

# Participant recruitment methods and statistical reasoning performance

---

---

**Gary L. Brase**

*University of Missouri–Columbia, Columbia, MI, USA*

**Laurence Fiddick**

*ESRC Centre for Economic Learning and Social Evolution,  
and James Cook University, Townsville, Queensland, Australia*

**Clare Harries**

*University College London, London, UK*

Optimal Bayesian reasoning performance has reportedly been elusive, and a variety of explanations have been suggested for this situation. In a series of experiments, it is demonstrated that these difficulties with replication can be accounted for by differences in participant-sampling methodologies. Specifically, the best performances are obtained with students from top-tier, national universities who were paid for their participation. Performance drops significantly as these conditions are altered regarding inducements (e.g., using unpaid participants) or participant source (e.g., using participants from a second-tier, regional university). Honours-programme undergraduates do better than regular undergraduates within the same university, paid participation creates superior performance, and top-tier university students do better than students from lower ranked universities. Pictorial representations (supplementing problem text) usually have a slight facilitative effect across these participant manipulations. These results indicate that studies should take account of these methodological details and focus more on relative levels of performance rather than absolute performance.

One of the best known Bayesian reasoning tasks is the *medical diagnosis problem*, which originally was in the following form:

If a test to detect a disease whose prevalence is 1/1,000 has a false positive rate of 5%, what is the chance that a person found to have a positive result actually has the

disease, assuming that you know nothing about the person's symptoms or signs? \_\_\_%

This original version was correctly answered by only 18% of medical students and doctors, which seemed to be uniformly agreed upon as poor performance (Casscells, Schoenberger, & Graboys, 1978); the correct posterior probability for this

---

Correspondence should be addressed to Gary L. Brase, Department of Psychological Sciences, 210 McAlester Hall, University of Missouri–Columbia, Columbia, Missouri, 65211, USA. Email: braseg@missouri.edu

The authors would like to thank the University of Sunderland and the ESRC for funding that supported Experiment 4, and Kimberly Sapp and Rebecca Miller for assistance in collecting some portions of the data. We also thank Sandra Brase and Mike Oaksford for advice and support regarding this research.

problem is .02 (.019627 to be precise), which can be derived from Bayes theorem:  $(.001*1)/[(.001*1) + (.999*.05)]$ . Subsequent rewordings and clarifications of this same problem, however, have elicited correct answers from up to 92% of university students (Cosmides & Tooby, 1996), which is similarly construed by almost all researchers as good performance. In between these two extremes lie many studies that have found intermediate levels of performance and, based on these results, have reached various interpretations of either good or poor statistical competence and rationality.

Early work was characterized by findings that human statistical judgements, including Bayesian reasoning tasks such as the medical diagnosis task, were not accurate (as many assumed) but instead quite poor. Against assumptions that people were generally competent and rational, Casscells et al. (1978, p. 1000) concluded of their participants: "Formal decision analysis was almost entirely unknown and even commonsense reasoning about the interpretation of laboratory data was uncommon." This result served to reinforce earlier pronouncements that "in his evaluation of evidence, man is apparently not a conservative Bayesian: He is not a Bayesian at all" (Kahneman & Tversky, 1972, p. 450).

With the fallibility of statistical judgement abilities established as the orthodoxy, the situation became the reverse of the initial assumption of statistical competence. People were assumed to generally lack crucial statistical insights and computational competencies, and findings of statistical competence became the counterintuitive result. For example, Cosmides and Tooby (1996) converted the numerical format of the medical diagnosis problem and found a large facilitative effect of using natural frequencies. They concluded that "It now appears that this conclusion was premature. Frequentist problems elicit bayesian reasoning" (p. 62). Similarly, Gigerenzer and Hoffrage (1995) found that "frequency formats made many participants' inferences strictly conform (in terms of outcome and process) to Bayes' theorem without any teaching or instruction" (p. 698).

Researchers in the limited human statistical competence tradition have assailed these findings

of enhanced performance in several ways, one of these being a repeated failure to obtain the "high-water" marks of performance found by Cosmides and Tooby (1996) and Gigerenzer and Hoffrage (1995). For instance, Macchi and Mosconi (1998) used Cosmides and Tooby's version of the medical diagnosis task but observed levels of performance that failed to approach those obtained previously:

This is possibly because of a difference in the degree of knowledge of statistics and Bayesian reasoning between the two groups: Our subjects were statistically naïve, whereas theirs were "physicians and fourth-year medical students" in a first experiment, and probably also in the experiment we are considering here (although this is not specified in their article). (p. 83)

Actually, the participants in Cosmides and Tooby (1996) were all Stanford University undergraduates (see p. 23, second paragraph), so this explanation is both factually wrong and dubious. Sloman, Over, Slovak, and Stibel (2003) noted, without further comment, a 21-percentage-point drop in performance between Cosmides and Tooby's (1996) results and their own on the same version of the medical diagnosis task (from 72% to "only 51%"; p. 300).

