Available online at www.sciencedirect.com
ScienceDirect
Cognitive Psychology 57 (2008) 293–334
Cognitive Psychology
www.elsevier.com/locate/cogpsych

The allure of equality: Uniformity in probabilistic and statistical judgment ☆

Ruma Falk [a,*], Avital Lann [b]

[a] *Department of Psychology and School of Education, The Hebrew University of Jerusalem, 3 Guatemal Street, Apartment 718, 96704 Jerusalem, Israel*
[b] *Faculty of Social Sciences, The Hebrew University of Jerusalem, Israel*

Available online 5 May 2008

Abstract

Uniformity, that is, equiprobability of all available options is central as a theoretical presupposition and as a computational tool in probability theory. It is justified only when applied to an appropriate sample space. In five studies, we posed diversified problems that called for *unequal* probabilities or weights to be assigned to the given units. The predominant response was choice of *equal* probabilities and weights. Many participants failed the task of partitioning the possibilities into elements that justify uniformity. The uniformity fallacy proved compelling and robust across varied content areas, tasks, and cases in which the correct weights should either have been directly or inversely proportional to their respective values. Debiasing measures included presenting individualized and visual data and asking for extreme comparisons. The preference of uniformity obtains across several contexts. It seems to serve as an anchor also in mathematical and social judgments. People's pervasive partiality for uniformity is explained as a quest for fairness and symmetry, and possibly in terms of expediency.

Keywords: Uniformity heuristic; Partitioning the space; Weighted mean; Self-weighting; Inverse weighting

☆ This study was supported by the Sturman Center for Human Development, The Hebrew University, Jerusalem, Israel. We thank Raphael Falk for his continued interest and dedicated help in all the stages of this project.
* Corresponding author. Fax: +972 2 6426016.
E-mail addresses: rfalk@cc.huji.ac.il (R. Falk), avital.a8@gmail.com (A. Lann).

0010-0285/$ - see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.cogpsych.2008.02.002

1. Introduction

How probabilities are intuitively assigned to events has always been the focus of the judgment-under-uncertainty school of research. In statistical language, the problem is that of the frequencies, or weights, to be attached to the various attributes or values. Evidently, the simplest method to fall back on is to embrace *equality*. Distributing the probabilities equally over the available options is the easiest, and it seems to be fair by being indifferent and doing justice to all contenders. This basic strategy, which we label the *uniformity heuristic*, has been described in historical accounts of the emergence of probability theory and statistics, and it is being mentioned in contemporary studies of psychologists, statisticians, and educators. However, relative to many specific heuristics and biases that have been studied in recent decades, this fundamental tendency has not received enough attention. Our purpose is to systematically explore the workings and manifestations of the uniformity construct and its degree of compellingness and resistance to contradictory evidence.

Carnap (1953, p. 132) presented the principle of indifference (as phrased by Jeffreys in 1939): "If there is no reason to believe one hypothesis rather than another, the probabilities are equal." The proviso in the first half of the sentence is critical. We conjecture that people sometimes apply the principle even when the proviso does not hold, and this is what we set about to examine in the literature review as well as in the experiments. The experimental tasks ask, either directly about event probabilities, or indirectly about means of distributions. In addition, several experimental manipulations investigate which conditions favor or check the reliance on uniformity. Finally, we have a look at the extent of generality of the uniformity tendency and at possible reasons for its supremacy.

1.1. Early roots of uniformity assumptions

Historically, according to Hacking (1975, chap. 14), the epistemological conception of equal probabilities goes back perhaps as early as to Aristotle. Leibniz defined probability as the ratio of favorable cases to the total number of *equally possible* cases in 1678, and so did Laplace around the end of the 18th century. The definition persisted in full vigor also a century after Laplace and it is still viable. According to Gigerenzer et al. (1989, p. 31 and 167), the heart of Laplace's interpretation of probability was viewing all cases with respect to which we are ignorant in the same way as equally possible. This enabled computing the probability of any given event. The uniformity assumption has been known throughout the years under various names, such as the *principle of indifference*, *principle of insufficient reason*, or *equal ignorance*. See Fine (1973, chap. 5), Keynes (1943, chap. 7), Stigler (1986, chap. 3) and Zabell (1988) on the history and philosophy of uniformity arguments.

1.2. Contemporary manifestations of the uniformity belief

In three notorious probability teasers—the *three cards*, the *three prisoners*, and the *three doors* (Monty's dilemma)—three a priori equally likely alternatives are presented. Then some datum (observation), which rules out one of the three, while lending unequal probabilities to the remaining two possibilities, is provided. The puzzles ask either to compare the chances of these two options or to assess the probability of one of them. Typically, most solvers keep adhering to uniformity, insisting also on posterior equiprobability of the remaining two possibilities (e.g., Bar-Hillel & Falk, 1982; Falk, 1992; Krauss & Wang,

2003; Nickerson, 1996, 2004; Shimojo & Ichikawa, 1989). In all three cases, the conditional probabilities (likelihoods) of the observation under the two remaining alternatives are different. Hence, *there is sufficient reason* for deeming these two alternatives *not* equally likely anymore. Evidently, the inventors of these puzzles counted on people's spontaneous proclivity to endorse equiprobability when creating these problems that fail their respondents.

The compellingness of the equiprobability belief was further revealed by the vehement objections of many lay readers as well as mathematicians and academics to vos Savant's (1990, 1991) correct solution of Monty's dilemma. Their support of the equality of probabilities of the two remaining options was decisive and confident; their uniformity belief seemed stable and indestructible (Falk, 1992). Furthermore, several (separate) debiasing attempts made by Krauss and Wang (2003), who introduced sensible changes in the formulation of Monty's dilemma, failed to achieve that end (only a combination of several such manipulations had the synergistic effect of correcting participants' perspective). Inappropriate uniformity responses are apparently hard to extinguish.

An extreme misapplication of uniformity occurs when it is applied to an *infinite* sample space. This is wrong, because the sum of all the probabilities in a discrete probability distribution, and the total area under the probability-density curve in the continuous case, should be equal to one, whereas for an unbounded and uniform distribution these totals would be infinite. Presuming constancy of probabilities, or densities, over an infinite sample space may result in contradictory consequences and acute paradoxes. Falk and Samuel-Cahn (2001) and Portnoy (1994) analyzed one problem of Lewis Carroll (1895/1958, problem 58) where he started by sampling three points *at random* (i.e., uniformly) on an infinite plane. One had to compute the probability that the triangle formed by those three points is obtuse. Carroll's answer was .64. Falk and Samuel-Cahn showed that reliance on the same self-contradictory assumptions may result also in a probability of 1 and even in a "probability" of 1.5, and Portnoy obtained a probability of .82 for the same event. These discrepant results were based on computations relating to "the triangle." However, presupposing the existence of such a triangle is illegitimate. This triangle was born in sin since it could *not* have been constructed, in the first place, by random selection of points from an infinite sample space. No less disturbing absurdities arise in the so-called *exchange paradox* as a result of assuming unlimited equal ignorance (e.g., Christensen & Utts, 1992; Nickerson & Falk, 2006).

Outside the realm of probability puzzles there are some direct research results showing a similar tendency. Tune's (1964) classic review of response preferences listed many probability-learning experiments in which participants had initially held an *equiprobable* hypothesis. Albert's (2003) students assumed that experimental outcomes would be equally likely even when the assumption was inappropriate. They chose the probability 1/2 for binary events *because there were only two possibilities*. Such statements abound in daily discourse and in the media (see Bruine de Bruin, Fischhoff, Millstein, & Halpern-Felsher, 2000, Table 1).

Fox and Rottenstreich (2003) found that participants' likelihood judgments were biased toward assigning equal credence to all mutually exclusive events considered by the judge. When the number of these events was changed by manipulating the problems' formulation, the participants kept distributing the probabilities equally over the set of events that they were led to consider. Fox and Clemen (2005) and See, Fox, and Rottenstreich (2006) confirmed the same tendency. Johnson-Laird, Legrenzi, Girotto, Legrenzi, and Caverni

(1999) maintained that individuals construct mental models for the truth of various possibilities. They apply equiprobability to these models, so that the probability of an event depends on the proportion of mental models in which it occurs.

Fox and Rottenstreich (2003) listed several other well-documented psychological findings that might be interpreted as manifestations of preference of uniformity. Thus, people's repeatedly observed tendency to overestimate proportions smaller than .5 and underestimate proportions greater than .5 could have been affected by their being drawn to the 50–50 anchor. The authors viewed this tendency as an instance of a general cognitive strategy that appears also outside the domain of probabilistic judgment. Roch, Lane, Samuelson, Allison, and Dent (2000) showed, for example, that in the economic context of a resource sharing task, participants anchored on the *equality heuristic* as a first stage. Likewise, Benartzi and Thaler (2001) found that, in choosing investments, many people used simple rules of thumb. One such rule was the *diversification heuristic*, or its extreme form: $1/n$ heuristic. Harris and Joyce (1980) found in three experiments that a sizable minority (37%) of students, who had been asked to allocate shares of the final outcomes of a group effort as fairly as possible, suggested giving all participants the same outcome, regardless of differential contributions.

A distinct inappropriate 'blip' at 50% was observed by Fischhoff and Bruine de Bruin (1999) in the distribution of participants' estimates of the probabilities of relatively unlikely events. The authors interpreted the number 50 in this case, not as a quantitative estimate of probability, but as a translation of the common phrase *fifty–fifty* that is equivalent to *absolutely no idea*. Indeed, this phrase would not have been coined, if not for the strong psychological link between epistemic uncertainty and equal probabilities. Konold (1989) reported that some of his participants viewed a 50% probability not as a prediction of long-range relative frequency, but rather as an admission of total ignorance, meaning "I really don't know" (p. 68).

The most stable and consistent finding of the vast research on subjective randomness is people's tendency to identify chance in binary sequences (and grids) with an excess of alternations between the symbol types (Falk & Konold, 1997; Nickerson, 2002). Kahneman and Tversky (1972) explained it by *local representativeness*: A random binary sequence is expected to exhibit the equiprobability of the two symbols not only globally but also locally in shorter sequences. This precludes long and medium runs of identical outcomes and results in overalternations. It reflects a wish to see *local uniformity*, or equality of frequencies everywhere along the series (Kareev, 1992).

Reliance on equally likely outcomes is still the most powerful tool in the calculus of probability, as presented in textbooks. When uniformity can legitimately be assumed, the computation of the probability of any event, E, reduces to that of a *proportion*:

$$P(E) = \frac{\text{number of favorable cases}}{\text{total number of cases}} = \frac{\text{number of elements in } E}{\text{total number of elements}} = \frac{n(E)}{N}, \tag{1}$$

where *cases* and *elements* in Eq. (1) refer to *elementary outcomes* (sample points) of the statistical experiment under discussion (Feller, 1957, chap. 1).

