Journal of Behavioral Decision Making
J. Behav. Dec. Making, **21**: 457–470 (2008)
Published online 5 June 2008 in Wiley InterScience
(www.interscience.wiley.com) **DOI**: 10.1002/bdm.611

The Public's Probabilistic Numeracy: How Tasks, Education and Exposure to Games of Chance Shape It

RALPH HERTWIG*, MONIKA ANDREA ZANGERL, ESTHER BIEDERT and JÜRGEN MARGRAF
Department of Psychology, University of Basel, Switzerland

ABSTRACT

As we navigate a world full of uncertainties and risks, dominated by statistics, we need to be able to think statistically. Very few studies investigating people's ability to understand simple concepts and rules from probability theory have drawn representative samples from the public. For this reason we investigated a representative sample of 1000 Swiss citizens, using six probabilistic problems. Most reasoned appropriately in problems representing pure applications of probability theory, but failed to do so in approximations of real-world scenarios – a disparity we replicated in a sample of first-year psychology students. Additionally, education is associated with probabilistic numeracy in the former but not the latter type of problems. We discuss possible reasons for these task disparities and suggest that gaining a comprehensive picture of citizens' probabilistic competence and its determinants requires using both types of tasks. Copyright © 2008 John Wiley & Sons, Ltd.

KEY WORDS probabilistic reasoning; cognitive illusions; education; representative sample

INTRODUCTION

The English novelist H. G. Wells, the author of science fiction classics such as *The Time Machine* and *The War of the Worlds*, also wrote pamphlets attacking the Victorian social order. In *Mankind in the Making* (1903), he argued that it is ''the first and most universal function of the school to initiate a smaller or greater proportion of the population into the ampler world ... of the reading and writing man. [...] there is also almost universally in schools instruction in counting, and wherever there is a coinage, in the values and simpler computation of coins.'' With some likelihood it is this quotation that later morphed into the famed prediction,

* Correspondence to: Ralph Hertwig, Department of Psychology, University of Basel, Missionsstrasse 60/62, 4055 Basel, Switzerland.
E-mail: ralph.hertwig@unibas.ch

attributed by scores of authors to Wells, that ''Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write.''[1]

About a century later, a number of people investigated how far we had – or rather had not – come in this respect. The mathematician Paulos (2001), for instance, diagnosed a widespread *innumeracy* – ''an inability to deal comfortably with the fundamental notions of number and chance'' (p. 3). He related the story of a weather forecaster on American television who reported that there was a 50% chance of rain on Saturday and a 50% of rain on Sunday, concluding that there was a 100% chance of rain that weekend. The inability to reason appropriately about chance is by no means solely an American affliction. In a survey, 1000 Germans were asked what ''40%'' means: (a) one quarter, (b) 4 out of 10, or (c) every 40th person. About one-third of respondents got the answer wrong (see Gigerenzer, 2002). Recently, Lipkus, Samsa, and Rimer (2001) developed a numeracy scale, testing basic mathematical and probability concepts and operations, and found that in a sample of highly educated participants, on average, merely 18% and 32% of participants answered correctly all of the general and expanded numeracy scale items.

Numeracy matters

Various researchers have argued that some numerical ability fosters good decisions in financial, medical, and other domains. Greater numeracy, for instance, appears to be linked with more accuracy in judgments about probabilities associated with prostate cancer screening, with accurately gauging the benefit of mammography (Schwartz, Woloshin, Black, & Welch, 1997), and more reliance on numeric rather than verbal risk information from physicians (Black, Nease, & Tosteson, 1995; for more references see Peters, Västfjäll, Slovic, Mertz, Mazzocco, & Dickert, 2006).[2]

The present study has two goals. First, we investigate the extent of *probabilistic numeracy* (or lack thereof) in a *representative sample of Swiss citizens*. With probabilistic numeracy, we mean the knowledge of basic concepts and rules of probability such as the multiplication rule, the conjunction rule, the empirical law of large numbers, and the ability to perform, if necessary, the adequate calculations. Probabilistic numeracy focuses exclusively on statistical reasoning, unlike Lipkus et al.'s (2001) more encompassing numeracy scale.[3] The second goal is to investigate three factors that may determine people's probabilistic numeracy, namely, (a) the nature of the tasks, (b) people's level of education, and (c) their exposure to games of chance.

THE CITIZEN: AN INTUITIVE STATISTICIAN OR BEDEVILED BY COGNITIVE ILLUSIONS?

During most of the second half of the 20th century, two groups of researchers were concerned with the extent to which people's intuitive statistical reasoning conforms to laws of probability theory and statistics. In their programmatic review of empirical work conducted in the 1950s and 1960s, titled *Man as an intuitive statistician*, Peterson and Beach (1967) concluded that ''probability theory and statistics can be used as the basis for psychological models that integrate and account for human performance in a wide range of inferential tasks'' (p. 29). In their view, the psychological laws of people's statistical reasoning and the laws

---

[1]This statement is quoted, for instance, in *How to Lie with Statistics* (Huff, 1954/1993), where it serves as an epigraph. No reference is given, also in none of the many other texts in which one can find it to be quoted.

[2]For the development of a ''subjective'' numeracy scale that is less reminiscent of a mathematics test but is still predictive of risk comprehension, see Fagerlin, Zikmund-Fisher, Ubel, Jankovic, Derry, and Smith (2007) and Zikmund-Fisher, Smith, Ubel, and Fagerlin (2007).

[3]Their numeracy scale tests basic abilities such as (a) converting one quantity into another (e.g., 1% to 10 in 1000), (b) discerning differences in magnitudes of risks (e.g., which of the following numbers represents the biggest risk of getting a disease? 1%, 10%, 5%), and (c) performing simple operations involving percentages and proportions (e.g., If person A's risk of getting a disease is 1% in ten years, and person B's risk is double that of A's, what is B's risk?).