The Gigerenzer and Hoffrage (1995) results have been subject to similar findings and criticisms. Mellers and McGraw (1999) obtained levels of performance that were not even half those obtained by Gigerenzer and Hoffrage, and they attributed this difference to a combination of practice effects and/or cultural differences in basic mathematical training (i.e., Germany versus the US). In a reanalysis of their data, Gigerenzer and Hoffrage (1999, Footnote 1) did not find support for the practice effect explanation, but left the later explanation open. In keeping with this possibility, in psychology laboratory classes using the medical diagnosis task, the reasoning of those students who have a background in science prior to studying psychology is facilitated by frequency and picture presentations of the problem compared with other formats, but the reasoning of students with arts backgrounds prior to studying psychology is not (McClelland & Cheng, 2004). Another study (Sedlmeier & Gigerenzer, 2001) was repeated by Ruscio (2003),

but once again the performance of Gigerenzer's participants was higher than that in the subsequent research. Ruscio (2003, p. 327) proposed a number of methodological differences between the two studies to account for these differences, while noting that the overall pattern of results was very consistent across the two studies.

### What is "good" performance?

One theme running through all this previous research is that of relative comparisons. Given identical or isomorphic stimuli in different studies of human statistical competence, how do the study results compare to one another? This is an appropriate consideration and indeed is a fundamental aspect of scientific development, but the key issue is not just whether performance can be raised to some absolute criterion (e.g., as observed in study *X*), but whether performance can be raised relative to what might otherwise be the case under similar circumstances (e.g., comparing across control and experimental conditions within the same study). Specifically, there are reasons to suspect that the 92% correct performance in Cosmides and Tooby (1996) is at or close to an optimum result. Besides the fact that it is in striking distance of 100% correct, it involved two methodological factors that were potentially conducive to good performance:

1. The participants in Cosmides and Tooby (1996) were undergraduates at Stanford University, one of the top universities in the United States (US News & World Report, 2004).
2. The participants in Cosmides and Tooby (1996) were paid for their participation.

These participants, in general, would have had a history of practice and success in solving academic-style tasks (having been admitted to and attending Stanford), and they were placed in a setting (i.e., payments for participation) that encouraged more focused attention to the study task. Bayesian reasoning tasks are not only similar in format to many university assessments (e.g., paper and pencil format, with a single

correct answer), but are also typically given within an academic context. Furthermore, prior research has found that level of motivation is related to college academic achievement, as measured by overall grades, class performance, or specific exam results (Brownlow, Gilbert, & Reasinger, 1997; Lin, McKeachie, & Kim, 2001; Phillips, Abraham, & Bond, 2003; Talbot, 1990). Thus, to the extent that these methodological details selected or created greater motivation to perform well on the study tasks, it is reasonable to propose that changes in these methodological details could have affects on the obtained levels of performance.

The effect of paying research participants, as compared to using other forms of inducements (most commonly, connecting participation to some aspect of university coursework), has recently been a topic of specific interest. Hertwig and Ortmann (2001; see also Rydval & Ortmann, 2004) documented that psychological experiments, compared to research in economics, are much less likely to use monetary incentives. They proposed that this is one of the reasons for differing results across these two fields. Camerer and Hogarth (1999) similarly found that financial incentives can yield different patterns of performance, particularly in tasks that are responsive to increases in effort (e.g., to counteract when intrinsic motivations wane). Bayesian reasoning tasks—as noted in these studies—appear to fit this criterion: They are difficult enough that reaching the correct answer requires some diligence and sustained attention for most people. However, both Hertwig and Ortmann (2001) and Camerer and Hogarth (1999) reviewed past research, and therefore the effects that they proposed to explain were not controlled (because they were between-experiment effects). For instance, these studies did not provide data on within- and across-institutions comparisons of performance on the same task, with only the method of participant inducement manipulated. These studies also point out the contrasts between participation inducements (e.g., as part of a course requirement or course extra credit), participant monetary payments for mere participation ("show-up" fees), and

performance-based payments. The present study focuses on the first two of these methods (course-related inducements versus payment for participation) because these are the methods that have almost exclusively been used in prior psychological research on statistical reasoning and, hence, are key to understanding past variations in performance.

In summary, there are reasons to expect that at least some variation in Bayesian reasoning (and the medical diagnosis task in particular) is due to differences in the recruitment methods and participant pools used, rather than differences in the tasks themselves. To whatever extent results of these methodologically heterogeneous studies are used to attribute greater or lesser amounts of statistical competence, these attributions may be over- or underestimates of competence because of these unacknowledged effects.

## EXPERIMENT 1

Over the years, the original medical diagnosis problem has been criticized on a variety of grounds. Some of these criticisms have been generally accepted with little controversy. For example, Cosmides and Tooby (1996) added clarifications in the wording to ensure that participants understood that there are no false negative results to consider (“Every time the test is given to a person who has the disease, the test comes out positive”) and to clarify the meaning of false positives (“But sometimes the test also comes out positive when it is given to a person who is completely healthy”). Macchi and Mosconi (1998) pointed out that the original problem uses a limit-case base-rate (1 in a 1,000), and this can make different types of responses difficult to dissociate (e.g., taking the base rate into account versus ignoring the base rate will result in similar responses). This possibility was also pointed out by Gigerenzer and Hoffrage (1995) in their description of a “rare-event shortcut” that was used by some of their participants to produce very close approximations of more complex calculations. Although some real-life

situations are, in fact, limit-case problems (e.g., finding the perpetrator of a specific crime), and understanding any special consideration of such situations is important (e.g., use of a rare-event shortcut), for research purposes it is desirable to have clearly distinguishable responses. This can be achieved by simply changing to a higher base rate (e.g., 85 out of 1,000).