1.3. Equiprobability as an expedient heuristic

The justifiability of assuming uniformity has been for long the subject of discussion by philosophers and statisticians (Carnap, 1953; Keynes, 1943). It has been contended that

this assumption is logically unavoidable and the only possible one. It has also been claimed that the equal probabilities are due to experience; von Mises (1928/1957, pp. 68–73) justified uniformity by means of relative frequencies in the long run. According to Feller (1957), associating probability 1/2 with either head or tail, when tossing a "good" coin, is a convention: "We preserve the model not merely for its logical simplicity, but essentially for its usefulness and applicability. In many applications it is sufficiently accurate to describe reality" (p. 19). Stigler's (1986) historical account also agreed that "doing this was more calculational expediency than metaphysical assumption" (p. 103). Two examples follow.

In the classic birthday problem, assuming that all birth dates are *equally likely* easily yields the result that the probability of a shared birthday for two or more people exceeds 1/2 when the size of the group reaches 23. Berresford (1980) compared this result with results based on *actual* births' distribution—for which the assumption of uniformity did not hold—and found that a group of 23 was still required to raise the probability of a shared birthday above one half. That probability is in fact minimal when the distribution of births is uniform (Zabell, 1988).

Empirical psychological research found that predicting numerical variables of interest (e.g., ratings) by using linear combinations in which all the predicting variables are *equally weighted* is superior to experts' clinical predictions (Dawes, 1979). Moreover, selecting weights from a rectangular distribution does no worse than proper (regression) models obtained via optimization by statistical criteria.

Presuming equal probabilities, or weights, can thus be construed as a heuristic rule that may work. Quite often it serves its purpose. Polya (1981) considered it a principle of plausible inference in problem solving that often enables forecasting a solution. However, as compellingly demonstrated by Tversky and Kahneman (1974), heuristics that are highly economical and usually effective sometimes lead to systematic and predictable errors. We will show that uniformity, if overused, or misapplied, may result in distinct fallacies.

1.4. The problem of the appropriate sample space

The determination of the set of options to which equiprobability should apply becomes a crucial issue for computing event probabilities. When uniformity is postulated, different finite sample spaces result in different numbers in the denominator of Eq. (1). The *partition* into possible outcomes, determines that number. There is, however, no simple answer to the question how to reasonably divide the space into equally likely events. This question had troubled the founders of probability theory. Stigler (1986, p. 103) regarded Laplace's suggestion that if the options under one specification were known not to be equally likely, one should respecify them, say, by subdividing the more likely cases, as fraught with inherent difficulties.

We coach our students from the beginning to construct uniform probability spaces for computing event probabilities as relative frequencies. When asked about the probabilities of the *numbers of sons* in a 2-children family, although the possible values are 0, 1, and 2, one should consider the 4-point sample space of *ordered pairs* of children—*BB*, *BG*, *GB*, *GG*—and then apply Eq. (1) for obtaining the required probabilities. One notorious historical example of *not* following the above advice is that of D'Alembert's error (Todhunter, 1865/2001, pp. 258–259). In 1754, D'Alembert obtained a probability of 2/3 for the event of at least one *H* in two tosses of a fair coin. He considered only the outcomes that

were necessary for determining whether the target event had or had not occurred, namely, *H*, *TH*, *TT*. Since the first two of these three events were favorable he got the answer 2/3 by relying on equality of the probabilities of his 3-point space. This error is a quintessential instance of presupposing uniformity over the wrong sample space.

Although in D'Alembert's case it is agreed that the 4-point uniform sample space of ordered pairs should have been considered (yielding the answer 3/4), in other cases the definition of the "right" sample space is not that simple. That decision should, as a rule, depend on some background information, which is not always available. Physicists have described *real* behaviors of particles, which can be modeled by different uniform spaces depending on latent microphysical factors (Basano & Ottonello, 1996; Feller, 1957, pp. 38–40).

Uniformity has thus to apply to the "correct" sample space, the determination of which is not always self-evident. The complex nature of this issue has been repeatedly demonstrated by showing the absurdities that may arise if one attributes equiprobability liberally to various conceivable sample spaces (Carnap, 1953, p. 132; Fox & Rottenstreich, 2003; Keynes, 1943, chap. 4; Nickerson, 2004, pp. 35–37 and 204–205; von Mises, 1928/1957, p. 77). Consider a variation of one of Gardner's (1982, pp. 107–108) examples: Suppose one only knows that a cube's side is between 2 and 10 cm long. Applying the principle of indifference to the range 2–10, the cube's side is as likely to be above or below 6 cm long. Therefore, its volume is equally likely to be either greater or less than 6^3=216 cc. However, if uniformity is applied to the cube's volume across the range 2^3–10^3, or 8–1000 cc, then this entails equal probabilities for the volume being greater or less than $1008/2 = 504$ cc, which is obviously a contradiction. This apparent paradox results from the equivocality concerning the variable to which uniformity should apply. Equiprobability is *not* invariant under simple transformations though it seems offhand that it should be (Kotz & Stroup, 1983, pp. 154–155).

Often there is no ultimate answer to the question of the definition of the sample space that could rightfully be considered uniform (Keren, 1984). This choice depends on nonmathematical context considerations, and it should best be decided by judicious consensus. We chose the simple problem of the three cards as our first experimental stimulus because it explicates the underlying procedure in terms that—when being read carefully—leave no doubt as to which sample space should be considered uniform. This could, however, either agree or clash with people's intuitions, and their responses may indicate which sample space they regard as uniform. Keren found that perception of the relevant sample space can be manipulated effectively by apt experimental instructions. The question was also investigated by Brase, Cosmides, and Tooby (1998) and Fox and Levav (2004). Gavanski and Hui (1992) showed that people's natural sample spaces might differ from those prescribed by probability theory, but people can be helped in accessing appropriate sample spaces. Fiedler (2000) stressed the point that latent properties of the environment with which the individual interacts determine the samples to which one relates. These might be biased and entail biased judgments.

1.5. Uniform versus differential weighting

The frequencies (or weights) in an empirical distribution depend on the precise procedure employed in collecting the data. Their estimation, as well as that of the mean value, must therefore rest on knowledge of the methods of ascertainment (Fisher, 1934; Lann &

Falk, 2006). When all the weights are equal, the weighted mean is the simple arithmetic mean of the values. The converse is not always true. For example, the weighted mean equals the arithmetic mean of a nonuniform symmetric distribution. If, on the other hand, greater (smaller) values are systematically weighted more heavily, the resultant weighted mean is greater (smaller) than the arithmetic mean.

People's intuitions concerning differential weights have been scantily investigated. Pollatsek, Lima, and Well (1981) asked students to combine the means of two groups of different sizes into an overall mean of the union group. Many students failed to weight the two given means by the group sizes (that had been explicitly given) and responded by the simple arithmetic mean. On the other hand, Levin (1974) reported that participants' overall mean showed some sensitivity to correct weighting. Methodological differences can reconcile the ostensibly discrepant results: Many people may indeed disregard the differential weights (as reported by Pollatsek et al.), whereas some others weight their responses in the right direction so that group responses (as found by Levin) result in analysis of variance showing a main effect of sample size. Hawkins, Jolliffe, and Glickman (1992) maintained (like Pollatsek et al.) that students commonly fail in the assignation of different weights and they resort to the simple arithmetic mean.

The same fallacy occurs in averaging rates (i.e., means of dichotomous variables). People often disregard the different denominators of the given rates, and they average them via the simple arithmetic mean. Huck and Sandler (1984, p. 8) caution against this tendency. They note that the appropriate weights of several batting averages (for quantifying a baseball player's effectiveness) should reflect the different denominators, that is, different numbers of times at bat. Averaging the rates obtained in subgroups of different sizes by calculating their arithmetic mean, namely, weighting them all equally, does not generally yield the rate for the whole group. An intuitive erroneous belief that these two magnitudes are equal accounts for the difficulty that people demonstrably experience with Simpson's notorious paradox (Bickel, Hammel, & O'Connell, 1975; Simpson, 1951).

1.5.1. Self-weighted sampling

We call a procedure, in which the frequency (probability) of sampling each value in a given set is *equal* (*proportional*) to that value, *self-weighted sampling* (SWS). The arithmetic mean (expectation) of the sampled values is the *self-weighted mean* (SW) of the original set of values (Lann & Falk, 2005). Formally, when observations from a population of n (nonnegative) values, $x_1, x_2, \ldots, x_n$—whose *arithmetic mean* is $A = (x_1 + x_2 + \cdots + x_n)/n$— are obtained via SWS, the expected mean of the observations is:

$$SW = \frac{x_1^2 + x_2^2 + \cdots + x_n^2}{x_1 + x_2 + \cdots + x_n}. \tag{2}$$

Because in SW the weight attached to x_i is x_i—so that the larger the value the larger its weight—SW is generally greater than A. If and only if all the xs are equal, so are SW and A. It can easily be derived from the definitions of the arithmetic mean A, and the variance, σ^2, that

$$SW = A + \frac{\sigma^2}{A}. \tag{3}$$

This means that for a fixed A, the excess of SW over A is proportional to the variance of the population (Jenkins & Tuten, 1992; Patil, Rao, & Zelen, 1988).

Real-life examples of SWS are often encountered in medicine (van Dijk, 1997; Zelen & Feinleib, 1969), genetics (Fisher, 1934), demography (Keyfitz, 1985, chap. 10), and many other areas (Basano & Ottonello, 1996; Patil & Rao, 1978; Stein & Dattero, 1985). For example, many schools advertise a modest number as their average class size, yet, most students find themselves in quite larger classes (Hemenway, 1982). This happens because the mean class size for the *school* is generally less than the mean for the *student*. Each class contributes one addend to the calculation of the mean for the school, which equals A of the class sizes. However, each class size is added as many times as that size when listed for student by student in computing the mean per student, which equals SW of the class sizes. When individuals are organized in units, the average size per unit (family, class, and city) is not always the same as the average size per individual (child, student, and inhabitant). The latter is generally greater than the former (Smith, 1979). When a community comprises families of *different sizes*, the average child does not come from the average family (Jenkins & Tuten, 1992; Nickerson, 2004, pp. 150–151). Recording the size for each child in the community yields an upward-biased estimate of the mean per family whenever there is variability in family-sizes. Not only does it exclude childless families, but it overrepresents families with many children. Note that if the average family has three children, the average child has more than two siblings (Keyfitz, 1985, pp. 285–288).