DOI: 10.1002/bdm

of probability theory are not in conflict. In contrast, the subsequent heuristics-and-biases research program has identified an extensive catalogue of deviations between people's statistical reasoning and rules from probability theory and statistics. These deviations – ''cognitive illusions'' – include the conjunction fallacy, insensitivity to sample size, base-rate neglect, and overconfidence (e.g., Gilovich, Griffin, & Kahneman, 2002; Kahneman, Slovic, & Tversky, 1982). People are subject to these cognitive illusions because they tend to rely on a limited number of heuristics that ''reduce the complex task of assessing probabilities and predicting values to simple judgmental operations'' (Tversky & Kahneman, 1974, p. 1124).

How can the citizen's mind appear to be akin to that of an intuitive statistician and yet error-prone because it is powered by quick and dirty heuristics? Although it is tempting to attribute the drastically different views about human rationality to the drastically different political climates of the 1960s and 1970s – revolutionary social changes and the promise of unlimited economic and technological progress (e.g., the first man on the moon) in the sixties versus growing disillusionment toward institutions (e.g., Watergate), the oil crisis and economic depression in the seventies – more prosaic factors are likely to have triggered the darker view of the seventies.

One factor is the kind of tasks used to probe probabilistic numeracy. Urns and balls, coins and cards – that is, games of chance – have long been the staple of probabilists. For instance, the famous exchange of letters between the French mathematicians Blaise Pascal and Pierre Fermat, in which the fundamental principles of probability were first formulated, was exclusively devoted to gambling problems (Gigerenzer, Swijtink, Porter, Daston, Beatty, & Krüger, 1989). The famous St. Petersburg paradox, according to which people are willing to pay only small amounts to take a gamble that is assumed to have an infinite expected value, involves nothing but a fair coin. In the 1960s, gambling problems made their way into the laboratories of psychologists such as Peterson and Beach or Edwards, although often in a more contemporary terminology, for instance, as bookbag-and-poker chips problems (e.g., Peterson & Beach, 1967, p. 32). In the context of those games of chance, Peterson and Beach's concluded, respondents' statistical reasoning is akin to that of an ''intuitive statistician.''

In the 1970s the kind of problems posed by experimenters changed. The problems were no longer pure applications of probability theory, and they typically did not embody games of chance. Although still of the textbook variety, the problems designed by Tversky and Kahneman (1974) and others approximated real-life situations. Compared with the bookbag-and-poker chips problems, they are more interesting and engaging but at the expense of being more ambiguous in their interpretation. We return to these costs of the new tasks shortly.

In our study, we used six tasks; four are pure applications of probability theory (henceforth ''theory problems'') and two approximate real-world situations (henceforth ''real-world problems''). Table 1 lists them and the simple concepts and rules from probability theory that they represent. The theory problems probe people's understanding of the probability of a simple event A and of its complement $\neg A$ (bag and ball task, the die and die_C tasks), and the multiplication rule for two independent events (coin task). The two real-world problems are two classic tasks from the heuristics-and-biases research program, the Linda and the maternity-ward tasks.

The Linda task was designed to investigate whether people's judgments accord with the ''simplest and most fundamental qualitative law of probability'' (Tversky & Kahneman, 1983, p. 294), the conjunction rule. This rule holds that the probability of a conjoint hypothesis (A&B) cannot exceed that of either of its constituents, that is, $p(\text{A\&B}) \leq p(\text{A}), p(\text{B})$. In conflict with this rule, in numerous studies most people ranked Linda to be more likely a bank teller and active in the feminist movement than a bank teller. The suggested explanation was that when confronted with a difficult task, people often answer an easier one instead, for example, they judge representativeness rather than probability (thus employing the representativeness heuristic). Indeed, Linda resembles a prototypical feminist bank teller more than she resembles a prototypical bank teller. Studying intuitive thinking in terms of representativeness was also the focus of the second real-world problem that we included, the maternity-ward problem (Table 1). In Tversky and Kahneman's

DOI: 10.1002/bdm

Table 1. Probabilistic reasoning tasks: the first four tasks represent pure applications of probability theory (''theory problem'') and the last two tasks represent approximation of real-world situations (''real-world problem'')

Tasks	Text	Response alternatives	Concepts and Rules
Bag and ball	Imagine a bag filled with 10 red balls and 10 blue balls. What is the probability of drawing a red ball?	10% 20% **50%** Don't know	Definition of the probability of a simple event *A*: Number of favorable events/number of total possible events
Coin	You toss two coins. What is the probability that both coins come up heads on this toss?	**25%** 35% 50% Don't know	Multiplication rule for independent events: $p(\mathrm{AB}) = p(\mathrm{A})p(\mathrm{B})$
Die	Imagine a regular die. What is the probability of rolling a ''3'' at the first toss?	1:3 (a third) 1:2 (half) **1:6 (one sixth)** Don't know	Definition of the probability of a simple event *A*: Number of favorable events/number of total possible events
Die_C	Imagine a regular die. What is the probability of not rolling a ''3'' at the first toss?	1:3 (a third) 2:3 (two thirds) **5:6 (five sixths)** Don't know	Definition of probability of event complementary to simple event *A*: $p(\neg \mathrm{A}) = 1 - p(\mathrm{A})$
Linda	Linda is 31 years old, and single. As a student, she was vehemently involved in fighting different forms of discrimination, and also participated in anti-nuclear demonstrations. Which of the following statements is more probable?	**Linda is a bank teller**. Linda is a bank teller and active in the feminist movement? I don't know	Conjunction rule: $p(\mathrm{A} \wedge \mathrm{B}) \leq p(\mathrm{A}), p(\mathrm{B})$
Maternity ward	A town is served by two hospitals. In the larger hospital about 50 babies are born each day, and in the smaller hospital about 10 babies are born each day. On average, half of all babies are boys. On some days, however, more boys than girls are born. In which hospital do such days occur more frequently?	In the small hospital **In the large hospital** Equally frequent in both hospitals I don't know	Cumulative binomial probability

The bold response represents the correct alternative or what the authors of the task proposed to be the correct answer (see text for our discussion of the Linda task).