Changing the format in which the numerical information is presented in the medical diagnosis task can have dramatic effects (Cosmides & Tooby, 1996; Gigerenzer & Hoffrage, 1995). Although there is still controversy about the theoretical interpretation of this phenomenon, the use of naturally sampled frequencies, or “natural frequencies” (i.e., a series of nonnormalized frequency statements such as “ $x$  events out of  $y$  trials” that are categorically related within a natural sampling framework; Kleiter, 1994), facilitates the generation of correct Bayesian reasoning. For the present purposes, the *why* of this phenomenon is not as important as the fact that this format has been widely used and consistently produces enhanced levels of correct performance. Furthermore, the addition of pictorial representations also seems to facilitate Bayesian reasoning (Brase, Cosmides, & Tooby, 1998; Cosmides & Tooby, 1996). The effects of pictorial representations has not elicited as much controversy as the use of natural frequencies, but again the main concern at the moment is that it is a fairly consistent result.

Finally, Macchi and Mosconi (1998) proposed that, aside from the use of natural frequencies, the use of homogeneous data formats (that is, using a homogeneous reference class size) can facilitate Bayesian reasoning. Although we cannot use both natural frequencies and homogeneous reference classes within the same task (Macchi & Mosconi, 1998, specifically exclude tasks that employ natural frequencies as relevant to their proposal, because they involve different computational demands), we can compare the two types of task to discover which is more facilitative. All else being equal, heterogeneous data formats are more likely to make a task harder than are homogeneous data formats: If one piece of information is a relative frequency

(e.g., 5%), another is a single-event probability (e.g., .05), and the third is a total frequency (e.g., 1 out of 20), then a preliminary step has been added of converting the data all to one format (assuming that the computational machinery involved requires a consistent representational format). Macchi and Mosconi (1998), however, use the term homogeneous data to refer to frequency information that is expressed using a consistent, common reference class (e.g., a common denominator).

Experiment 1 was a head-on comparison of different versions of the medical diagnosis problem that served several purposes. First, there was an attempt to replicate the results of Macchi and Mosconi (1998) with a homogeneous data (i.e., common denominator) format. This was compared to a parallel-problem version that uses a natural-frequency presentation format (none of the problems used by Macchi and Mosconi, 1998, is actually a correctly formed natural frequency format) and a parallel problem that used natural frequencies plus a pictorial representation of the problem. These results will inform the interpretation of subsequent experiments, both in terms of which of these conditions would be most useful for further experiments and as a baseline level of performance for comparison.

## Method

### *Participants*

The participants were 120 undergraduates at a regional university in the south-eastern United States (96 females and 24 males). This institution is ranked as a second-tier, regional Master's-granting university (US News & World Report, 2004), and the participants consented to participate in exchange for extra credit in psychology classes. Their average age was 24.0 years.

### *Materials and procedure*

Each participant was given a single sheet with some brief instructions and one of three versions of a statistical reasoning problem (full text of problems are in the Appendix). An equal number of participants (40) received each of the three problem versions. The first version was a

replication of the final "homogeneous-data" problem constructed by Macchi and Mosconi (1998; their Version D). The second version was a revision of this problem that made it conform to a natural-frequency format. The third version was a further revision, again posed in natural frequencies, but also presenting the statistical information pictorially. There was no time limit for completing this study, and all participants finished within 10 minutes.

## Results

Table 1 shows the percentages of participants providing various answers to each problem version. The performance on the homogeneous data version generally replicates the findings of Macchi and Mosconi (1998) for both the number of correct answers and the number of incorrect answers of various types. A general Pearson chi-square test found that there were significant differences between the conditions, in the predicted direction:  $\chi^2(2) = 5.098$ ,  $p = .039$ , one-tailed. Pairwise difference of proportion tests found that the level of correct performance on the homogeneous-data version (worst performance) was significantly different from that in the natural-frequencies-plus-pictures condition (best performance; 7.5% vs. 25%,  $z = 2.12$ ,  $p = .02$ ,  $\eta = .24$ ), although the other conditions were not significantly different themselves (7.5% vs. 12.5%,  $z = 0.75$ ,  $p = .23$ ,  $\eta = .08$ , and 12.5% vs. 25%,  $z = 1.43$ ,  $p = .08$ ,  $\eta = .16$ , all one-tailed). The patterns of various incorrect responses provide some indications of how reasoning strategies differed across the conditions. Most notably, the responses of " $x$  out of 1,000", where  $x$  is something other than the base rate (85/1,000), the false positive rate (50/1,000), or the total positive rate (135/1,000), was much more common for the homogeneous-data condition than in the other conditions. Within this category of responses the answer "35 out of 1,000" was most prevalent, suggesting that the participants did understand that some calculations were needed but were unable to integrate the given information successfully.