Suppose cars travel along a freeway, each with a constant speed. Measuring the speeds of cars passing a fixed point, during a given time unit, in order to assess the average speed (A) of cars on the freeway is expected to give an upward-biased answer (SW instead of A). The faster cars will be sampled by this method more frequently than their share among the traveling cars, because in a given time unit the probability of recording a car's speed is proportional to the distance traversed by the car, which is proportional to its speed (Falk, Lann, & Zamir, 2005; Haight, 1963, pp. 114–116; Stein & Dattero, 1985). Likewise, if you get, at a random moment, to a bus stop in which three buses stop per hour—on average once every 20 min—the untutored intuition that your expected waiting time will be 10 min is generally faulty. The 10-min average wait would be correct only if all intervals between consecutive buses were exactly 20 min. If there is any variation in the interarrival times, the chances of arriving during a longer interval are greater, and the expected wait will be greater than 10 min (Lann & Falk, 2005). Your expected mean wait will be SW of all half intervals. Eq. (3) shows that it will be larger the greater the variance of the intervals' lengths. As flatly put by van Dijk (1997), "variation is the villain" (p. 28).

Considering people's shortcomings in taking differential weights into account, even when they are explicitly given (as in Pollatsek et al., 1981), it stands to reason to expect that self-weighting (as in the above examples) will a fortiori be harder to embrace. In many SWS cases, only the values are known and the proportionality of the probabilities of sampling them is not explicated. A few examples follow.

Stein and Dattero (1985) reported that students frequently suggest polling their classmates (or stopping people on the street) to ask about the number of children in their family in order to estimate the average size of a *family* in the population. And they fall prey to the same sampling bias in thinking that the average speed of cars on the freeway can be obtained by measuring the speeds of cars passing a fixed point. Jenkins and Tuten (1992) related that their colleagues believed that if the average number of children per family is three, then the average child has two siblings. They described also some distortions of historians and social scientists that had relied on the average size for the unit when the average for the individual should have been employed and vice versa. For

example, statistics indicated that in 1850 most American slaveholders owned few slaves, whereas in fact most slaves lived in large peer groups. Likewise, psychological studies of children's intelligence erred in drawing conclusions about families without acknowledging that large families had been overrepresented by sampling children (Bytheway, 1974; Smith, 1979).

Suppose all *men* in a large group write down the number of boys (including themselves) and the number of girls in their family. We then add the reported numbers of each kind and get the two respective sums B and G. Rao (1977) reported that people believe that the expected proportion $B/(B+G)$, is one half. This is wrong. The expected proportion is greater than 1/2, because all-daughter families are not represented in this survey, and the more boys there are in a family, the more likely they are to be represented in the group and then report a larger number of boys (Falk, 1982). This is a case of SWS.

Most of the evidence cited heretofore consisted of sporadic studies, anecdotal examples, introspection, and teachers' reports. Educators' experience, as well as introspection, is often a source of important insights. However, we endeavor to obtain orderly experimental evidence of people's performance in pertinent tasks that call for self-weighting.

1.5.2. Inversely weighted sampling

Unequal weights may be heavier for smaller values. In particular—in contrast to the situation in which each value is weighted in proportion to its magnitude—a value might be weighted in *inverse* relation to its size. In analogy to SWS, we refer to this case as *inversely weighted sampling* (IWS).

Suppose a car travels a distance of S km, from A to B, at 50 kph and back from B to A at 100 kph. We wish to know the car's average speed for the round trip ABA. This average should equal the total traversed distance ($2S$) divided by the total time of travel in the two directions. The time (in hours) of going from A to B is $S/50$, and from B to A it is $S/100$. The total distance can be obtained by adding the products of the two speeds and their respective durations of travel. Hence the desired average (in kph) is:

$$\frac{50\frac{S}{50}+100\frac{S}{100}}{\frac{S}{50}+\frac{S}{100}}=\frac{50\frac{1}{50}+100\frac{1}{100}}{\frac{1}{50}+\frac{1}{100}}=\frac{2}{\frac{1}{50}+\frac{1}{100}}.$$

This is a weighted mean of 50 and 100 in which each value is weighted by its reciprocal. The result is 66.7 kph, less than A of the speeds. This is an instance of IWS. If one samples a random moment from the total duration of the trip, for measuring the car's speed, the probability of sampling a given speed is directly related to the length of time that the car spends traveling at that speed, which is inversely related to the speed itself. In averaging the speeds, this probability works as a weight for its respective speed. This results in the harmonic mean of the speeds.

The *harmonic mean*, H, of n positive numbers, $x_1, x_2, \ldots, x_n$, is traditionally defined as the inverse of the arithmetic mean of the reciprocals of the n values. It is easy to see that H equals the weighted mean of the xs, where each x is weighted by its reciprocal:

$$H=\frac{n}{\frac{1}{x_1}+\frac{1}{x_2}+\cdots+\frac{1}{x_n}}=\frac{x_1\frac{1}{x_1}+x_2\frac{1}{x_2}+\cdots+x_n\frac{1}{x_n}}{\frac{1}{x_1}+\frac{1}{x_2}+\cdots+\frac{1}{x_n}}. \quad (4)$$

Comparing the weight of each x_i in the definition of SW—Eq. (2)—with its respective weight in the definition of H—Eq. (4)—and considering the equal weights in the arithmetic

mean[1] A, it becomes clear that $H \leqslant A \leqslant SW$ (Lann, 2008). When averaging only two values, a and b, H and SW are equally distant from A. The reciprocality between the corresponding weights in H and SW can be utilized for correcting sampling bias: If we are interested in the arithmetic mean of a population of positive values, but the values are sampled via SWS, then, to debias the inflated frequencies of larger observations, we can compute H of the sampled observations to obtain the desired answer (Patil et al., 1988; Stein & Dattero, 1985).

The harmonic mean pops up in diverse contexts. H has applications to wildlife populations and human demography (Keyfitz, 1968, pp. 382–384, 1985, pp. 335–337; Patil & Rao, 1978); it is of interest in traffic flow analysis (Falk et al., 2005; Haight, 1963, p. 22), and it has, since the days of Pythagoras, a musical connection. It is therefore important to know to what extent people recognize the need to weight values in converse relation to their size. There are a few suggestions that the arithmetic mean prevails as the chosen average, also under IWS. It has been repeatedly observed by teachers that, given two different speeds at which a car travels the same distance forth and back, students tend to compute A (rather than H) of these speeds as the car's average speed for the round trip (Gorodetsky, Hoz, & Vinner, 1986). Puzzlemakers like to include problems of this type in recreational-math collections. They can trust the illusion to work (Gardner, 1982, p. 142; Huck & Sandler, 1984, pp. 3 & 7).

1.6. Overview of the experiments

We conducted two sets of experiments. In the first (Study 1), a probabilistic experiment—that of the three cards—was described, and participants were asked about the conditional probability of a certain outcome. Uncritical application of the uniformity heuristic results in this case in an erroneous answer. However, uniformity may lead to the correct answer if the sample space is appropriately defined. Therefore, the research question may be viewed as finding out to which sample points participants ascribe equal probabilities. Study 1 comprised several variations of the same problem in order to find out whether different descriptions of the situation produce considerable shifts in the responses (Tversky & Kahneman, 1981), and which framing promotes a correct insight. Although a series of studies have established the relative superiority of formulating problems in frequentist rather than probabilistic terms (Cosmides & Tooby, 1996; Gigerenzer & Hoffrage, 1995), we limited this inquiry to questions about probability. The three-cards problem is *prototypically probabilistic*; a frequentist formulation would render it rather pointless, because it concerns a unique event.

In the second class of experiments (Studies 2–5), the descriptions of the statistical experiment implied assigning different weights to several given values. Out of the unlimited repertoire of nonuniform distributions, we selected two diametrically opposed ones as our targets: In one case, the weights had to be equal (or proportional) to their respective values (under SWS), and in the other, they had to be inversely proportional to the values (under IWS). Participants' subjective weights were to be inferred from their estimates or comparisons of means of distributions. Several framings of the problems were devised in an

[1] To be consistent with the notation SW, the *uniformly* weighted arithmetic mean and the *inversely* weighted harmonic mean could have been denoted UW and IW, respectively. However, because the terms *arithmetic* and *harmonic* are widely used, we retain the symbols A and H.

attempt to present increasingly transparent formulations with respect to the correct weighting.

In all the experimental forms, the givens and the underlying background were described unequivocally. This should have left no doubt as to the sampling procedures involved, unless participants project their own preconceptions on the process.

1.6.1. Participants

The participants were *statistically naive* students of the Hebrew University from diverse disciplines, in their first year of study. Altogether, we recruited more than 1200 students by entering classes in three successive academic years. The students participated voluntarily (they got credit when required by their departments). This large sample included students of psychology, education, social work, economics, law, business administration, political science, international relations, biology, and pharmacology. The participants showed interest and were eager to answer correctly. Many held lengthy discussions with us after handing in their forms. Hence, potential mistakes cannot be attributed to laziness or carelessness, but rather to faulty intuitions.

1.6.2. Procedure

The experiments took place in a class setting. The task was always to respond in writing to (Hebrew) forms. The total number of different forms was 32. About half the participants got two forms, and the rest got only one. A total of 1915 regular filled forms was collected (about 3% were disqualified because of incomplete or indistinct responses). Whenever two forms were paired, they belonged to different studies and never to analogous conditions, so that there could be no conceivable interplay between them. The order between two such forms was reversed for half the participants. Some forms appeared in alternative variants that differed only in the given numbers or in the order between multiple answers; their allocation was counterbalanced.

Each of the forms appeared on one page that started with the instructions to read the problem carefully and answer intuitively. We repeated these instructions orally, encouraging the students to rely on their common sense (even without carrying out computations). Respondents had either to circle one out of a given set of (numerical or verbal) options, or to provide a number. Space was left for an optional short written explanation. The experiment lasted about 15–20 min. Then we either briefed the class by explaining the correct answers, or handed out feedback sheets.

1.6.3. Analyses of results

In all the tables of results, boldface numbers designate the percentage of the modal choice in the column, and shaded percentages are those of the correct answer. The percentages in the tables of the results speak for themselves. They are often unequivocal concerning the use of the uniformity heuristic. No tests of significance are necessary to "prove" similarities or differences. Because of the misleading interpretation of a "significant" result that is too often believed to mean that H_0 has been rendered improbable (Cohen, 1994; Falk & Greenbaum, 1995; Gigerenzer & Murray, 1987, pp. 24–25), we rely essentially in drawing our conclusions on persistent *replications* based on large numbers. The 32 forms, in the five studies, investigate broadly the same question in many variations of context and task. The conclusions rest on studious evaluation of the obtained rates and on the extent of their consistency.