Journal of Behavioral Decision Making, 21, 457–470 (2008)
DOI: 10.1002/bdm

(1974) study, most participants thought that the number of days in which more babies born were boys was about the same in the small and the large hospital, ''presumably because these events are described by the same statistic and are therefore equally representative of the general population'' (p. 1125). Sedlmeier and Gigerenzer (1997) pointed out that the maternity-ward problem has been studied in two versions, one involving a frequency distribution and the other involving a sampling distribution. The former is a distribution of values from one sample. In contrast, the latter is a distribution of means from independent samples of fixed size, drawn from the same population. Our version of the maternity-ward problem involves a sampling distribution.[4]

Before we turn to possible determinants of the public's probabilistic numeracy, let us return to the potential costs of tasks that approximate real-life situations. One of the costs is an ongoing controversy concerning the normative status of some of the cognitive illusions documented (e.g., Gigerenzer, 1996, 2001; Kahneman & Tversky, 1996; Vranas, 2000). The Linda task is a case in point. Gigerenzer (1996) has argued that characterizing violations of the conjunction rule (and various other rules) as an unequivocal ''cognitive illusion'' ignores the existence of conflicting interpretations of probability theory. According to frequentistic interpretation of probability, for instance, a single-event probability judgment, as required by the Linda task, is simply meaningless. But even if single-event probabilities are seen as meaningful, as they are by subjectivist interpretations of probability, the semantic and pragmatic ambiguities inherent in the Linda task make violations of the conjunction rule defensible (see e.g., Adler, 1984, 1991; Chase, Hertwig, & Gigerenzer, 1998; Hertwig & Gigerenzer, 1999; Hertwig, Benz, & Krauss, 2008; Hilton, 1995; Politzer & Noveck, 1991; but see also e.g., Mellers, Hertwig, & Kahneman, 2001; Tentori, Bonini, & Osherson, 2004). Related concerns have also been raised with regard to other tasks within the heuristics-and-biases program such as the Bayesian reasoning tasks (e.g., Birnbaum, 1983). For the purpose of this paper, this controversy suggests that there is no consensus in the literature as to whether or not people who violate the conjunction rule in the Linda task are guilty of having committed a cognitive illusion.

WHAT DETERMINES CITIZENS' PROBABILISTIC NUMERACY?

To the best of our knowledge, hardly any study has investigated the degree of the public's probabilistic numeracy using a *representative sample of citizens*.[5] Henrich (2001) highlighted both psychologists' and economists' ''extreme reliance ... on university students as subjects. Many scholars from both fields are guilty of reaching conclusions about 'human reasoning' ... by relying entirely on this very weird, and very small, slice of humanity'' (p. 414). Although our study is not representative of humanity – it does not cut across cultures – it goes far beyond university students or specific professions. In a representative sample of Swiss citizens we tested three (by no means exclusive) hypotheses.

According to the first hypothesis, the public's probabilistic numeracy depends on the kinds of tasks used (*task hypothesis*). Specifically, we predict that faced with pure applications of probability theory, people will

[4]In the original Tversky and Kahneman (1974) version of the task, respondents were told: ''For a period of 1 year, each hospital recorded the days on which (more/less) than 60% of the babies born were boys. Which hospital do you think recorded more such days?'' In this task, the empirical law of large numbers—the intuition that larger samples generally lead to more accurate estimates of population means (Freedman et al., 1991)—prescribes that the days on which more baby boys are expected to be born is greater in the small than in the large hospital. In a pilot test, we noticed that this version is difficult to understand for people, talking on the phone. Therefore, we chose a slightly simpler version based on the wording used of Evans and Dusoir (1977, p. 133–134). In this version, the representativeness heuristic still predicts that respondents reason that the number of days in which more babies born were boys is about the same in the small and the large hospital. The statistically correct answer, however, is the large hospital.

[5]Although there are numerous studies investigating particular professionals such as physicians, investment bankers and meteorologists (e.g., Eddy, 1982; Murphy & Winkler, 1984), and few studies have employed the Lipkus et al. (2001) numeracy scale in a representative sample (e.g., Fagerlin et al., 2007).

DOI: 10.1002/bdm

more likely reason in accordance with probability theory. In contrast, approximations of real-world situations will yield lower levels of correct responses. This hypothesis is derived from the discrepancy between the results reviewed in Peterson and Beach (1967) and those results in the tradition of the heuristics-and-biases program.

According to the second hypothesis, the public's probabilistic numeracy will also depend on the level of education (*education hypothesis*). This hypothesis is derived from Stanovich and West's (2000) investigations of individual differences in reasoning and the finding that, across numerous probabilistic reasoning tasks (investigated within the heuristics-and-biases research program), there is a positive association between performance in the tasks and cognitive ability, as measured in terms of SAT scores. They explained this association by the assumption of two reasoning systems, one associative and the other rule-based, and that students with higher cognitive ability are more likely to give rule-based responses. Rather than using SAT scores (SATs do not exist in Switzerland), we use the level of education as a proxy for cognitive ability. We also investigate whether the association between performance in the reasoning tasks and cognitive ability is equally robust across theory and real-world tasks, respectively. If not, this would suggest that the different kinds of tasks indeed elicit different cognitive processes, which, in turn, could have triggered the discrepant findings in the traditions of Peterson and Beach (1967) and heuristics-and-biases.

