .056$) even though it was not statistically significant, perhaps due to that study's smaller sample size.

Study 4 (Smithson) also compared imprecision in assignments for observed events (<u>red</u> marble) with imprecision for unobserved events (<u>orange</u> or <u>any new</u>). The effects of N and type of event on imprecision were tested in a 2x3 repeated measures ANOVA. The results indicated significant main effects for N ($F(1,102) = 78.160$, $p < .0005$, $\eta^2 = .434$) and for event type ($F(2,102) = 3.527$, $p = .033$, $\eta^2 = .065$) but no interaction effect ($F(2,101) = 2.137$, $p = .123$).
Post-hoc contrasts indicate that imprecision differs between red and the other two kinds of events (red/orange contrast $F(1,102) = 5.423$, $p = .022$ and red/any new contrast $F(1,102) = 6.726$, $p = .011$). Both unobserved events' upper and lower probabilities are closer together (more precise) than those for the observed event.

Study 5 (Bartos), as mentioned earlier, used a game-like scenario to elicit estimates from subjects over time. For each subject, N ranged from 0 to whenever their game ended. Here, we compare subjects' imprecision in estimating the lower and upper probability of finding the target organism on the next turn when the organism had not yet been observed versus when it had already been observed. The data analyzed here are restricted to $N \leq 4$, since cases where the subject had not yet seen the target organism became rare for larger N. The results indicated significant main effects for N ($F(3,510) = 4.556$, $p = .004$, $\eta^2 = .026$) and for event type ($F(1,510) = 7.334$, $p = .007$, $\eta^2 = .014$) and no interaction effect ($F(1,510) = 0.169$, $p = .918$).

5.2 Summary and Discussion

Studies 2-4 give a consensual picture of a tendency for people to be less imprecise (and therefore less cautious) about probability estimates for unobserved events than for observed events. The effect-sizes in all studies are fairly similar, whereas for Study 5 the effect-sizes are rather small. Table 5 displays the means and standard errors for the imprecision levels found for each task in the three studies.

	Event	Mean	s.e.
Study 2	Red (observed)	0.558	.048
	Orange (unobserved)	0.437	.052
Study 3	Red (observed)	0.381	.013
	Any New Color	0.324	.016
Study 4	Red (observed)	0.292	.020
	Orange (unobserved)	0.254	.024
	Any New Color	0.251	.023
Study 5	Target observed	0.406	.013
	Target not observed	0.371	.013

Table 5: Mean Imprecision Results in Studies 2-5

Again, although Study 4 elicited parameter estimates rather than next-event probabilities, the overall pattern of results is consistent with those for the other studies. The balance of evidence thus far does not indicate any interaction between the number of observations made and the tendency to be less imprecise for heretofore unobserved event probabilities.

6 Effect of N and Event Type on s-values

Since imprecision is $P^+ - P_- = s/(N+s)$, we may impute s-values to subjective lower and upper probability estimates if we know N, by rearranging this formula to yield
$s = N(P^+ - P_-)/(1 - P^+ - P_-)$.
Although imprecision was found to decrease with N in studies 2-5 as might be reasonably expected, this says nothing about the behavior of s as a function of N. In this section we focus on how large s-values are and whether they vary with N. Studies 2, 3, and 5 provide direct evidence concerning s, while the data from Study 4 may be used for suggestive comparisons.

6.1 Findings from Studies 2-4

In Study 2, 20 subjects had $P^+ - P_- = 1$ for at least one estimation task, so this analysis is based on the 25 subjects for whom s was a defined value. Repeated-measures ANOVA indicated that s did not increase with N, but there was an interaction effect such that s did increase for the observed event (red) and decreased slightly for the unobserved event (orange). That pattern may be seen in Table 6 below.
Study 3 contained much fewer ‘pathological’ values for s (only 11 out of 118 subjects had undefined s-values). For the remaining 107 subjects, ANOVA results indicated that the value of s increased with N. There was also an interaction effect such that s increased more rapidly for observed than for unobserved events. As can be seen in Table 6, three of the mean s-values for Study 3 are fairly similar to their counterparts in Study 2. The exception is for the unobserved event when N = 16.