As often happens, some participants neglected to complete the verbal justifications, and some explanations were indistinct. Still, many were to the point, and they shed additional light on the quantitative analyses. We quote a few examples.

2. Study 1—The three-cards problem

2.1. Rationale

The problem was originated by Joseph Bertrand in 1889 and had since been widely circulated in several variations. One is told that three cards—the first, red on both sides (RR); the second green on both sides (GG); and the third, red on one side and green on the other one (RG)—are well shuffled. Then somebody blindly draws one card and, still blindly, puts it on the table. We observe a red face up (this observation is denoted r), and the question is what is the (conditional) probability that the bottom side of this card is also red.

Using the above notations, one is asked about $P(RR|r)$. The authors of this teaser set a trap for the respondents: Those who take into account the conditioning event, r, and understand that the card on the table cannot be GG, usually figure out that of the two remaining options, RR and RG, one has red and one has green on the other side—hence they give the mistaken answer 1/2 (Bar-Hillel & Falk, 1982; Brase et al., 1998; Fox & Levav, 2004). However, although RR and RG were a priori equally likely to be the ones on the table, they are *not* anymore so after observing r. The correct answer is $P(RR|r) = 2/3$, because the double-red card (RR) is twice as likely to yield a red face up (it is a certainty) than is the RG card, whose chances of showing either red or green as the upper face are equal (this argument can be formalized by Bayes' theorem).

The answer 2/3 might be better justified by noting that the description of the procedure guarantees equal probabilities of appearing as an upper face on the table to all *six sides*. The question of the probability of red on the bottom side reduces to whether the observed red side was a singleton or a "doubleton" (one of a pair). Two of the three *equally probable red sides* are doubletons with red also on the other side, only one is a singleton with green on the reverse side. This reasoning relies on uniformity as well, but it applies to the sample space of the sides rather than to that of the cards.

The three cards can be viewed as a core problem representing many of its kind. The advantage of using this problem as an experimental tool lies in its simplicity. Unlike Monty's dilemma and the three prisoners, where implicit assumptions have to be spelled out to disambiguate the situation (Falk, 1992; Nickerson, 1996), the story of the three cards is unequivocal. Furthermore, whereas the solutions of Monty and the prisoners are often rather complex—using tree diagrams, hierarchic tables, or Bayes' formula (Krauss & Wang, 2003; Shimojo & Ichikawa, 1989)—the three-cards' resolution can easily be achieved by resorting to the uniform sample space of the six sides.

It remains to be seen whether participants would intuitively regard the sides, rather than the cards, as the elements of their uniform sample space, and, if they do, under what conditions. For this end, we tried several variations of the problem that seem more conducive to the required insight. One version was meant to confront the respondent with the contradiction in assigning probability 1/2 to same color on the reverse side independently of whether the obverse is red or green, and at the same time assigning probability 2/3 to the event that a card has identical colors on both sides. The other versions attempted to direct solvers' attention to sides by separating them and by lending them

more individuality, up to "humanizing" them. Separation and individuation manipulations were described by Brase et al. (1998) and Fox and Levav (2004); their outcomes will be compared with those of our attempts.[2]

The answers 1/2 versus 2/3 distinguish between applying equiprobability to the cards and to the sides—provided that one has correctly deduced from the observation *r* that the *GG* card is ruled out. There are, however, grounds for concern that participants might ignore the given information, *r*, and respond, instead, with the prior probability (base rate) of the event that both sides of the drawn card are red. In that case, their answer will be 1/3. This is the probability of the *intersection* of the events *upper side is red* and *bottom side is red*. As teachers of many years, we have repeatedly encountered students who had been asked about $P(A|B)$, but for some reason or other "did not believe us" that B had occurred and computed $P(A \cap B)$, namely, the probability that *both* A and B will occur. Beyth-Marom (1977) established that participants often confuse $P(A|B)$ with $P(A \cap B)$. Falk, Lipson, and Konold (1994) and Teigen and Keren (2007) described situations in which participants, who were given the prior probability of a target event, ignored subsequent relevant information and insisted on adhering to the given base rates. Contrary to many studies that had documented base-rate neglect (e.g., Tversky & Kahneman, 1982), they found a "reverse base-rate fallacy". Such responses incur loss of some information, because our purpose was to compare the rates of the answers 1/2 and 2/3 among those who ruled out the *GG* card. In subsequent studies, we avoided confounding our intended examination with the issue of dealing with conditional probabilities.

2.2. Methods

Each of 486 participants responded to one of seven forms that presented variations of the three-cards puzzle. The versions were all mathematically analogous. They differed in details of their cover story. Participants were asked to circle one probability out of the following scale of given options:

0 1/6 1/5 1/4 1/3 1/2 2/3 3/4 4/5 5/6 1

The correct answer was always 2/3, whereas the answer 1/2 reflected the uniformity fallacy of adhering to equality of probabilities of the two remaining units (cards or their equivalents).

2.2.1. Traditional three-cards problem-stem

Three of the versions started with the same opening paragraph that described a statistical experiment, supposedly conducted by the experimenter:

> I hold three cards in my hand: One is red on both sides, another one is green on both sides, and a third one is red on one side and green on the other.
>
> I shuffle the three cards well, pick one card with closed eyes, and, still blindly, place it on the table.
>
> Now, I open my eyes, and then...

[2] Our manipulations were done independently. Due to oversight, we learned about these studies only after ours had been completed. Yet, we deem it important to report our results, since science is about the accumulation of information and testing by replication (Ross, 1985). Conducting replications is highly advocated and hardly ever implemented.

The continuation differed among the versions *Control*, *Standard*, and *Conflict manipulation*:

Control. To get an assessment of the extent to which participants understand the basic situation, we offered a control ending in which the specific color on the upper side of the drawn card was *not* mentioned:

> ...We all see *the color* of the upper side of this card.
> In your opinion, what is the probability that when I turn the card over, the bottom side will be of *the same color*?

Standard. This version ended with the classic formulation of the three-cards puzzle:

> ...We all see that the upper side of this card is *red*.
> In your opinion, what is the probability that when I turn the card over, the bottom side will be *red* as well?

Conflict manipulation. We reasoned that if participants will be confronted, in the same form, with the givens "red on the upper side," "green on the upper side," and "we see the color on the upper side" (without mentioning the identity of the observed color), and will be asked three times about the probability of finding the same on the bottom side, they will see the paradox in giving the answer 1/2 in the two former cases and 2/3 in the latter, and will realize that, irrespective of whether the upper color is red or green, the required probability is 2/3. The form comprised three parts. Each part was followed by the above scale of probabilities. The problem-stem, describing the experiment, was followed by:

> ...Three different outcomes are described below. Answer separately, in each case, according to the described outcome.
> (*a*) We all see that the upper side of this card is *red*.
> In your opinion, what is the probability that when I turn the card over, the bottom side will be *red* as well?
> (*b*) We all see that the upper side of this card is *green*.
> In your opinion, what is the probability that when I turn the card over, the bottom side will be *green* as well?
> (*c*) We all see *the color* of the upper side of this card.
> In your opinion, what is the probability that when I turn the card over, the bottom side will be of *the same color*?

Four alternative forms presented the three parts in four different permutations—*a b c*, *b a c*, *c a b*, and *c b a*—so as to counterbalance for possible order effects. Only participants who chose the same probability for the red and the green observations (parts *a* and *b*) were included (three participants who answered these parts differently were among the disqualified).

2.2.2. Individuation and separation manipulations

To draw participants' attention to sides, as possible sample points, we offered a variant of the problem in which each of the six *sides* was *labeled*. Because the physical tie between

the two sides of a card might obstruct regarding the sides as distinct units, we *separated* the elements and presented pairs of colored balls in urns, instead of sides of cards. This manipulation was intensified by *humanizing* the elements that became men and women in rooms, and more so, by *naming* these individuals.

Labeling sides. The formulation was largely as before, but for listing the six sides with an individual symbol attached to each:

> I hold three cards in my hand:
> One is red on both sides; one side is denoted R_1, and the other R_2. Another is green on both sides; one side is denoted G_1 and the other G_2. The third one is red on one side, denoted R_3, and green on the other one, denoted G_3.
> I shuffle the three cards well, pick one card with closed eyes, and, still blindly, place it on the table.
> Now, I open my eyes, and we all see that the upper side of this card is *red*. We do not know whether this upper side is R_1, R_2, or R_3.
> In your opinion, what is the probability that when I turn the card over, the bottom side will be *red* as well?

Labeling the sides is, in fact, a weak form of separation. In the next versions the separations were physical.

Balls in urns. Three opaque urns that look identical are in front of me. Each one contains two balls:
One urn contains *two red balls*.
Another urn contains *two green balls*.
And a third urn contains *one red and one green ball*.
I pick one urn with closed eyes, and, still blindly, draw one ball from this urn.
Now I open my eyes, and we all see that I drew a *red* ball.
In your opinion, what is the probability that when I draw the other ball out of the same urn, it will be *red* as well?

Human elements. Without loss of generality, men and women in rooms can replace red and green balls in urns. Humans might presumably be perceived as more unique, and thus play the role of elementary events:

> Three rooms, whose doors are closed, are in front of me. In each one are two persons:
> In one room there are *two men*.
> In another room there are *two women*.
> And in a third room there are *a man and a woman*.
> I don't know which pair is in which room.
> I happen to see by chance that a *man* gets out of one of the rooms.
> In your opinion, what is the probability that there is another *man* in the same room?

In an alternative form, a *woman* was observed getting out of a room and one was asked about the probability that there is another *woman* in that room.

Named humans. To impart more individuality to the six persons in the rooms, the same problem introduced each individual by his/her name. In two variants, either a man or a woman (unnamed in both cases) was observed getting out of one of the rooms.

2.3. Results and discussion

2.3.1. Overall outcomes

Table 1 summarizes the results. The correct answer 2/3 was the majority choice in *Control*, indicating that participants understood that there were two favorable cards among the three equiprobable cards. In all the other cases, the fallacious answer 1/2 was the most frequent. The next frequent error was 1/3, which is in fact the prior probability of the target event. All other erroneous answers were essentially blind guesses, with no comprehensible justifications. They are grouped together in the Other category.