The third hypothesis specifies another factor that may underlie citizens' probabilistic numeracy (*exposure hypothesis*). Not only early probabilists such as Pascal, Fermat, and Huygens took advantage of games of chance to develop probability theory: Partaking in games of chance may also provide Jane or Joe Average with a learning vehicle with which to test their intuitions about chance. If so, it is reasonable to expect that those people with more exposure to games of chance have had more opportunity to sharpen their intuitions about chance and probabilities.

METHODS

One thousand respondents participated in this study. A random-quota method according to region (75.6% German speakers, 24.4% French speakers; 68.2% participants from urbanized vs. 31.8% from rural areas), sex (49.4% men, 50.6% women), age (15–29 years: 23.8%; 30–49 years: 41.1%, and 50–74 years: 35.2%), and household size (1–2 people: 43.6%; 3–4: 40.4%; and 5 and more people: 16%) was used to proportionally represent the Swiss population (over age 14), based on the census in 2000 (Swiss Federal Office of Statistics). To represent the basis of these census figures, samples were weighted by the quota method using a weighting variable (values ranged from 0.64 to 1.74). Therefore the data can result in rounding discrepancies. A survey research firm, the GfK group, experienced in telephone surveys performed the interviews (in September 2004). Respondents were randomly selected by using computer-generated telephone numbers, and up to 50 attempts were made to contact a telephone number (see Zangerl, Munsch, Meyer, & Margraf, 2008).

The interview was designed to investigate in a representative sample of Swiss citizens (a) their probabilistic numeracy, (b) their exposure to games of chance, and (c) the prevalence of pathological gambling. To probe pathological gambling, the standardized revised South Oaks Gambling Screen (SOGS-R; Lesieur & Blume, 1993), adapted for a telephone survey, was administered to those respondents who reported to have gambled in the last year. In addition, all respondents were asked to report their experience with each of eleven ''game of chance'' activities (Table 2), and they responded to the six tasks measuring probabilistic numeracy (Table 1). Due to a mistake in administering the French version of the Linda and the maternity-ward tasks, we only include the 756 German-speaking respondents for those two tasks.

RESULTS

Overall, answers to the SOGS-R indicated that 1.6% respondents in our sample are probable pathological gamblers, and 1.8% are potential pathological gamblers. The prevalence rate of probable pathological

DOI: 10.1002/bdm

Table 2. Exposure to games of chance: item that probed citizens' direct involvement (item 1) with games of chance

Items	Options/response categories	Results
1. Have you ever participated (in whichever form) in one of the following gambling activities?	Playing lotto/toto, games in casino, cards for money, slot machine, the stock market or option market, bingo for money, ability games for money, dice for money; buying lottery scratch tickets; betting on sports, racing	26% (256) *never* participated during life-time; 74% (744) participated in at least one gambling activity during life time

gambling is about double that found in an earlier Swiss epidemiological study (Bondolfi, Osiek, & Ferrero, 2000). This sample of gamblers does not differ in education or in probabilistic numeracy from the remaining sample nor are lower levels of education a risk factor for pathological gambling (Zangerl et al., 2008). We will not elaborate further on the epidemiological and clinical aspects (see Zangerl et al.) but turn to citizens' probabilistic numeracy and the three hypothesized determinants.

The impact of task

Does citizens' probabilistic numeracy hinge on the kind of task? Clearly it does. Table 3 shows the proportion of correct responses averaged across levels of education for the four theory tasks. Proportion correct ranges between 45.5% (coin problem) and 79.7% (bag-and-ball problem), with an average score of 64.7% (German-speaking respondents: 65.8%; French-speaking respondents: 61.3%). In contrast, in the two real-world tasks – Linda and maternity ward – merely 12.9% and 28.8% of respondents, respectively, arrived at an answer consistent with the rule (Table 4). In other words, citizens' probabilistic numeracy, or lack thereof, depends on the kind of tasks used to probe their competence. Moreover, the level in the theory tasks is relatively high, compared with chance performance (33%).

To test the robustness of this task disparity in probabilistic numeracy we investigated another sample of respondents, consisting of 328 first-year psychology students (45 male and 283 female students; 134 from the University of Geneva and 194 from the University of Basle). They responded to the same six probabilistic reasoning tasks used in the telephone interview. The proportions correct in the four theory tasks were 94%, 79%, 95%, and 92%, respectively, and thus, on average, slightly higher than in the high-education group of

Table 3. Percentage of correct responses in the four ''theory problems'' (level of chance performance per task: 33%; $N = 1000$ respondents)

Level of education/tasks	Bag and ball	Coin	Die	Die_C	Average proportion of correct responses	Proportion of ''perfect score''[a]
Low (Swiss elementary school, secondary school, junior high school); $N = 157$	65.6	44.6	52.9	46.5	52.4	19.1
Medium (vocational school, high school); $N = 606$	79.5	41.7	68.9	59.2	62.3	21.9
High (college, university); $N = 237$	89.6	55.7	91.1	79.2	78.9	44.1
Average proportion of correct responses	79.7	45.5	71.7	62.0	64.7	26.7

Note: In order to not overestimate the public's level of numeracy, we treated ''I don't know'' and missing responses as incorrect responses (the combined percentage of ''I don't know'' and missing responses ranged between 0.8% and 12.1% across cells).
[a]A person has a ''perfect score'' if her/his responses are all correct.