	Event	N	Mean	s.e.
Study 2	Red (observed)	8	6.168	1.225
	Red (observed)	16	9.712	1.548
	Orange (unobserved)	8	4.682	1.427
	Orange (unobserved)	16	3.559	1.067
Study 3	Red (observed)	8	5.305	0.276
	Red (observed)	16	9.303	0.549
	Any New Color	8	4.702	0.384
	Any New Color	16	7.250	0.600

Table 6: Mean s Values in Studies 2 and 3

Study 4, like Study 2, found no main effect for N but strong evidence of an interaction effect. A repeated measures ANOVA yielded a main effect for event type (Multivariate $F(2,81) = 14.227$, $p < .0005$, $\eta^2 = .260$) and for the interaction ($F(2,81) = 3.840$, $p < .025$, $\eta^2 = .087$), but no main effect for N ($F(1,82) = 1.340$, $p = .250$). Post-hoc contrasts reveal that the difference is really due to the higher mean for Red when N = 4 and when N = 16. Moreover, the s-value for red increases for higher N. The results are summarized in Table 7.

Event	N	Mean	s.e.
Red (observed)	4	2.288	0.286
Red (observed)	16	3.238	0.309
Orange (unobserved)	4	1.737	0.293
Orange (unobserved)	16	1.603	0.229
Any New Color	4	1.803	0.297
Any New Color	16	1.872	0.270

Table 7: Mean s Values in Study 4

6.2 Findings from Study 5

For each subject in this study, N ranged from 0 to whenever their game ended, thereby providing a range of N’s for testing its effect on s. Figure 1 shows how log(s) increased with N for those values of N with sufficiently many cases to provide reliable estimates. Log(s) is used here mainly to decrease the influence of outliers on the analysis. A simple ANOVA with log(s) as the dependent variable yielded $F(12,1127) = 23.416$ with $p < .0005$. The effect-size was $\eta^2 = .200$, so the effect is sizeable in terms of variance explained. The graph shows a near-monotonic increasing trend, with some evidence of leveling-off around N = 10.

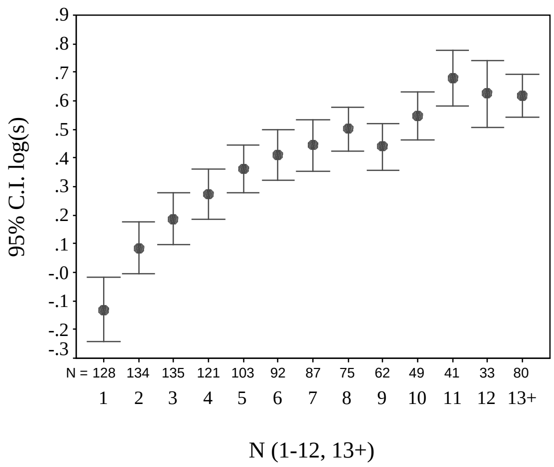

Figure 1: Log(s) Values increasing in N

Now we turn to the comparison of observed with unobserved events. As in the imprecision analysis, these data are restricted to $N \leq 4$, since cases where the subject had not yet seen the target organism became rare for larger N. As in the other studies, s-values are lower for unobserved events (the CAUB-like bias). Figure 2 displays the findings.

Unlike studies 2 and 4, s increased with N for both observed and unobserved events. Neither main effect is large in terms of variance, but they are significant ($F(3,510) = 8.581$ with $p < .0005$ for the effect due to N, and $F(1,510) = 6.808$ with $p = .009$ for the (un)observed event effect; η^2 for the former is .048 and for the latter is .013). Contrary to the findings in studies 2-4, there was no interaction effect.