Despite the majority of correct choices in *Control*, 25% of answers 1/2 seem unfeasible as just arithmetic errors. The explanations of this answer showed that, for the most part, uniformity accounted for the wrong answer. Some assumed a given upper color (as in *Standard*): "Suppose the card is red on one side, then the chances that it will be red on the other side are, as one can see from the givens, 1:2." Others resorted to downright uniformity: "If we pick a card of a certain color, we know that the other side could be *either the same or different* [italics added]".

The answer 1/2, selected in *Standard* by 75% of 107 participants, confirmed the expectation of the makers of this puzzle. That the remaining two *cards* stay *equiprobable* also after learning about the observation of an upper red face, was stated in no uncertain terms: "At first, when the three cards were in our hands, the probability of drawing a red card on both sides was 1/3. But, once it is clear that this is not the green card on both sides, the probability reduced to 1/2" (the mistaken "reduced" is fun). A few participants articulated the uniformity principle in its ultimate form – "because there are two colors"—without even mentioning cards or sides. The minority who chose the probability 1/3 simply ignored the observed conditioning event: "The chances of seeing red also on the other side are the random chances of choosing the red-red card from the beginning, that is, the probability is 1/3." The refusal to update the prior of the double-red card reflects a reverse base-rate fallacy (Falk et al., 1994).

Prima facie, the manipulations (starting with *Conflict*) appear beneficial: The rate of the wrong answer 1/2 decreased (relative to *Standard*). However, this decline was not system-

Table 1
Percentages of chosen probabilities in seven versions of the three-cards problem, in Study 1 ($N = 486$)

Probability	The version						
	Traditional problem-stem			Individuation and separation manipulations			
	Control	Standard	Conflict manipulation[a]	Labeling sides	Balls in urns	Human elements	Named humans
1/3	2.8	11.2	32.4	25.8	21.1	18.8	38.3
1/2	25.4	**74.8**	**57.7**	**51.5**	**55.3**	**47.9**	**40.4**
2/3	**63.4**	9.3	7.0	18.2	3.9	10.4	4.3
Other	8.4	4.7	2.8	4.5	19.6	23.0	16.9
n	71	107	71	66	76	48	47

Note. The probability chosen is that of the reverse side being red, given that the obverse was red, and of the equivalent events in all versions. The correct answer is always 2/3.

[a]The percents are those of the probability of red/green on the reverse side, given the obverse was red/green (parts *a* & *b*).

atically matched by a correspondent rise in the rate of the correct answer 2/3. In *Standard* there were about 16% of answers other than 1/2 or 2/3, whereas in response to all the manipulations, the rate of such answers was steadily at least 30%. Conceivably, the conflict, separation, and individuation hints did help respondents to view individual sides, balls, or persons as elementary events, but they had also to consider the organization of these elements within more inclusive units (cards, urns, rooms), and to take into account the effect of the conditioning observation. This additional cognitive load might have caused more participants either to dismiss the impact of the conditioning event or to grope in the dark for the required probability.

In *Conflict manipulation*, 63% of the participants were *inconsistent* in their responses to the cases of specified and unspecified obverse color (parts *a* and *b* vs. *c*). Only 4 of 71 participants gave the correct answer 2/3 to all three parts. The others were consistent in answering three times 1/2 or 1/3. Apparently, many participants were *not* strongly disturbed by giving identical probabilities for same color on the reverse side when red and green were given as the obverse colors, and a different probability for the same event, when the obverse color was not specified. The modal inconsistent triple answer 1/2, 1/2, 2/3 (47% of the inconsistent triplets) replicated the most popular answers to *Standard and* to *Control* even when confronted with them side by side, despite the contradiction. A typical explanation was *a*: "1/2, because one of the 2 cards with one red face has red also on the other side;" *b*: same as *a*, but for replacing "red" with "green;" *c*: "2/3, because there are 2 cards out of 3 with identical colors on the front and back sides." The second most frequent inconsistent triplet was 1/3, 1/3, 2/3. The majority answer to part *c*—the question of seeing the same unspecified color on the reverse side—was 2/3. However, 25% who circled 1/2 were not sloppily mistaken in dividing 2 by 3, they were swayed by their own answer to parts *a* and *b*. Sensing the need for consistency, they carried over to part *c* their wrong answers to parts *a* and *b*. This was betrayed by some participants who had first circled the probability 1/2 in parts *a* and *b* and 2/3 in part *c*, and then crossed the 2/3 and circled 1/2 instead. They explained their answer to part *c*: "This case is not different from parts *a* and *b*, although it seems so at first." The popular inconsistent 1/2, 1/2, 2/3 and consistent 1/2, 1/2, 1/2 indicate that either no conflict was felt due to the inconsistency, or, when a conflict was sensed, the wish for uniformity of the cards had the upper hand.

Labeling apparently did not help the 34 participants who still answered 1/2. They either ignored those tags altogether and explained their answer the standard way, or, despite noticing them, they kept relying on the equiprobability of the two remaining cards: "The green card (G_1G_2) is, in fact, out of the game. So two cards remain (R_1R_2 and G_3R_3), and the drawn card is one of them" Conversely, the labels had been often put to use by the participants who circled the probability 2/3: "If the side facing up is red, there are 3 possibilities for the identity of this card, out of them, 2 are that this is the double-red card, i.e., R_1 R_2. Two out of three are therefore the chances."

The modal answer was 1/2 also for *Balls in urns*: "It is possible that the third group was chosen, and then a green ball will be picked, and it is possible that the first group was chosen, and then a red ball will be picked. Two possibilities—50%." The percentage of the other wrong answers rose considerably at the expense of the correct answer, 2/3, that hit bottom (4%). Separating the elements within an urn apparently succeeded in directing participants' attention to the balls, but failed in helping them find the required conditional probability. Many solvers were unable to integrate the information about urns and balls with the impact of the conditioning event and were evidently at a loss.

In the combined results of the 95 participants who responded rather similarly to *Human elements* and *Named humans*, the percentage of the most frequent error, 1/2, dropped to 44%. However, here too there was no rise in the rate of the correct answer, and the percentage of other mistaken choices was maximal among all conditions, perhaps because of increased confusion concerning the two levels (persons and rooms) and their prior and posterior probabilities. The mistaken 1/2 was explained either by postulating equiprobability of the two remaining rooms: "It is certain a man was seen leaving the room, therefore we reduce the rooms to 2," or by outright uniformity applied to the six individuals without any consideration of their arrangement in rooms or of the given observation: "Half the people are men."

Brase et al. (1998) obtained results similar to ours and to Bar-Hillel and Falk's (1982) in an analogous problem in which candy canes replaced cards and two different types of edges replaced the two colors. Once these edges were presented as two different candies, or were physically broken, the percent of correct answers rose (unlike the case of our separation manipulations), but only to somewhat less than 25% or 35% (see the results of "Single Event" in Brase et al., Fig. 3). Their manipulation was thus only partially successful; it did not turn the tide. Our separation attempts managed only to lower the rate of the incorrect answer 1/2.

Fox and Levav (2004) numbered the six sides of the three cards. They also separated and humanized the sides by framing the problem in terms of three companies represented by two men, two women, or a man and a woman. As in our experiment, both their manipulations involved a decrease in the rate of the answer 1/2, and an increase in the answer 1/3 (Tables 5 and 8 in Fox & Levav). Their percents of the correct answer 2/3 were higher than ours, though they did not form a majority there either. Their relative advantage in eliciting the correct answer might be accounted for by differences in the problems' formulations. We adhered to the same target question also in *Labeling sides*: "What is the probability that when I turn the card over, the bottom side will be *red* as well?" and the same was done in *Human elements*, whereas Fox and Levav eased the target question in both cases. They asked: "Given that the side showing is red, what is the probability that it is side Red_1 or Red_2?" (p. 632).

The question about a conditional probability was apparently a drawback of the design because it interfered with examining the application of uniformity. However, if we confine our consideration only to the 339 participants who had chosen either the probability 1/2 or 2/3, the picture does not change. Barring control, 1/2 was throughout the majority choice of the respondents who took account of the conditioning event.

2.3.2. Unrefined scale of probabilities

All in all, participants were hardly willing to abandon the initially justified belief in the equiprobability of the given events after receiving information that reduced their number to two and undid their equiprobability. Many ruled out the double-green card (or its equivalents), showing that they took notice of the conditioning event. They were able to distinguish between possible and impossible events, but they failed to discern varying degrees of possibility between the remaining uncertain events and kept considering them equally probable. This phenomenon accords with Konold's (1989) report that some of his participants interpreted probability statements as definite predictions of what is going to happen. High (low) probabilities were taken as *certain* predictions of occurrence (non-occurrence) of the event in question. But when *uncertain*, they translated their state of knowledge into uniform probabilities that convey the message "I have no idea." These

people had in mind a *triple-edged probability scale* with no distinctions within the middle "don't know" category.

2.3.3. Conclusions

Though puzzles and paradoxes are often trifled as just piquancy, they can be invaluable in diagnosing our cognitive habits. The three-cards problem and its many isomorphs proved instrumental in pinpointing our tacit, untested assumptions and inbuilt bias toward uniformity. Being familiar with the compelling force of that problem, we have anticipated the prevalence of the answer 1/2. We were surprised, however, that a quarter of the participants erred in responding similarly even to *Control*. The finding that the apparently sensible corrective manipulations were to no avail came as a surprise as well. These results attest to the stubbornness of the uniformity heuristic. Switching people's attention from the units on which they had initially focused (cards) to elements of a more refined partition (sides) proved not easy to accomplish. As argued by Brase et al. (1998), "If whole objects are the natural unit of analysis for our statistical inference mechanisms, then people should have difficulty with any problem whose solution requires one to count arbitrary parsings of intact whole objects" (p. 9).

Yet, the results suggest that labeling and personification have a somewhat positive effect. To further explore whether individuation hints help in considering an appropriate partition, we employed them again in Study 2 without confounding weighting with conditionality. If the effect of individuation, as found by Fox and Levav (2004) and hinted by the results of this study, would replicate, this might validate its expediency. It would also confirm Nisbett and Ross' (1980, chap. 3) thesis that vivid information has greater impact on people's inferences.

The three cards, *GG*, *GR*, and *RR*, can also be identified by their numbers of red sides, namely, 0, 1, and 2. The three respective posterior probabilities—0, 1/3, and 2/3—are directly proportional to these numbers. Hence, the three-cards experiment represents a case of SWS. In Study 2, we have a look at people's responses to another situation in which the sampling procedure imparts self-weighting of the values.