DOI: 10.1002/bdm

Table 4. Percentage of correct responses in the two ''real-world problems'' (level of chance performance per task: 33%; $N = 756$ German-speaking respondents)

Level of education/tasks	Linda[a]	Maternity ward	Average proportion of correct responses	Proportion of ''perfect score''[b]
Low (Swiss elementary school, secondary school, junior high school); $N = 124$	14.5	38.4	26.5	7.3
Medium (vocational school, high school); $N = 461$	11.9	29.1	20.5	2.0
High (college, university); $N = 171$	14.0	21.1	17.6	3.5
Average proportion correct responses[a]	12.9	28.8	20.9	3.2

Note: In order to not overestimate the public's level of numeracy, ''I don't know'' and missing responses were classified as incorrect responses (the combined percentage of ''I don't know'' and missing responses ranged between 1.8% and 16.4% across cells).
[a]See text for the controversy over whether people violating the conjunction rule have committed a cognitive illusion.
[b]A person has a ''perfect score'' if her/his responses are all correct.

our representative sample (90% vs. 78.9%).[6] In contrast, the performance in the Linda and maternity-ward tasks was 13% and 12%, respectively. Thus, the disparity between the two types of tasks was even larger in the student than in the representative citizen sample (77% vs. 43.8% points). We return to possible explanations of this disparity in the discussion.

The impact of education

Does people's probabilistic numeracy hinge on their level of education? Table 3 shows the number of correct responses for the four theory tasks, separately for respondents' highest level of education attained. Except in the coin task, probabilistic numeracy consistently increases with education, supporting the education hypothesis. In the high-education group, on average, 78.9% of all responses were correct, compared with a still much better than chance performance of 52.4% average score in the low-education group. Table 3 also shows how many respondents at each level of education answered all four theory tasks correctly. In the high-education group 44.1% of respondents achieved a perfect score, relative to 19.1% in the low-education group. An analysis of variance showed levels of education for the four theory tasks to be significant ($F(2,1016) = 50.609$, $p < .001$): with a higher level of education the number of correct responses increased linearly (linear contrast $p < .001$).

Results look drastically different in the two real-world tasks, however. Averaged across the four theory tasks, respondents with a high level of education scored 26.5% points more correct responses than those with a low level of education (Table 3). Averaged across the two real-world tasks, this education performance edge disappeared and was, in fact, even slightly reversed (due to the relatively high score of the low-education group in the maternity-ward problem) (Table 4). The negligible impact of education is also manifest in the ''perfect score'': merely 7.3% (low), 2.0% (medium), and 3.5% (high education) of respondents arrived at correct answers in both tasks, respectively. Relatedly, among those highly educated respondents who conformed to the multiplication rule in the coin task, merely 10% (17 out of 171) conformed to the related conjunction rule in the Linda task. The lack of a positive effect of education in the real-world tasks appears inconsistent with Stanovich and West's (1998) findings, an issue to which we return in the discussion.

[6]This difference is likely to be due to the fact that the first-year psychology students were in the middle of attending Statistics 101, whereas for most people in the high-education group academic training dated back numerous years (average age = 44 years).

DOI: 10.1002/bdm

Table 5. Sex differences in percentage correct as a function of tasks and levels of education, respectively (positive differences indicate better male than female performance)

	Level of education		
Tasks	Low	Medium	High
Bag and ball	29.9	5.5	14.0
Coin	−9.9	4.9	10.4
Die	14.9	13.7	3.5
Die_C	13.3	13.3	8.7
Linda	13.2	−0.8	−2.1
Maternity ward	−2.3	3.9	−15.1
Average proportion correct responses	9.9	6.8	3.2

The impact of exposure to games of chance

To investigate the extent to which exposure to games of chance affects probabilistic numeracy, we probed citizens' respective exposure (Table 2). Nearly three fourths (744 out of 1000) reported to have participated in at least one gambling activity and 26% reported never to have gambled at all. To find out whether people with exposure are more likely to give correct answers, we focus on the four theory problems. (We omit the two-real world problems for this analysis because there is little variance between people on them, making it difficult to find any association.) Consistent with the exposure hypothesis, we found that people with no exposure solved, on average, 2.4 problems of the four theory problems, whereas people with exposure solved 2.6 ($t_{(998)} = -2.513$; $p = .006$). The effect size of this difference is small ($d = 0.18$; Rosenthal & Rosnow, 1991).

Finally, we also examined another variable that has been reported to be associated with mathematical quantitative skills in general, namely, sex (for a excellent review of the findings and candidate explanations; see Halpern, Benbow, Geary, Gur, Hyde, & Gernsbacher, 2007). Table 5 reports sex differences across all six tasks times three education levels. In 13 of the 18 differences and across levels of education, males performed better than females, although sometimes merely by a small margin (Binomial test: $p = .03$).

DISCUSSION

Our investigation of a representative sample of Swiss citizens showed that they have some basic knowledge of simple probability concepts. In problems that represent pure applications of probability theory, such as determining the probability of drawing a red ball from a bag with 10 red and 10 blue balls, 80% of all responses were correct. One should, however, not overestimate the level of competence. Consider the coin task that requires the multiplication of two independent probabilities and is thus the most complex of the theory tasks. It resulted in 46% correct responses (Table 3; 79% in the student sample).

The level of performance was quite different in the two tasks that represented approximations of real-world situations. Merely 21% of responses conformed to the probability rules and a tiny proportion of 3.2 respondents gave correct responses to both tasks (Table 4). There are at least three explanations for this striking disparity in probabilistic numeracy that we discuss next.

Real-world tasks are designed to elicit errors

One possible reason for the disparity in probabilistic numeracy is that the tasks used in the heuristics-and-biases program, unlike the theory tasks, were designed with the explicit goal of triggering errors in probabilistic reasoning. Just as vision researchers construct situations in which the functioning of the visual system leads to incorrect inferences about the world (e.g., about line lengths in the Müller-Lyer

DOI: 10.1002/bdm

illusion), researchers in the heuristics-and-biases program construct problems such that reasoning by cognitive heuristics leads to violations of probability theory. In their work on violations of the conjunction rule, Tversky and Kahneman (1983) mentioned this design feature explicitly: ''Our problems, of course, were constructed to elicit conjunction errors'' (p. 311).