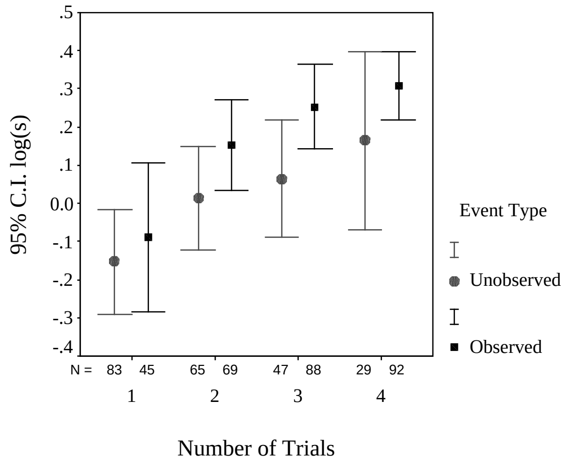

Figure 2: Observed vs. Unobserved Events

6.3 Summary and Discussion

In all studies thus far, the typical s-value for a subject is considerably higher than Walley's recommendation of s = 1 (or at most, 2). The only possible exception is in the "improper" estimation tasks in Study 4, and even then only for unobserved events! One possible implication of this finding is that people are less decisive than normative frameworks would advise them to be, in the face of sample space ignorance or uncertainty.

All four studies found that for events that have been observed, s increased with N even though imprecision declined with N. For unobserved events, the findings were less clear. Studies 3 and 5 found an increase in s with N, whereas studies 2 and 4 did not. Studies 2-4 found evidence that the tendency for s to increase with N is larger for observed than unobserved events, but Study 5 did not reproduce that finding, at least, not for values of N from 1 to 4. It would seem that a study is required which, like Study 5, extends subjects' estimates over a larger range of N than the rather limited ones used thus far, and includes estimates of both specific a nonspecific unobserved event probabilities.

7 Future Directions for Research

Our studies thus far indicate that while subjective imprecise probability judgments agree in some aspects with Walleyan prescriptions (e.g., the Vacuous Priors and Symmetry Principles), they depart from them in several important respects (e.g., violations of the Representative Invariance Principle, a tendency to give less imprecise estimates for unobserved than for observed events, and a tendency for s to increase with N). These studies merely scratch the surface, and there is much to be done. The following list is not exhaustive, but comprises topics immediately suggested by Walley's framework and the investigations reported in this paper.

<u>Elicitation methods</u>:

As indicated at the outset of this paper, there are several outstanding issues concerning the elicitation of subjective lower and upper probabilities. These issues should be investigated with a goal of formulating an integrative framework on elicitation methods.

<u>Reference class</u>:

Comparisons between Studies 4 and 2 suggest that when people are asked for estimates of an unknown proportion in a population they tend to provide narrower intervals than when asked for estimates of an unknown probability regarding a draw from that population. The issue of reference class deserves further exploration and clarification, preferably via within-subjects experimental designs rather than the between-samples ad hoc comparisons used here.

<u>Similarity of unobserved to observed events</u>:

The hypothesis that the upper (and perhaps lower) probability of an unobserved event would be judged to be greater if the event is similar to one that has been observed could easily be tested. This hypothesis is suggested by Tversky & Kahneman's [7] work on the representativeness heuristic, which indicates that the subjective probability of an event often is based on its similarity to events that have already occurred.

<u>Calibration and coherence</u>:

Coherence in the sense of [9] has not yet been tested directly. However, a related topic, calibration, could be addressed by the data from studies 2-5. Space limitations preclude elaborating on this point here. Calibration assessment is based on the following relationship that holds in any e-contaminated Walley's [10] framework: $P_-/(1-P^+) = n/(N-n)$. Similar odds-like relationships can be shown for certain related imprecise probability schemes. Since these relationships are independent of s, they provide a calibration benchmark independent of cautiousness. That is, we may assess how close the odds-like expression is to n/(N-n) and whether there is any systematic bias towards over- or under-estimation.