3. Study 2—Self-weighting problems

3.1. Rationale

The procedure of collecting numerical data in this study (SWS) entailed weights (frequencies) equal to the values. In most of the problems, equal number of families with one, two, three, or four children were given. Participants were asked about the mean number of children in the family of *a child* in this population. Let the *number of children in a family* be referred to as the *family-size* (not including the parents). The distinction between the mean family-size *per family* and the mean *per child* is important. The former is the simple arithmetic mean, A ($=2\frac{1}{2}$), of the distribution of families, and the latter is the self-weighted mean, SW ($=3$), of that distribution. One can learn from participants' chosen mean how they intuitively weight the different family-sizes. A mean greater than A, indicates that the balance of the weights tends upwards.

In analogy to the question of relating to sides or to cards, respondents' mean per child may indicate which sample space they consider uniform: that of the families or that of the children. Based on suggestions in the literature (Jenkins & Tuten, 1992; Nickerson, 2004, pp. 150–151; Stein & Dattero, 1985), we hypothesized that many participants will consider

the families equiprobable and will answer the question of the mean per child by A of the family-sizes, that is, by the mean per family.

Modifications of the same problem included facilitating features: By asking both about mean family-size per family and per child, we tried to *raise* participants' *sensitivity* to the implications of the difference between the two sampling methods. In a stronger manipulation they had to *compare* the means obtained by recording family-size per family and per child. Participants were also required to mentally conduct a thought-experiment and *produce an example* of a typical sample of family-sizes recorded for the children prior to selecting their answer of the mean family-size per child. Envisaging the sampling process might encourage relating to children and realizing that a given family-size is multiply recorded according to its magnitude. Furthermore, we offered versions with *only one family of each of the four sizes*. These included variants in which we intensified the salience of individual children by *naming* or even *depicting* them. To amplify the effect, two other conditions presented *extreme* distributions comprising only two families with widely disparate numbers of children. A strategy of "extremization" is often effective in problem solving (e.g., Gardner, 1978, pp. 54–55; Polya, 1981, Vol. 1, p. 10). Finally, to find out whether participants are sensitive to the *effect of the variance* in family-sizes (Eq. (3)), we presented two pairs of families with the same arithmetic mean of their sizes; in one case the two sizes were close to each other and in the other they were widely discrepant. Participants were asked to compare the mean family-size per child in these two cases.

3.2. Methods

Thirteen forms dealt with problems concerning mean family-size. Altogether, 771 regular forms were collected. Our main dependent variable was participants' *assessed mean family-size per child*, denoted M_c.[3] The *assessed mean family-size per family* is denoted M_f. Participants' ability to distinguish between the two, and to realize that the former is generally greater than the latter, was the target of the investigation. In all the cases it was true that $M_c = SW > M_f = A$. One version concerned the *assessed mean number of siblings per child*, denoted M_{sib}. This should equal $SW - 1$ and be greater than $A - 1$.

The first eight versions asked participants to *assess mean family-size*—M_f in the first and M_c in the other seven. They differed in the given distribution of families and in the vividness of the description. In the first seven of these versions, the population consisted of equal number of families with one, two, three, or four children, and participants were asked to circle one mean out of the following scale of options:

$$1 \qquad 1\tfrac{1}{2} \qquad 2 \qquad 2\tfrac{1}{2} \qquad 3 \qquad 3\tfrac{1}{2} \qquad 4$$

The *correct answers* are $M_f = A = 2\frac{1}{2}$, and $M_c = SW = 3$. The other five versions, required *comparison of means*—four obtained by different methods from the same population, and the fifth by the same method from different populations.

[3] When asking about M_c, we clarified that the recorded child is *included* in the family-size, and that one should assume that all the numbers recorded for children are accurate.

3.2.1. Simple distribution

All four versions in this category started with the *same problem-stem*:

> In a certain community, all families have children (one or more).
> There are *equal numbers of families* with **one**, **two**, **three**, and **four** children (there are no families with more than four children).

The continuation varied among the forms.

Control—mean per family (M_f). To make sure that participants got the families' distribution and its arithmetic mean right, they were asked:

> A researcher addresses *each family* in the community and records its *number of children*. The researcher computes the *mean of the recorded numbers*. In your opinion, what is the value of this mean?

Standard—mean per child (M_c). A researcher addresses *each child* in the community and records the *number of children* in the child's family.
The researcher computes the *mean of the recorded numbers*. In your opinion, what is the value of this mean?

Sensitization—mean per child and per family (M_c and M_f). Participants were asked about both M_c *and* M_f in the same form. Researcher **A** addresses *each family*, as in *Control*, and Researcher **B** addresses *each child*, as in *Standard*. The above scale of multiple answers was given below each of the two questions. In two alternative forms the order of the two questions was reversed. We were interested only in the M_c responses. A few participants whose M_f was wrong were deleted because of failing the control test.

Imagine a sample (M_c). The question was as in *Standard*. However, before the question paragraph (after the problem-stem), participants were asked:

> Suppose you sample some children from this community, so that all the children in the community have equal chances to be sampled. You record the *number of children* in the family of every child in the sample.
> Please write in the allotted space a *representative example* of the numbers that may comprise the sample:

__ __ __ __ __ __ __ __ __ __ ... ...

This format enables participants who apprehend the self-weighting to produce the exemplary sample 1, 2, 2, 3, 3, 3, 4, 4, 4, 4; whereas if they embrace equal weights, they could repeat each of the four numbers three times.

3.2.2. Reduced distribution: one family of each kind

Four versions—with minor variations—were simplified by presenting only one family of each size.

Standard—mean per child (M_c). *All* the children of four families play in a community center. One family has **one** child, one has **two** children, one has **three** children, and one has **four** children.
A researcher addresses *each child* in the community center and records the *number of children* in the child's family. The researcher computes the *mean of the recorded numbers*. In your opinion, what is the value of this mean?

Labeling children (M_c). Repeating the previous formulation, we added the children's *names* immediately after mentioning their number in each family.

Depicting children (M_c). Every child in every family was represented by a sketch, with the child's name written above the figure. Otherwise, everything was the same as before.

Extreme distribution(s) (M_c). Only two families, with a wide gap between their sizes, were described. Two alternative forms of this version were composed. The wording was as in *Standard* of this section, except for the following changes: *Four families* were replaced by *two families*; the numbers of children in the two families were 3 and 15 in one form ($A = 9$; $SW = 13$), and 2 and 6 in the other form ($A = 4$; $SW = 5$). The scale of multiple answers for the first form was:

3 4 5 6 7 8 9 10 11 12 13 14 15

And for the second it was:

2 $2\frac{1}{2}$ 3 $3\frac{1}{2}$ 4 $4\frac{1}{2}$ 5 $5\frac{1}{2}$ 6

3.2.3. Comparing means

In four versions, participants were required to *compare* their M_c with the mean *per family* (M_f or A), either when the distribution of families was given or when it was not known. The former case included an extreme case of two widely disparate family-sizes, and the latter included comparison of the assessed mean number of siblings per child (M_{sib}) with $A - 1$. The fifth version required comparing M_cs of two distributions, each consisting of two families with the same A but with considerably different variabilities.

Simple distribution—mean per child versus mean per family (M_c vs. M_f). The problem-stem distribution was presented again. As in *Sensitization*, two procedures of data collection were described, but here they had to be compared with each other:

> Two researchers visit the community.
> Researcher **A** addresses *each family* and records its *number of children*,
> Researcher **B** addresses *each child* and records the *number of children* in the child's family.
> Each of the two researchers computes the *mean of the recorded numbers*. In your opinion, what will be the *relation between the means* obtained by the two researchers?
>
> 1. There will be no difference between the two means.
> 2. The mean of Researcher **A** will be *greater* than that of Researcher **B**.
> 3. The mean of Researcher **B** will be *greater* than that of Researcher **A**.

In two alternative forms, the multiple answers appeared in different orders.

Extreme distribution(s)—mean per child versus the arithmetic mean of family-sizes (M_c vs. A). Only two families with a considerable gap between their sizes as in the assessment task in *Extreme distribution(s)* (M_c) were presented:

> *All* the children of two families play in the yard. One family has **2** children and the other one has **6** children.
> A researcher addresses *each child* in the yard and records the *number of children* in the child's family.

> The researcher computes the *mean of the recorded numbers.*
> In your opinion, which of the following statements is correct?
>
> 1. The researcher's mean will equal 4.
> 2. The researcher's mean will be less than 4.
> 3. The researcher's mean will be greater than 4.

In another variant, the two family-sizes were 3 and 15 and M_c had to be compared with $A = 9$. Each variant appeared in two alternative forms, with changed order among the three multiple answers.

No distribution—mean per child versus mean per family (M_c vs. M_f). This version resembled the story of the two researchers in *Simple distribution*. However, the only information concerning the distribution of the families was that "There are families of *different* numbers of children in the community." Note that this suffices for deducing that $SW > A$, and hence M_c should be greater than M_f.

No distribution—mean number of siblings per child versus one less than the mean of family-sizes (M_{sib} vs. $A - 1$). Only the crucial information that *not* all family-sizes were identical was given:

> In a certain community, families have different numbers of children. The *mean number of children of a family* in the community is **3**. A researcher addresses *each child* in the community and records the child's *number of siblings* (brothers and sisters).
> The researcher computes the *mean of the recorded numbers* (i.e., the mean number of siblings of a child in the community).
> In your opinion, which of the following statements is correct?
>
> 1. The researcher's mean will be less than 2.
> 2. The researcher's mean will be equal to 2.
> 3. The researcher's mean will be greater than 2.

In three alternative forms, different orders between the possible answers were presented.

Two families—mean per child for different variances (M_c^L vs. M_c^S). The assessed mean family-size per child in the case of *large* variability between the two families is denoted M_c^L, and in the case of *small* variability it is denoted M_c^S. The distributions to be compared had the same arithmetic mean (A) of family-sizes:

> *All* the children of two families play in Yard **A**. One family has **8** children and the other has **10** children.
> Researcher **A** addresses *each child* in Yard **A** and records the *number of children* in the child's family.
>
> *All* the children of two other families play in Yard **B**. One family has **3** children and the other has **15** children.
> Researcher **B** addresses *each child* in Yard **B** and records the *number of children* in the child's family.
> Each researcher computes the *mean of his/her recorded numbers*. In your opinion, what will be the *relation between the means* obtained by the two researchers?

1. There will be no difference between the two means.
2. The mean of Researcher **A** will be *greater* than that of Researcher **B**.
3. The mean of Researcher **B** will be *greater* than that of Researcher **A**.

In a second variant, the numbers in Yard **A** were 3 and 5, and in Yard **B** they were 1 and 7. Each variant appeared with two different orders between the multiple answers.