If this error propensity in real-world tasks were the reason for the disparity in probabilistic numeracy, it would have interesting implications for the debate on how dysfunctional people's probabilistic reasoning is. The conclusions drawn from the study of visual illusions often differ sharply from those drawn by proponents of the heuristics-and-biases program. Vision scientists do not conclude from the robustness of the Müller-Lyer illusion, for instance, that people are generally poor at inferring object lengths. In contrast, sweeping conclusions have been drawn about people's ability to reason probabilistically, both by researchers within and outside of the heuristics-and-biases program. Slovic, Fischhoff, and Lichtenstein (1976), for instance, concluded: ''It appears that people lack the correct programs for many important judgmental tasks ... it may be argued that we have not had the opportunity to evolve an intellect capable of dealing conceptually with uncertainty'' (p. 174). The conjunction fallacy impelled paleontologist Stephen Jay Gould (1992, p. 469) to speculate: ''Our minds are not built (for whatever reason) to work by the rules of probability.'' It is worth pointing out that Tversky and Kahneman (1983) cautioned that because their problems were constructed to elicit violations of the conjunction rule, ''they do not provide an unbiased estimate of the prevalence of these errors'' (p. 311).

We suggest that studies involving representative samples and a wide range of tasks (ideally, a representative sample of tasks gauging probabilistic numeracy; Dhami, Hertwig, & Hoffrage, 2004) promise to provide us with a less biased estimate of the public's knowledge of basic rules of probability.

Real-world tasks may have more than one ''correct'' solution

Another possible reason for the disparity in the public's probabilistic numeracy relates to the issue we discussed earlier: the debate regarding the normative status of violations of the conjunction rule in the Linda task. One group of researchers has argued that these violations are defensible; in fact, that they represent intelligent and socially rational inferences (e.g., Hertwig & Gigerenzer, 1999). Another group of researchers has interpreted them as genuine reasoning errors (e.g., Tentori et al., 2004). The theory problems, pure application of probability theory, rendered it crystal-clear that these tasks are about probabilistic reasoning. In the Linda task, however, participants have to infer what the experimenter means by ''probable,'' a term that in natural language has multiple, related meanings, most of which cannot be reduced to mathematical probability (e.g. ''plausible'' or ''conceivable''). In light of this and other ambiguities in the Linda task, it is also noteworthy that 135 of the 328 (41%) university students selected the ''don't know'' option in the Linda task, whereas in all other tasks this option was chosen by fewer than 10% of student respondents. One interpretation of this high number of ''don't know'' responses is that many people are reluctant or even cautious to generalize their knowledge of a basic rule of probability to the Linda task.

One further way to test the ambiguity explanation is to convert ambiguous real-world tasks into less ambiguous theory tasks, and study whether accordance with the assumed norm increases. For the Linda task, Tversky and Kahneman (1983) did so by omitting Linda's personality and thus turning the task into a pure application of the conjunction rule. As a consequence ''almost all respondents obeyed the conjunction rule'' (p. 305). They also studied a task that admittedly could be seen as a theory task, namely, the die task (Tversky & Kahneman, 1983, p. 303), in which people judged the likelihood of three sequences of throws of a two-colored die. The majority of respondents (63%) violated the conjunction rule. The concern, however, has been raised that this task also suffers from ambiguities that participants need to resolve, and which may account for the violations of the conjunction rule (Macdonald, 1986).

Admittedly, semantic and pragmatic ambiguity can explain the discrepancy in performance between the theory tasks and the Linda task in our study, but it leaves the low performance in the maternity-ward task

DOI: 10.1002/bdm

unexplained. One possible interpretation is that the intuitive repertoire of people's probabilistic reasoning competence does not include sampling distribution. This in fact is the conclusion of Sedlmeier and Gigerenzer (1997), who argue that people are endowed with the valid intuition that estimates and predictions based on larger samples tend to be more accurate. In contrast, people's intuition does not seem to extend to sampling distributions that are hard to grasp. They suggested that one reason is that ''frequency distributions are involved in everyday problems of estimation and prediction, whereas the rule that the variability of a sampling distribution decreases with increasing sample size seems to have only few applications in ordinary life'' (p. 46).

Real-world tasks may be more difficult

Still another possible explanation for the disparity in probabilistic numeracy is that the Linda and the maternity-ward problems may simply be more difficult than the four theory problems. This is likely to be the case for the maternity-ward problem that involves two sampling distributions and requires a judgment of their respective variance – not a trivial task (Sedlmeier & Gigerenzer, 1997). The Linda task, however, is not more complicated than the corresponding coin task. The coin task requires determining the probabilities first and then multiplying them; the Linda task requires deducing that the probability of each of two single events cannot exceed the probability of their conjunction. One may object to our conclusion, as one reviewer did, that the Linda task involves uncertain constituent probabilities, whereas those of the coin tasks are certain (i.e., 50% of getting heads vs. tails). Therefore, the Linda task can be deemed to be more difficult than the coin task, because combining uncertain probabilities is more demanding than combining certain probabilities. This objection assumes that the Linda task requires first deriving estimates of the uncertain constituent probabilities and then calculating the conjoint probability. This, however, was not required of our respondents, who were merely asked to choose which of two statements is more probable (see Table 1). All that is required is a sense of the two statements' relative likelihoods.