**Table 1.** Percentages of participants producing various answers to the tasks in Experiments 1 and 2

	<i>Macchi and Mosconi, 1998</i>	<i>Experiment 1</i>			<i>Experiment 2</i>
		<i>HD (n = 40)</i>	<i>NF (n = 40)</i>	<i>NF + P (n = 40)</i>	<i>NF + P Honors (n = 23)</i>
Correct (85/135)	10 <sup>a</sup>	7.5 <sup>a</sup>	12.5	25	43.5
Base-rate (85/1,000)	30	27.5	32.5	27.5	30.4
False pos. rate (50/1,000)	0	10	10	10	4.3
Total pos. rate (135/1,000)	47	7.5	12.5	7.5	0
Other forms of $x/1,000$		37.5	15	10	4.3
Other answers		10	17.5	20	17.4

*Note:* Columns contain the results (reading left to right) from Macchi and Mosconi's (1998) homogeneous data conditions (their "Version D"), the present replication of the homogeneous data (HD) results, the problem converted to natural frequencies (NF), the natural-frequencies problem with a pictorial representation added (NF + P), and the natural-frequency- with-pictures version administered to undergraduate honours students (NF + P honours).

<sup>a</sup>Pseudo-Bayesian. The truly correct answer for the HD condition is 85 out of 130.75; using Bayes' theorem:  $(.085*1)/[(.085*1) + (.915*.05)]$ . However, neither this response nor anything approaching it (i.e., 85/130 or 85/131) was found in our study. The "pseudo-Bayesian" response of 85 out of 135, which was apparently derived from the same calculations as those in the other conditions, is the closest response to the strict Bayesian answer.

## EXPERIMENT 2

Aside from replicating the pattern of results noted for Macchi and Mosconi's (1998) study, Experiment 1 also replicated an incidental finding in that study: much lower levels of performance on the medical diagnosis task than found in previous research (e.g., Cosmides & Tooby, 1996). If, as hypothesized here, this lower performance is due to differences in intrinsic motivation/ability and extrinsic motivation (such as monetary incentive), then it should be possible to raise performance on this task in a number of ways. One way to improve performance is to select participants who have higher levels of internal motivation and ability—even within the same institution.

## Method

### *Participants*

The participants were 23 undergraduates (13 females and 10 males) at the same regional university as that in Experiment 1, but who were enrolled in a freshmen honours course. All consented to participate as part of a class demonstration during a guest lecture. Their average age was 18.1 years. Compared to the participants in

Experiment 1 (regular undergraduates at the same institution, for whom the average Grade Point Average (GPA) was 3.4, and the 25th–75th percentile range for SAT scores was 1,000–1,200 and for ACT scores was 19–23; Scholastic Aptitude Test and American College Testing, respectively), these participants were younger, had higher grades (3.83 mean entering GPA) and higher admission test scores (1,250 mean SAT, 30 mean ACT), and demonstrated higher academic ambitions, having voluntarily enrolled in the honours programme.

### *Materials and procedure*

Each participant was given a single sheet with some brief instructions and the same natural-frequencies-plus-pictures task as that used in the Experiment 1. There was no time limit for completing this study, and all participants finished within 10 minutes.

## Results

Whereas 25% of the participants reached the correct answer to this task in Experiment 1, 43% of the participants in Experiment 2 reached the correct answer to this same task (Table 1). This

difference is marginally significant ( $z = 1.52$ ,  $p = .07$ ,  $\eta = .19$ ), with an effect size on a par with the various structural changes made to tasks in any of the studies reported in this paper. The pattern of different responses suggests that these participants were much less likely (than the participants in Experiment 1) to consider the total positive rate (135/1,000) or some other intermediate calculation ( $x/1,000$ ) as the final answer.

### EXPERIMENT 3

Experiment 1 demonstrated that unpaid participants from a regional university performed better on a Bayesian reasoning task when the data were posed in natural frequencies and better still when these natural frequencies were supplemented with a pictorial representation. Freshmen honours students at the same institution (Experiment 2) performed even better than the general student population, indicating a role for internal motivation and ability. Experiments 3 and 4 were designed to further explore two factors: (a) interinstitution differences in internal motivation/ability and (b) external motivation due to monetary payments, as well as how these two motivational states are related. Specifically, Experiment 3 utilizes unpaid participants at two different types of university: a second-tier regional university and a top research university.

#### Method

##### *Participants*

A total of 55 of the participants were undergraduates at a second-tier regional university (47% with A levels for entry) in the north-east of England (40 females and 15 males). All consented to participate as partial fulfilment of an introductory psychology (i.e., first year) course requirement. Their average age was 22.7 years. The other 67 participants (45 females and 22 males) were potential undergraduates at one of the top universities in England (83% with A levels for entry). All consented to participate as part of general induction activities for the university. Their average age was 19.9 years.