3.3. Results and discussion

3.3.1. Overall outcomes

The outcomes of *Control* show that 29 of 32 participants correctly assessed $M_f = A = 2\frac{1}{2}$. The 91% correct answers confirm the simplicity of the case of the mean per family. Having established that participants largely equated M_f with A, we subsequently compared M_c responses with A of the given distribution without asking about M_f.

Table 2 presents the results of seven versions of assessing mean family-size per child (M_c), giving the percents of participants whose M_c equaled SW (the correct answer) and A (the uniformly weighted mean). However, applying a more lenient criterion of admissibility, any M_c greater than A could be considered adequate, because it indicates a tendency to weight greater values more heavily. The main finding was that, in the absence of powerful facilitating hints, participants failed to self-weight the given values. They weighted all the sizes uniformly and answered by A, the mean family-size per family, as in *Control*. As in Study 1, the uniformity fallacy prevailed in *Standard*. At the same time, it was reduced by apt formulation of the mean-per-child problem. Considering only one family of each size was, on the whole, helpful, especially when the children were individualized, and when there were only two families with an amplified gap between their sizes.

Choosing $A = 2\frac{1}{2}$, instead of $SW = 3$ in *Standard* of *Simple Distribution*, seemed self-evident: "If there is an equal number of families with 1, 2, 3, and 4, then the mean will be $(1 + 2 + 3 + 4):4 = 2\frac{1}{2}$." The correct choice was explained: "Every child counts himself and his siblings; and then we divide by the number of children that were interrogated: 1; 2, 2; 3, 3, 3; 4, 4, 4, 4." In *Sensitization*, participants who had circled A for both M_c and M_f justified that double choice by: "The answers to both parts are identical, because, in fact, this is the same question in different words." They were not sensitive to the implications of the different methods of data collection. Yet, the need for differential weighting was not beyond the grasp of 30% who had responded in the right direction: "The big families will give more big answers and the mean will increase."

In *Imagine a sample,* sorting the generated samples as indicative of either SWS or *uniformly weighted sampling* (UWS) was unequivocal for 51 out of the 69 participants: they split into 36 UWS and 15 SWS. The remaining 18 samples were sorted *undecided*. Of the participants whose sample was definitely categorized, 90% selected M_c *consistent* with their sample. The group's modal choice was still the uniformly weighted mean A. The undecided group had 72% of A answers. The increase of the rate of $M_c > A$ answers relative to *Standard* might be explained by a limited degree of success of the attempt to think of a sample. In a typical response, the sample consisted of three of each of the numbers 1, 2, 3, and 4 (in mixed order) and—consistent with UWS—$2\frac{1}{2}$ was the chosen M_c: "Because equal numbers of children from all types of families are included in the sample." Most of those who answered correctly produced a perfect, or close to perfect, self-weighted sample and based their explanation on it.

Table 2
Percentages of assessed mean family-size per child (M_c), according to its relation to means of the families' distribution in seven versions, in Study 2 ($N = 422$)

Participants' chosen mean per child (M_c)	The version						
	Simple distribution[a]			Reduced distribution: One family of each kind			
	Standard	Sensitization[b]	Imagine a sample	Standard[a]	Individuation[a]		Extreme distribution(s)[c]
					Labeling children	Depicting children	
$M_c < A$	10.0	12.8	5.8	9.3	5.6	0.0	1.8
$M_c = A$	**72.2**	**57.4**	**66.7**	**51.9**	37.0	43.4	18.2
$A < M_c < SW$	—	—	—	—	—	—	9.1
$M_c = SW$	13.3	21.3	20.3	29.6	**46.3**	**47.2**	**69.1**
$M_c > SW$	4.4	8.5	7.2	9.3	11.1	9.4	1.8
n	90	47	69	54	54	53	55

Note. In all versions families have different sizes. A and SW denote, respectively, the arithmetic mean and the self-weighted mean of the given distribution of families. M_c should always equal SW.

[a]Four family-sizes—1, 2, 3, and 4 (with equal frequencies)—were given. Their means are $A = 2\frac{1}{2}$, $SW = 3$. There was no value between A and SW in the scale of options.

[b]M_c responses were counted only for those whose M_f was right.

[c]Two alternative family-sizes—2, 6, or 3, 15—were given. Their means are $A = 4$, $SW = 5$ or $A = 9$, $SW = 13$, respectively.

Reducing the distribution to *One family of each kind* was not enough for tipping the balance in *Standard* in favor of the correct answer: "The total of families is 4, the total of children 10, therefore the mean is $2\frac{1}{2}$," However, the effect of *Individuation* was stronger here than in Study 1. The outcomes of *Labeling* and *Depicting* children showed that these measures helped participants to consider each child personally and weight the family-sizes accordingly. Some listed the children one by one, naming each child and specifying its family-size. The *Extreme distribution* of only two families swayed the responses toward weighting the family-sizes in the right direction. The common error of selecting A was reduced and the rates of answers greater than A peaked to 80%. The efficacy of the manipulation was revealed in an explanation of the choice $M_c = SW = 5$ (for families of sizes 2 and 6): "If there are 2 families in the community center—of 6 children and 2 children—then the mean will be $(6 \times 6 + 2 \times 2)/8 = 5$. I should note that at first glance I assumed that it was 4 children." Comments similar to this indicate that some learning took place while considering the extreme version.

Comparison is apparently more potent than sensitization. Table 3 presents the results of the four comparisons between M_c and A (or M_f). When participants were asked to judge which of two sampling procedures will yield a greater mean, they were apparently driven to consider the difference between the two methods. The rates of the mistaken $M_c = A$ dropped in comparison tasks relative to their analogous assessment tasks. There was also a rise in the correct answers ($M_c > M_f$ or $M_c > A$). Yet, we cannot know, in the case of correct comparisons, whether the participant's M_c was numerically accurate (equal to SW) or just greater than A. Some comparisons between M_c and M_f in *Simple distribution* and *No distribution* were well explained: "Since Researcher **B** addresses each child he'll probably count the same family several times." Still, the modal choice was $M_c = M_f$: "The families and the children will report the same number of children."

Coupling the amplified gap between the two families with a comparison task in *Extreme distribution(s)* proved highly effective in clarifying the distinction between self- and equal-weighting: "There are greater chances that the researcher will ask one of the 6 children of the second family. This means that the number 6 will appear more times than the number 2, and therefore the mean will be greater than 4 which is the mean of the ratio 1:1 between 2 and 6."

Table 3
Percentages of comparisons between assessed mean per child (M_c) and mean per family (M_f or A) in four versions, in Study 2 ($N = 257$)

Participants' choice	The version			
	Simple distribution M_c vs. M_f	Extreme distribution(s) M_c vs. A	No. distribution	
			M_c vs. M_f	M_{sib}[a] vs. $A - 1$
$M_c < M_f$ or $M_c < A$	15.2	3.9	8.5	5.4
$M_c = M_f$ or $M_c = A$	**44.3**	7.8	**46.5**	**83.9**
$M_c > M_f$ or $M_c > A$	40.5	**88.2**	45.1	10.7
n	79	51	71	56

Note. In all the versions families have different sizes. The assessed mean per family (M_f) should equal the arithmetic mean (A) of the families' distribution. The assessed mean per child (M_c) should equal the self-weighted mean (SW) of that distribution. The correct answer is always $M_c > M_f$ or $M_c > A$.

[a] M_{sib} = assessed number of siblings per child. The correct answer is $M_{sib} > A - 1$, which is equivalent to $M_c > A$.

Ostensibly, whoever understands that M_c should be greater than the mean per family, $A = 3$, can also see that the assessed mean number of siblings per child ($M_{sib} = M_c - 1$) should be greater than $A - 1 = 2$. However, most participants had apparently focused on the need to subtract one from the number of children in the family to get a child's number of siblings and thus circled 2 as their M_{sib}. Many regarded that choice self-evident: "If there are three children on the average, logic implies that each child has 2 siblings on the average."

Participants sensed the increase in SW with an increased variance in family-sizes. When comparing M_c of two families in distributions with *different variances* (not included in Table 3), 75% of 60 participants circled the category $M_c^L > M_c^S$ and only 18% chose $M_c^L = M_c^S$. Many participants mentioned only the bigger family of the more disparate pair: "There are more children in a family of 15 children, therefore the relative part of bigger numbers is greater." Only one participant who chose correctly wrote: "Because the dispersion of the children between the two families in Yard **B** is more extreme than in Yard **A.**" Contrarily, a participant who had opted for no difference noted: "The means are the same, only the standard deviations are different."

3.3.2. Conclusions

People's bias in weighting the given options equally was prominent in the standard tasks of assessing mean family-size *per child*, much as in the case of the three-cards problem. The rates of relying on (posterior) uniformity of cards or families were rather similar in the two studies (cf. Tables 1 and 2). Recall that the three cards can also be construed as a self-weighting problem, because cards were sampled via red sides, just as families were sampled via children. In both studies, participants were mostly oblivious of the need for self-weighting and they fell back on the simplest solution of equal weights, often without explicit awareness of their weighting decisions. Just as cards (not sides) were subjectively considered the focal units of analysis in Study 1, so were families (not children) primarily viewed as "whole objects" (Brase et al., 1998) in this study. It took specific manipulations to shift participants' attention to the more refined uniform sample space of children. Corrective effects of individuation, as hinted in Study 1 and found by Fox and Levav (2004), were found. Most expedient was the *amplification* (extremization) of the variability in family-sizes. The best performance in this study, close to 90% of correct judgments, was obtained for a *comparison* involving *extreme* gaps. We therefore employed a combination of comparison and extremization also in the next study, where the correct weights should be inversely related to the weighted values.

4. Study 3—Inverse-weighting problems

4.1. Rationale

Participants had to figure out the mean of two values by weighting them in converse relation to their size. Inverse weighting is exemplified when a car travels there and back along the same distance at different speeds and one has to find the *overall mean speed* for the round trip. We used this framework for studying people's weighting decisions under IWS. Weighting lower speeds more heavily decreases the mean below A. Specifically, weighting each speed exactly by its reciprocal yields H—see Eq. 4.

In Study 2, where SWS was due, many participants resorted to equal weights. It stands to reason to expect that a similar or even stronger tendency will be observed under IWS,

because one has to inverse each value prior to using it as a weight. Wilkening (1981) argued—in a developmental study about conceptions of velocity, time, and distance—that concepts involving direct relations are easier to understand than those involving inverse relations. Falk and Wilkening (1998) found that, *ceteris paribus*, children succeeded better in solving probability problems when the number to be figured out was in the numerator than when it was in the denominator of the odds in favor of the target event.