<u>Modeling lower and upper probability judgments</u>:

Study 5 provided data suitable for modeling lower and upper probability judgments, and these results, along with those for calibration, will be reported elsewhere. There are two related enterprises involved: Modeling lower and/or upper subjective probabilities as a function of n/N; and using lower and upper subjective probabilities to model subjective 'best estimate' probabilities. This is an area where much more work is needed, preferably integrating mathematical imprecise probability frameworks with psychological work on weighted-average and anchoring-and-adjustment models.

<u>Subadditivity</u>:

Pointwise subjective probability estimates often exhibit subadditivity (cf. Tversky & Koehler [8]). Are subjective lower probability estimates also subadditive (which would contradict Walley's and others' frameworks)?

Accuracy-informativeness tradeoff:
Yaniv and Foster [11] have tested models of how people assess tradeoffs between accurate (but imprecise) intervals and inaccurate (but precise) ones. This line of research could be pursued in studying how people decide on the level of imprecision when accuracy is unknown, or how they trade off imprecision against decisiveness.

The 'dilation' problem:
Under some conditions in various imprecise probability frameworks,
$P^+(A|B) > P^+(A) \geq P_-(A) > P_-(A|B)$.
Seidenfeld and Wasserman [6] coined the term "dilation" for this issue, and also raised it in the Discussion in [10]. A related issue is that conditionalization can make imprecise probabilities more imprecise, so that
$P^+(A|B) - P_-(A|B) > P^+(A) - P_-(A)$.
Do people's judgements adhere to this rule when it is indicated by the IDM, for instance? Or are they always less cautious about imprecision when given more information? When additional information would be regarded by people as nondiagnostic and increase imprecision, under what conditions do they prefer not to obtain that information?

The 'event horizon' problem:
In Walley's framework, predicting further ahead than one turn (e.g., what's the probability of getting at most y A-events in the next Y trials) widens imprecision to an asymptotic limit that is well below 1. Do people exhibit a limit such as this for an arbitrarily long prediction horizon, or do they eventually broaden out to vacuity?

The 'Monkey Trap' problem:
Under sample space ignorance, when do people prefer to stick with what they have or gamble on something better coming along? Note that this problem has a special formulation under sample space ignorance.

References

[1] C. Camerer and M. Weber. Recent developments in modeling preferences: uncertainty and ambiguity. *Journal of Risk and Uncertainty*, 7: 215-235, 1992.

[2] S.P. Curley and J.F. Yates. The center and range of the probability interval as factors affecting ambiguity preferences. *Organizational Behavior and Human Decision Processes*, 36: 273-287, 1985.

[3] R.M. Hogarth and H. Kunreuther. Decision making under ignorance: Arguing with yourself. *Journal of Risk and Uncertainty*, 8: 1-37, 1995.

[4] H. Reichenbach. *Theory of Probability*. Berkeley: University of California Press, 1949.

[5] J.E. Russo and K. Kozlow. Where is the fault in fault trees? *Journal of Experimental Psychology: Human Perception and Performance*. 20: 17-32, 1994.

[6] T. Seidenfeld and L. Wasserman. Dilation for sets of probabilities. *Ann. Statist.*, 21: 1139-1154, 1993.

[7] A. Tversky and D. Kahneman. Judgments of and by representativeness. In D. Kahneman, P. Slovic, and A. Tversky (eds.) *Judgment Under Uncertainty: Heuristics and Biases*. New York: Cambridge University Press, 1982.

[8] A. Tversky and D.J. Koehler.. Support theory: A nonextensional representation of subjective probability. *Psychological Review*, 101: 547-567, 1994.

[9] P. Walley. *Statistical Reasoning with Imprecise Probabilities*. London: Chapman Hall, 1991.

[10] P. Walley. Inferences from multinomial data: Learning about a bag of marbles. (with discussion) *Journal of the Royal Statistical Society, Series B*, 58: 3-57, 1996.

[11] I. Yaniv and D.P. Foster. Graininess of judgment under uncertainty: an accuracy-informativeness tradeoff. *Journal of Experimental Psychology: General*, 124: 424-432, 1995.