##### *Materials and procedure*

Each participant was given a single sheet with some brief instructions and one of two versions of a statistical reasoning problem: the natural sampling (NS) problem and the natural sampling with pictorial representation (NS + P) problem (see Appendix). A roughly equal number of participants received each of the versions, and there was no time limit for completing this study.

### Results

Table 2 (left half) shows the percentages of participants providing various answers to each problem version. At both universities, participants given the additional pictorial representation more often reached the correct answer than those given just the natural-frequency text. This improvement, however, was not statistically significant in either case (19.2% vs. 34.5%,  $z = 1.27$ ,  $p = .10$ ,  $\eta = .17$ , and 40.5% vs. 46.7%,  $z = 0.51$ ,  $p = .31$ ,  $\eta = .06$ ). There was a significant difference between the universities, with the participants from the top-tier national university doing significantly better than those from the second-tier regional university on the NF (text only) problem (19.2% vs. 40.5%,  $z = 1.79$ ,  $p = .04$ ,  $\eta = .23$ ), but this difference was ameliorated for the problem with pictorial representations added (34.5% vs. 46.7%,  $z = 0.95$ ,  $p = .17$ ,  $\eta = .12$ ). Like the comparison of different responses across Experiments 1 and 2, it appears that participants from the top-tier university were less likely to stop with the total positive rate (135/1,000) as the final answer than were the regional-university participants.

### EXPERIMENT 4

#### Method

##### *Participants*

A total of 49 of the participants were undergraduates at a second-tier regional university in the north-east of England (15 females and 34 males). These participants were paid £3 for their participation (the equivalent of about \$5). Their

**Table 2.** Percentages of participants producing various answers to the tasks in Experiments 3 and 4

	<i>Experiment 3 (unpaid)</i>				<i>Experiment 4 (paid)</i>			
	<i>Second-tier regional university</i>		<i>Top-tier national university</i>		<i>Second-tier regional university</i>		<i>Top-tier national university</i>	
	<i>NF</i> ( <i>n</i> = 26)	<i>NF + P</i> ( <i>n</i> = 29)	<i>NF</i> ( <i>n</i> = 37)	<i>NF + P</i> ( <i>n</i> = 30)	<i>NF</i> ( <i>n</i> = 23)	<i>NF + P</i> ( <i>n</i> = 26)	<i>NF</i> ( <i>n</i> = 25)	<i>NF + P</i> ( <i>n</i> = 24)
Correct (85/135)	19.2	34.5	40.5	46.7	47.8	30.8	64.0	70.8
Base-rate (85/1,000)	23.1	24.1	29.7	13.3	17.4	19.2	0.0	12.5
False pos. rate (50/1,000)	11.5	0.0	2.7	0.0	8.7	3.8	0.0	0.0
Total pos. rate (135/1,000)	19.2	13.8	2.7	6.7	8.7	11.5	8.0	4.2
Other forms of $x/1,000$	15.4	3.4	8.1	16.7	8.7	26.9	8.0	0.0
Other answers	11.5	24.1	16.2	16.7	8.7	7.7	20.0	12.5

*Note:* Columns contain the results (reading left to right) from unpaid participants (Experiment 3) at both a second-tier regional university and a top-tier national university, on both the natural-frequencies (NF) version of the task and the natural-frequencies problem with a pictorial representation (NF + P) version of the task. The columns to the right of those show the results from paid participants (Experiment 4) at the same universities, on the same tasks.

average age was 21.8 years. The other 49 participants (22 females and 27 males) were undergraduates at one of the top universities in England. These participants were paid £5 for their participation (the equivalent of about \$8) in completing a booklet of tasks that included this study (the entire booklet took no more than one hour to complete). Their average age was 22.6 years.

### *Materials and procedure*

Each participant was given a single sheet with some brief instructions and one of two versions of a statistical reasoning problem: the natural sampling (NS) problem and the natural sampling with pictorial representation (NS + P) problem (see Appendix). A roughly equal number of participants received each of the versions, and there was no time limit for completing this study.

## Results

Table 2 (right-hand side) shows the percentages of participants providing various answers to each problem version. Although the pictorial representation slightly facilitated performance once again at the top-tier national university (64.0% vs. 70.8%,  $z = 0.51$ ,  $p = .31$ ,  $\eta = .07$ ), it did not have this effect at the second-tier regional

university (47.8% vs. 30.8%,  $z = -1.22$ ,  $p = .11$ ,  $\eta = .17$ ). There was a significant difference between the universities, with the participants from the top-tier national university this time doing significantly better than those from the second-tier regional university on the problem version with pictorial representations added (30.8% vs. 70.8%,  $z = 2.87$ ,  $p = .003$ ,  $\eta = .40$ ), but this difference was much smaller for the NF (text only) problem version (47.8% vs. 64.0%,  $z = 1.13$ ,  $p = .13$ ,  $\eta = .16$ ). With the larger number of top-tier university participants providing the correct response, there is reciprocally less to evaluate in terms of patterns of other responses. Most of the remaining responses fell into the category of “other answers”: a catch-all category that included many nearly correct responses such as 86/136, 85/130, 85/125, and 85/145. If anything, then, the percentages of correct responses under these conditions may be underestimates of competence using more lenient performance criteria.