To conclude, our findings suggest an important addition to future investigations that probe citizens' probabilistic numeracy. The ability to reason statistically may not only be a function of education, and how uncertain information is represented (in terms of probability or frequencies; see the debate between Kahneman & Tversky, 1996, and Gigerenzer, 1996) but also whether a task represents a pure application of probability theory or an approximation of real-world situations. Consequently, one may investigate each statistical reasoning task using both representations in parallel, thus finding out how robust the performance gap between both kinds of tasks is and determining which of the possible causes we have discussed accounts for the gap. In addition, to reach an unbiased estimate of the public's probabilistic numeracy requires a representative sample of tasks.

Determinants of probabilistic numeracy

Beyond the nature of the task, we corroborated another determinant of citizens' probabilistic numeracy: education. People with a college or university degree are more likely to respond correctly to the theory problems (Table 3). The beneficial impact of education, however, disappeared in the Linda and maternity-ward problems (Table 4). The latter finding seems inconsistent with Stanovich and West's (1998) finding: in a study involving students, they found that mean SAT score of those who violated the conjunction rule in the Linda problem was 82 points lower than of respondents whose answers conformed to the rule. Although within each of the three education brackets in our representative sample, a positive association between cognitive ability and performance may exist, we did not find such a positive link across levels of education (Table 4). Finally, we did find a link, though small in terms of effect size, between exposure to games of chance and probabilistic numeracy. Those with exposure were more likely to reason appropriately in the theory problems, and independent of education, males appeared to perform slightly better than females.

DOI: 10.1002/bdm

CONCLUSION

At the beginning of the twentieth century, H. G. Wells is reported to have predicted that statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write. One can safely say that this day has arrived. Every day, citizens are faced with a baffling array of percentages and probabilities – from probabilities of precipitation (Gigerenzer, Hertwig, van den Broek, Fasolo, & Katsikopoulos, 2005) and probabilities of side effects of medication to the risks of household accidents (Hertwig, Pachur, & Kurzenhäuser, 2005). To find out more about the degree of the citizenship's probabilistic numeracy and its determinants, we need to indeed investigate the public. We took a step in this direction, and found that education, exposure to games of chance, and sex are associated with probabilistic numeracy. In addition, our findings point to the critical role of the tasks designed to probe probabilistic numeracy.

ACKNOWLEDGEMENTS

We thank Andrea H. Meyer for his assistance in analyzing the data and Laura Wiles for editing the manuscript, and the Airport Casino Basel AG for their financial support in conducting the study.

REFERENCES

Adler, J. E. (1984). Abstraction is uncooperative. *Journal for the Theory of Social Behaviour*, *14*, 165–181.

Adler, J. E. (1991). An optimist's pessimism: Conversation and conjunction. *Posnan Studies in the Philosophy of the Sciences and Humanities*, *21*, 251–282.

Birnbaum, M. H. (1983). Base rates in Bayesian inference: Signal detection analysis of the cab problem. *American Journal of Psychology*, *96*, 85–94.

Black, W. C., Nease, R. F., & Tosteson, A. (1995). Perception of risk and screening effectiveness in women younger than 50 years of age. *Journal the National Cancer Institute*, *87*, 720–731.

Bondolfi, G., Osiek, C., & Ferrero, F. (2000). Prevalence estimates of pathological gambling in Switzerland. *Acta Psychiatrica Scandinavica*, *101*, 473–475.

Chase, V. M., Hertwig, R., & Gigerenzer, G. (1998). Visions of rationality. *Trends in Cognitive Science*, *2*, 206–214.

Dhami, M., Hertwig, R., & Hoffrage, U. (2004). The role of representative design in an ecological approach to cognition. *Psychological Bulletin*, *130*, 959–988.

Eddy, D. M. (1982). Probabilistic reasoning in clinical medicine: Problems and opportunities. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 249–267). Cambridge, UK: Cambridge University Press.

Evans, J. St B. T., & Dusoir, A. E. (1977). Proportionality and sample size as factors in intuitive statistical judgement. *Acta Psychologica*, *41*, 129–137.

Fagerlin, A., Zikmund-Fisher, B. J., Ubel, P. A., Jankovic, A., Derry, H. A., & Smith, D. M. (2007). Measuring numeracy without a math test: Development of the subjective numeracy scale. *Medical Decision Making*, *27*, 672–680.

Freedman, D., Pisani, R., Purves, R., & Adhikari, A. (1991). *Statistics* (2nd ed.). New York: Norton.

Gigerenzer, G. (1996). On narrow norms and vague heuristics: A reply to Kahneman and Tversky. *Psychological Review*, *103*, 592–596.

Gigerenzer, G. (2001). Content-blind norms, no norms, or good norms? A reply to Vranas. *Cognition*, *81*, 93–103.

Gigerenzer, G. (2002). *Calculated risks: How to know when numbers deceive you*. New York: Simon & Schuster.

Gigerenzer, G., Hertwig, R., van den Broek, E., Fasolo, B., & Katsikopoulos, K. (2005). ''A 30% chance of rain tomorrow'': How does the public understand probabilistic weather forecasts? *Risk Analysis*, *25*, 623–629.

Gigerenzer, G., Swijtink, Z., Porter, T., Daston, L., Beatty, J., & Krüger, L. (1989). *The empire of chance: How probability changed science and everyday life*. Cambridge, UK: Cambridge University Press.

Gilovich, T., Griffin, D., & Kahneman, D. (2002). *Heuristics and biases: The psychology of intuitive judgment*. Cambridge, UK: Cambridge University Press.

Gould, S. J. (1992). *Bully for brontosaurus: Reflections in natural history.* New York: Norton.