The results of Experiments 3 and 4 can be combined, with the results dummy coded (1 for correct answer, 0 for incorrect), to permit a  $2 \times 2 \times 2$  (Type of Institution  $\times$  Type of Incentive  $\times$  Problem Version) analysis of variance (ANOVA). Two main effects were found to be

significant: paid versus unpaid inducement: overall, unpaid 36.1% versus paid 53.1%,  $F(1, 212) = 7.665$ ,  $p = .006$ ,  $\eta^2 = .035$ ; and type of institution: overall, regional university 32.7% versus national university 53.5%,  $F(1, 212) = 11.740$ ,  $p = .001$ ,  $\eta^2 = .052$ . Neither the remaining factor (presence of pictorial representation) nor any of the interactions were significant. Thus, there appears to be a simple, additive relationship between the factors of internal motivation/ability (as measured by institutional enrolment) and external motivation (as measured by type of inducement to engage in the task).

## GENERAL DISCUSSION

The high-water marks for correct performance in Bayesian reasoning come from Cosmides and Tooby (1996) and Gigerenzer and Hoffrage (1995). The participants in these studies came from very selective universities and were more likely to perform well on academic tasks (e.g., Stanford University). Furthermore, in all these studies the participants were paid for their participation, which provided strong external reinforcement for good performance (see, e.g., Camerer & Hogarth, 1999; Hertwig & Ortmann, 2001; Wright & Aboul-Ezz, 1988). The present results indicate that the much less impressive performance on the same tasks that was produced by participants in other studies appears to reflect the additive effects of one or both of these elements not being present (e.g., undergraduates from lower tier, regional universities, who were not paid). In fact, just by selecting differentially able and motivated participants (honours students) within a university, it was possible to document a substantial improvement in statistical reasoning.

The use of a pictorial representation to aid in Bayesian reasoning typically, but not uniformly, facilitated performance. Specifically, there was a small facilitation in Bayesian reasoning in four out of five comparisons between natural-frequency text and the same text with a pictorial aid.

The use of homogeneously presented data did not produce very good Bayesian reasoning, in

contrast to the claim of Macchi and Mosconi (1998). One way to reconcile their results with ours is a possible implicit difference in definitions of “Bayesian reasoning”. The present research, and much prior research, uses the more general definition—that Bayesian reasoning is the inference of a posterior probability. If Macchi and Mosconi used the more restrictive (and less common) definition that reasoning is only “Bayesian” if—and only if—the formal Bayes theorem is used (see Brase, 2002), then their original conclusions can be retained (although it then says little about human psychology beyond a very narrow form of competency). No doubt the process of Bayesian reasoning and the result of a Bayesian response can be distinguished (Villejoubert, 2003).

The best performance observed in this paper, produced by top-tier, national-university participants who were paid for their efforts and who had pictures to facilitate their performance, was 70.8%. This is impressive and is close to the 76% found by Cosmides and Tooby (1996) with their paid participants at a university of similar status in a similar “passive pictorial” condition. Some of this discrepancy can be accounted for by the relatively more stringent criteria used in this study; Cosmides and Tooby (1996) counted answers that were either exactly or approximately correct (i.e., frequency answers of “1 out of 50” or “1 out of 51” both were counted as answers of 2%). Some of the answers in this study were similarly close (e.g., 86 out of 136, 85 out of 130), and if answers within 5 percentage points of the correct answer are allowed, the percentage answering correctly in this condition increases to 79.2%. The 92% performance obtained by Cosmides and Tooby involved “active pictorial” conditions; the task required participants to actually circle parts of the picture to represent the base rate and false-alarm rates. This additional requirement, which further alters the task away from that of a simple text presentation, might have additionally facilitated performance in our participants.

Overall, the message from this research is that assessments of statistical competence (or incompetence) need to take into account the participant

pool and compensations/incentives used. As a practical matter, it is probably not feasible to have all research done at a particular tier of university or have all research conducted with paid participants (although, see Hertwig & Ortmann, 2001). It therefore makes more sense to advocate an approach that attends to the relative levels of performances across different conditions, within the same participant populations and using the same experimental methodology. So, for example, it can be noted that paying participants, compared to not-paying participants, produced an average improvement of 17 percentage points. Similarly, using top-tier, national-university participants produced about a 21-percentage-point improvement, as compared to the second-tier, regional-university students.

Further work in this area may proceed in a number of directions. Specific studies could include experiments in which monetary incentives for participants are tied to actual performance (e.g., participants who obtain the correct answer would receive larger payments) and repeated measures experiments in which the same participant is assessed under different incentive conditions (e.g., unpaid, paid, etc.).