DOI: 10.1002/bdm

Halpern, D. F., Benbow, C. P., Geary, D. C., Gur, C. R., Hyde, J. S., & Gernsbacher, M. A. (2007). The science of sex differences in science and mathematics. *Psychological Science in the Public Interest*, *8*, 1–51.
Henrich, J. (2001). Challenges for everyone: Real people, deception, one-shot games, social learning, and computers. *Behavioral and Brain Sciences*, *24*, 414–415.
Hertwig, R., Benz, B., & Krauss, S. (2008). The conjunction fallacy and the meanings of ''and''. Manuscript submitted for publication.
Hertwig, R., & Gigerenzer, G. (1999). The conjunction fallacy revisited: How intelligent inferences look like reasoning errors. *Journal of Behavioral Decision Making*, *12*, 275–305.
Hertwig, R., Pachur, T., & Kurzenhäuser, S. (2005). Judgments of risk frequencies: Tests of possible cognitive mechanisms. *Journal of Experimental Psychology: Learning, Memory and Cognition, 35*, 621–642.
Hilton, D. J. (1995). The social context of reasoning: Conversational inference and rational judgment. *Psychological Bulletin*, *118*, 248–271.
Huff, D. (1954/1993). *How to lie with statistics*. New York: Norton.
Kahneman, D., Slovic, P., & Tversky, A. (Eds.), (1982). *Judgment under uncertainty: Heuristics and biases*. Cambridge: Cambridge University Press.
Kahneman, D., & Tversky, A. (1996). On the reality of cognitive illusions. *Psychological Review*, *103*, 582–591.
Lesieur, H. R., & Blume, S. B. (1993). Revising the South Oaks Gambling Screen in different settings. *Journal of Gambling Studies*, *9*, 213–223.
Lipkus, I. M., Samsa, G., & Rimer, B. K. (2001). General performance on a numeracy scale among highly educated sample. *Medical Decision Making*, *21*, 37–44.
Macdonald, R. R. (1986). Credible conceptions and implausible probabilities. *British Journal of Mathematical and Statistical Psychology, 39*, 15–27.
Mellers, B., Hertwig, R., & Kahneman, D. (2001). Do frequency representations eliminate conjunction effects? *Psychological Science*, *12*, 269–275.
Murphy, A. H., & Winkler, R. L. (1984). Probability forecasting in meteorology. *Journal of the American Statistical Association*, *79*, 489–500.
Paulos, J. A. (2001). *Innumeracy: Mathematical illiteracy and its consequences*. New York: Hill and Wang.
Peters, E., Västfjäll, D., Slovic, P., Mertz, C. K., Mazzocco, K., & Dickert, S. (2006). Numeracy and decision making. *Psychological Science*, *17*, 407–413.
Peterson, C. R., & Beach, L. R. (1967). Man as an intuitive statistician. *Psychological Bulletin*, *68*, 29–46.
Politzer, G., & Noveck, I. A. (1991). Are conjunction rule violations the result of conversational rule violations? *Journal of Psycholinguistic Research*, *20*, 83–103.
Rosenthal, R., & Rosnow, R. L. (1991). *Essentials of behavioral research: Methods and data analysis* (2nd ed.). Boston: McGraw Hill.
Schwartz, L. M., Woloshin, S., Black, W., & Welch, G. (1997). The role of numeracy in understanding the benefit of screening mammography. *Annals of Internal Medicine*, *127*, 966–972.
Sedlmeier, P., & Gigerenzer, G. (1997). Intuitions about sample size: The empirical law of large numbers. *Journal of Behavioral Decision Making*, *10*, 33–51.
Slovic, P., Fischhoff, B., & Lichtenstein, S. (1976). *Cognitive processes and societal risk taking* In J. S. Carroll, & J. W. Payne (Eds.), *Cognition and Social Behavior.* (pp. 165–184). Mahwah, NJ: Erlbaum.
Stanovich, K. E., & West, R. F. (1998). Individual differences in framing and conjunction effects. *Thinking and Reasoning*, *4*, 289–317.
Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Sciences*, *23*, 645–726.
Tentori, K., Bonini, N., & Osherson, D. (2004). The conjunction fallacy: A misunderstanding about conjunction? *Cognitive Science*, *28*, 467–477.
Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*, 1124–1131.
Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, *90*, 293–315.
Vranas, P. M. (2000). Gigerenzer's normative critique of Kahneman and Tversky. *Cognition*, *76*, 179–193.
Zangerl, M. A., Munsch, S., Meyer, A. H., & Margraf, J. (2008). Prevalence and risk of pathological gambling in Switzerland after the legalization of games of chance and casinos in 2000. Manuscript submitted for publication.
Zikmund-Fisher, B. J., Smith, D. M., Ubel, P. A., & Fagerlin, A. (2007). Validation of the subjective numeracy scale: Effects of low numeracy on comprehension of risk communications and utility elicitations. *Medical Decision Making*, *27*, 663–671.

DOI: 10.1002/bdm

Authors' biographies:

Ralph Hertwig is Professor of Cognitive and Decision Sciences in the Department of Psychology at the University of Basel, Switzerland. His research focuses on models of bounded rationality, social intelligence, and methodology of the social sciences.

Monika Andrea Zangerl is a doctoral student of clinical psychology and psychotherapy in the Department of Psychology at the University of Basel, Switzerland. Her research focuses on epidemiology of pathological gambling in Switzerland.

Esther Biedert is a clinical psychologist in the Department of Clinical Psychology and Psychotherapy of the Department of Psychology at the University Basel, Switzerland. She leads the Center of Psychotherapy for Adults and her research focuses on evidence-based psychotherapy.

Jürgen Margraf is Professor of Clinical Psychology and Psychotherapy and Dean of the Department of Psychology at the University of Basel, Switzerland. His research interests include the etiology and treatment of mental disorders with a special focus on anxiety-related problems.

Authors' addresses:

Ralph Hertwig, **Monika Andrea Zangerl**, **Esther Biedert**, and **Jürgen Margraf**, Department of Psychology, University of Basel, Missionsstrasse 60-62a, 4051 Basel, Switzerland.

DOI: 10.1002/bdm