Here we have demonstrated that payments that are not contingent on performance have a significant effect on performance. Statistical reasoning, and indeed most forms of reasoning that psychologists study, is not inherently a matter of responding to incentives in order to maximize gain (though, of course, reasoning can be modelled in this way, e.g., Oaksford & Chater, 1994). Incentives may have a different, though facilitatory, influence on reasoning performance and motivation. It has been argued that performance-contingent payments in particular are conducive to enhanced levels of performance (e.g., Hertwig & Ortmann, 2001). However, results on this are mixed, and recent research has begun to illuminate the significant effects of different magnitudes of performance-based inducements as well as interactions with broader cognitive abilities (Rydval & Ortmann, 2004). Conclusions about the generalizability of these implications across all of psychology await more extensive research, but there appear to be potentially far-reaching ramifications.

Additionally, judgements with real-world consequences, as compared to the judgements made within a psychology experiment, may involve stronger motivations for accuracy (but asymmetric pay-off matrices, and the interaction between social and financial rewards may mean that this is not so). It would also be useful to investigate whether it is possible to construct benchmarks for estimating interuniversity differences, using factors such as average standardized admission tests scores or university rankings. Such a benchmark could facilitate comparisons between research results from different universities, ideally covering a wider range of institutions (top-tier regional, liberal arts college, comprehensive universities, and community colleges). More broadly, the observed internal motivation/ability effect in this research supports recent work on individual differences in cognitive abilities as an explanatory factor in reasoning and decision making (Stanovich & West, 1998a, 1998b, 2000). Finally, these effects are probably generalizable to a number of other psychological phenomena (i.e., those that involve somewhat effortful inferential procedures) and could help explain a variety of reported variations in experimental results.

It may be tempting to describe the differences in performance across institutions as differences in internal motivation (i.e., students' intrinsic drives to perform well on academic-style tasks) and to describe differences in performance based on incentive regimens as differences in external motivation (i.e., amount of focused attention paid to the study materials, based on the nature and value of the incentives for doing so). Although these may be fair characterizations, some interinstitutional differences could also be due to differences in types of educational background, experience with similar types of task, and other demographic factors. Similarly, it may be argued that students paid to participate differ from students who participate as part of a class in several respects besides the form of inducement (e.g., paid participants were more alert and open to the prospect of participation for money, were more likely to have engaged in similar activities in the past, and had other personal characteristics that made paid participation more attractive as

an activity). Finally, there may be intermediate processes, such as changes in emotional affect (Estrada, Isen, & Young, 1997; Isen, 1993) that can be taken into account. Separating out the various contributions these factors may (or may not) make to statistical reasoning performances is necessary before these differences can be ascribed to internal and external motivation per se.

Original manuscript received 13 September 2004

Accepted revision received 13 March 2005

PrEview proof published online 29 August 2005

## REFERENCES

- Brase, G. L. (2002). Ecological and evolutionary validity: Comments on Johnson-Laird, Legrenzi, Girotto, Legrenzi, & Caverni's (1999) mental model theory of extensional reasoning. *Psychological Review*, *109*, 722–728.
- Brase, G. L., Cosmides, L., & Tooby, J. (1998). Individuation, counting and statistical inference: The role of frequency and whole-object representations in judgment under uncertainty. *Journal of Experimental Psychology: General*, *127*, 3–21.
- Brownlow, S., Gilbert, N. M., & Reasinger, R. D. (1997). Motivation, personality preferences, and interests in college students. *Journal of Psychological Practice*, *3*, 128–140.
- Camerer, C. F., & Hogarth, R. M. (1999). The effects of financial incentives in experiments: A review and capital labor production framework. *Journal of Risk and Uncertainty*, *19*, 7–42.
- Casscells, W., Schoenberger, A., & Graboys, T. (1978). Interpretation by physicians of clinical laboratory results. *New England Journal of Medicine*, *299*, 999–1000.
- Cosmides, L., & Tooby, J. (1996). Are humans good intuitive statisticians after all?: Rethinking some conclusions of the literature on judgment under uncertainty. *Cognition*, *58*, 1–73.
- Estrada, C. A., Isen, A. M., & Young, M. J. (1997). Positive affect facilitates integration of information and decreases anchoring in reasoning among physicians. *Organizational Behavior & Human Decision Processes*, *72*, 117–135.
- Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review*, *102*, 684–704.
- Gigerenzer, G., & Hoffrage, U. (1999). Overcoming difficulties in Bayesian reasoning: A reply to Lewis and Keren (1999) and Mellers and McGraw (1999). *Psychological Review*, *106*, 425–430.
- Hertwig, R., & Ortmann, A. (2001). Experimental practices in economics: A methodological challenge for psychologists? *Behavioral and Brain Sciences*, *24*, 383–404.
- Isen, A. M. (1993). Positive affect and decision making. In M. Lewis & J. M. Haviland (Eds.), *Handbook of emotions* (pp. 261–277). New York: Guilford Press.
- Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology*, *3*, 430–454.
- Kleiter, G. (1994). Natural sampling: Rationality without base rates. In G. H. Fischer & D. Laming (Eds.), *Contributions to mathematical psychology, psychometrics, and methodology* (pp. 375–388). New York: Springer-Verlag.
- Lin, Y., McKeachie, W. J., & Kim, Y. C. (2001). College student intrinsic and/or extrinsic motivation and learning. *Learning & Individual Differences*, *13*, 251–258.
- Macchi, L., & Mosconi, G. (1998). Computational features vs. frequentist phrasing in the base-rate fallacy. *Swiss Journal of Psychology*, *57*, 79–85.
- McClelland, A., & Cheng, W. Y. (2004). *How to improve Bayesian reasoning: Culture, education and frequency formats*. Manuscript in preparation. University College London.
- Mellers, B. A., & McGraw, A. P. (1999). How to improve Bayesian reasoning: Comment on Gigerenzer and Hoffrage (1995). *Psychological Review*, *106*, 417–424.
- Oaksford, M., & Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review*, *101*, 608–631.
- Phillips, P., Abraham, C., & Bond, R. (2003). Personality, cognition, and university students' examination performance. *European Journal of Personality*, *17*, 435–448.
- Ruscio, J. (2003). Comparing Bayes's theorem to frequency-based approaches to teaching Bayesian reasoning. *Teaching of Psychology*, *30*, 325–328.
- Rydval, O., & Ortmann, A. (2004). How financial incentives and cognitive abilities affect task performance in laboratory settings: An illustration. *Economics Letters*, *85*(3), 315–320.

- Sedlmeier, P., & Gigerenzer, G. (2001). Teaching Bayesian reasoning in less than two hours. *Journal of Experimental Psychology: General*, *130*, 380–400.
- Sklaroff, A. (Ed.). (2004). America's best colleges. *US News & World Report*, April.
- Slovan, S. A., Over, D., Slovak, L., & Stibel, J. M. (2003). Frequency illusions and other fallacies. *Organizational Behavior and Human Decision Processes*, *91*, 296–309.
- Stanovich, K. E., & West, R. F. (1998a). Individual differences in rational thought. *Journal of Experimental Psychology: General*, *127*, 161–188.
- Stanovich, K. E., & West, R. F. (1998b). Who uses base rates and P(D/H)? An analysis of individual differences. *Memory and Cognition*, *26*, 161–179.
- Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Sciences*, *23*, 645–726.
- Talbot, G. L. (1990). Personality correlates and personal investment of college students who persist and achieve. *Journal of Research & Development in Education*, *24*, 53–57.
- U.S. News & World Report (2004). America's Best Colleges, 2004 Edition (A. Sklaroff, Ed.) *U.S. News & World Report*.
- Villejoubert, G. (2003). *Posterior probability judgements: Distinguishing numerical outcomes from their underlying reasoning processes*. Unpublished PhD thesis, University of Hertfordshire, Hatfield, UK.
- Wright, W. F., & Aboul-Ezz, M. E. (1988). Effects of extrinsic incentives on the quality of frequency assessments. *Organizational Behavior and Human Decision Processes*, *41*, 143–152.

## APPENDIX

### Text of the tasks used in Experiments 1–4

#### *Homogeneous-data version (HD)*

85 out of 1000 Americans have disease X. A test has been developed to detect when a person has disease X. Every time the test is given to a person who has the disease, the test comes out positive. But sometimes the test also comes out positive when it is given to a person who is completely healthy. Specifically, out of every 1000 people who are perfectly healthy, 50 of them test positive for the disease.

Imagine that we have assembled a random sample of 1000 Americans. They are selected by a lottery. Those who conducted the lottery had no information about the health status of any of these people.

How many people who test positive for the disease will actually have the disease?

\_\_\_\_ out of \_\_\_\_

#### *Natural-frequencies version (NF)*

85 out of 1000 Americans have disease X. A test has been developed to detect when a person has disease X. Every time

the test is given to a person who has the disease, the test comes out positive. But sometimes the test also comes out positive when it is given to a person who is completely healthy. Specifically, out of the remaining 915 perfectly healthy people in every 1000 Americans, 50 of them test positive for the disease.

Imagine that we have assembled a random sample of 1000 Americans. They are selected by a lottery. Those who conducted the lottery had no information about the health status of any of these people.

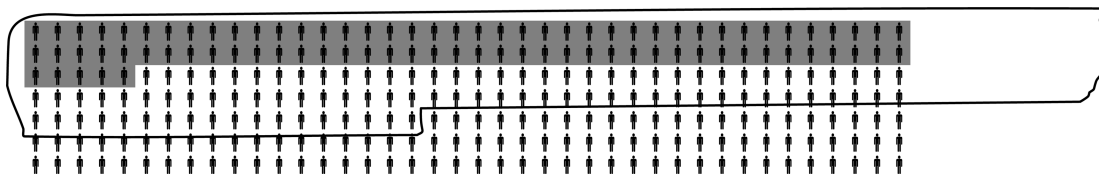
How many people who test positive for the disease will actually have the disease?

\_\_\_\_ out of \_\_\_\_

#### *Natural frequencies + pictorial representation version (NF + P)*

85 out of 1000 Americans have disease X. . . . [same introduction as NF version]

Another way to view this is shown below. 1000 Americans are represented by the 1000 figures printed below (in 25 rows of 40). Figures that are darkened are those persons with the disease. Figures that are circled are those persons who test positive for the disease.



[18 more rows shown like the row immediately above]

Imagine that we have assembled a random sample. . . . [remainder the same as NF version]