Considering the efficacy of combining comparison with extremization in the case of SWS, we included a version with an increased gap between the two speeds and asked participants to compare their assessed overall mean speed with the arithmetic mean of the two speeds. In addition, to extend the inquiry beyond the sphere of speed problems, we asked for an extreme comparison in another IWS problem concerning rates of success.

4.2. Methods

Five forms dealt with IWS. Altogether, we collected 210 regular responses. Participants' *assessed overall mean*, denoted M, should equal H of the two given numbers. Three of the five versions asked for assessments and two for extreme comparisons. Of the three assessment versions, two asked for M of the same two speeds and one specified the value of one speed as well as that of the overall mean speed (H of the two speeds) and asked for assessment of the second speed. The two comparisons pitted participants' assessed mean against the known value of A of two given numbers. In one case, these were speeds (of traversing the same distance), and in the other, they were probabilities of success per trial. Suppose one runs repeated, independent trials, such that the probability of success in each one is p. It can easily be ascertained that the expected number of trials until the first success (including that success) is $1/p$ (Keyfitz, 1985, pp. 335-336). If, now, two types of trials, with success probabilities p_1 and p_2, are each performed repeatedly until the first success, then the *overall expected proportion* of successes among the totality of trials of the two kinds is about $2/[(1/p_1) + (1/p_2)]$, that is, the *harmonic mean*, H, of p_1 and p_2. Comparing the assessed overall proportion, denoted P, with the arithmetic mean of the two proportions is also a case of comparing M with A, because the *proportion* of successes is the *mean* of the outcomes of all the trials when a success is encoded as one and a failure as zero.

4.2.1. Assessments

The two versions in which an assessment of the overall mean speed was required were analogous. The purpose of two mathematically equivalent questions was to make sure the problem is understood:

Standard 1—overall mean speed (M). The problem was worded so as to make clear that one is asked about the overall mean speed and not about the arithmetic mean of the two speeds:

> A man drives to work, in the morning, from city A to city B at a constant speed of 100 kph. On his way back he is more relaxed and he drives along the same way (this time from city B to city A) at a constant speed of 60 kph.
> In your opinion, what is the driver's mean speed (in kph) for the round trip, from A to B and back to A?
> Circle your answer and explain your choice:

70 75 80 85 90

Standard 2—overall constant speed (*M*). This version asked virtually the same question, offering the same scale of multiple-choice answers, but without mentioning the word "mean," to avoid the risk of inducing participants to resort to the arithmetic mean of the speeds:

> A man drives to work, on Monday morning, from city A to city B at a constant speed of 100 kph. On his way back he is more relaxed and he drives along the same way (this time from city B to city A) at a constant speed of 60 kph. On Tuesday, he decides to drive at the *same constant speed* both on his way to work and back, but to drive so that the total time of driving from city A to B and back to A will *equal* the total driving time of Monday.
> In your opinion, at what speed should he drive on Tuesday?

Missing speed. This version is exceptional in not asking for the overall mean speed—which is given—but rather for assessment of a missing constituent of that mean (Gardner, 1982, p. 142):

> A skier travels up to the top of the mountain in a cable car that advances slowly at 5 kph. The impatient skier feels annoyed. He decides to compensate by skiing faster on his way back, down the slope, so that his mean speed for the round trip, up and down, will be 10 kph.
> At what speed should he ski downward on his way back in order to achieve that goal (assuming equal distances on the way up and down)?
> (A blank space was marked for inserting the answer.)

There is a twist to this puzzle. The correct answer is that such a speed *does not exist*. The skier wants his mean speed to be double that of going up, but in order to do this he must cover twice the distance of the way up during the same length of time it took him to go up, and all this time has already been exhausted. No time is left for the way down at any speed whatsoever.

4.2.2. Comparisons—extreme gaps

Overall mean speed versus arithmetic mean of speeds (*M* vs. *A*). The same cover story as in *Standard 1—overall mean speed* (*M*) was used, but with amplified difference between the two speeds. These were (in kph) 80 and 20 in one variant ($A = 50$; $H = 32$); and 120 and 40 in another ($A = 80$; $H = 60$). The low speed on the way back was justified by an overheated engine. Participants were asked to compare the (assessed) overall mean speed with the known A of the two given speeds. Each form ended with:

> In your opinion, which of the statements concerning the driver's *overall mean speed (in kph) for the round trip*, from A to B and back to A, is true?

1. It is greater than A.
2. It is less than A.
3. It is equal to A.

The symbol A was replaced in one variant by 50 and in the other one by 80. Each variant appeared with two different orders between the three answers.

Overall proportion versus arithmetic mean of the proportions (*P* vs. *A*). The recurrent trial of this version was a basketball shot. The two proportions of success, .20 and .80, matched the 20 kph and 80 kph of one variant of the previous comparison problem.

> A group of 20 children meets for basketball training. They get, one by one, to the marked line at a fixed distance from the basket. Each child throws the ball repeatedly *until the first hit*, and then makes room for the next child in the line.
>
> The children's group comprises 10 "experts" and 10 "tyros". An expert's chance of a hit per throw is 80%, and a tyro's chance is 20%.
>
> The coach observes all the trials and records all the throws and their outcomes (hit or miss). Finally, he computes the percentage of hits out of the total throws of all the children.
>
> In your opinion, *which percentage of hits is expected* out of all the trials recorded by the coach?

1. 50%
2. Greater than 50%.
3. Less than 50%.

Two forms had different orders among the three options.

4.3. Results and discussion

The two *Standard* assessments of the overall mean speed proved equivalent. Their outcomes are reported together in Table 4, side by side with the results of the two extreme comparisons. In *Missing speed* (not included in Table 4), only 1 of 36 participants claimed that "there is no such speed." The majority gave the answer 15.

As can be seen in Table 4, the modal assessment was the arithmetic mean of the two speeds. The equality of the weights was taken for granted: "The mean is adding the 2

Table 4
Percentages of assessed overall means under inverse weighting and their comparisons with the arithmetic mean (*A*), in four [a] versions, in Study 3 ($N = 174$)

Participants' choice	The version		
	Assessments	Comparisons—extreme gaps	
	Standard overall mean,[a] *M*	*M* versus *A*	*P*[b] versus *A*
Less than *A*	13.6	**50.0**	35.0
(Less than *H*)	(3.4)		
(Equal to *H*)	(10.2)		
Equal to *A*	**84.1**	43.5	**60.0**
Greater than *A*	2.3	6.5	5.0
n	88	46	40

Note. The correct overall mean (and proportion) is the harmonic mean, *H*, of the given numbers, which is less than their *A*.

[a]The results of the two *Standard* versions are combined. *M* is the assessed overall mean speed for the round trip ABA for different speeds of travel from A to B and from B to A.

[b]*P* is the assessed overall proportion of successes among the totality of trials obtained by running trials until the first success, and doing it equal times for two different success probabilities.

numbers and dividing by 2. Without calculating, 80 is in the middle between 60 and 100." As hypothesized, participants' penchant for equal weights was more marked and harder to debias under IWS than under SWS. An extreme-comparison task, which had the greatest corrective effect under SWS, produced some positive shift but was less effective in this case: "The ways forward and backward are of equal distance and therefore each direction has the same relative weight in computing the mean: $(40 + 120)/2 = 80$." But also: "Since on the way back the man drives more slowly, this part *lasts longer*. The computation of the average speed is a function of the *time* and not only of the distance. Therefore the driver spends more time lazily on the road (at 20 kph) than driving at 80 kph, and the mean will be less than 50 kph." Averaging via the harmonic mean turned out more difficult when the elements to be averaged were success probabilities. Preference for symmetry was revealed in justifying the answer 50%: "Because the distance of 80% from 50 is 30, and 20% are also distant from 50 by 30, the chances will be exactly in the middle, which is 50%."

The conclusions that the uniformity bias is stronger, and that shaking it off is harder, under IWS than under SWS have been based on responses to different situations and different numbers. In the next study, we endeavor to find out whether that difference is sustained also when SWS and IWS relate to the same setup.

5. Study 4—Radar recordings and average speed

5.1. Rationale

We singled out the radar situation for a dual examination of averaging speeds under both SWS and IWS, while using the same numbers and relating to the same situation. Potential differences in the extent of uniformity responses could thus be justifiably attributed to the sampling methods rather than to contextual or numerical differences. The same two speeds that had been used in the standard assessment in Study 3 were employed. For one group, these were the equally frequent speeds of cars traveling along the road, and one had to assess the mean of the speeds recorded during an hour by a *radar* mechanism (M_R). For another group, these were speeds recorded with equal frequencies by the radar, during an hour, and one had to assess the mean *speed* of the cars traveling on the road (M_S). The correct means are SW in the first case, and H in the second.

The apparent symmetry of the situations, and the fact that SW and H of two numbers are equally distant from A, might support expecting about equal rates of fallacious A answers in both cases. However, the previous findings suggest otherwise. In particular, the extra cognitive load of having to think of self-weighting and then reverse the effect may take its toll and result in a lower rate of H than of SW answers. Because of expecting more difficulties when inverse weighting is due, we devised also a *two-stage* version in that case: An auxiliary step endeavored to prime respondents by urging them to consider the situation more carefully.

5.2. Methods

We got 220 regular responses to three versions that had asked about mean speed in different ways. The two speeds to be averaged were always 100 kph and 60 kph. The scale of multiple answers was:

60 65 70 75 80 85 90 95 100

Mean recording, given speeds (M_R). Many cars travel in the same direction on the highway. The same number of cars travel on each kilometer, *half* of them at 100 kph and *half* at 60 kph. A radar trap, located at a certain point on the roadside, records during one hour the speeds of all the cars that pass by.
In your opinion, what will be the *mean speed* (in kph) recorded by the radar during an hour?

Mean speed, given recordings (M_S). Many cars travel in the same direction on the highway. A radar trap, located at a certain point on the roadside, records during one hour the speeds of all the cars that pass by. *Equal numbers* of cars traveling at 100 kph and at 60 kph were recorded during an hour. No other speeds were recorded.
Assuming that this sample is representative, what is your assessment of the *mean speed* (in kph) of the cars on the highway?

Mean speed, given recordings, with priming (M'_S). Everything was the same as in the previous (M_S) version. The difference was that an additional question had been interjected *before* asking about the mean speed of cars on the highway:
In your opinion, which of the following conclusions from the radar measurements is correct?

1. The *same number* of cars on the highway travel at 60 kph and at 100 kph.
2. *More* cars on the highway travel at 60 kph than at 100 kph.
3. *More* cars on the highway travel at 100 kph than at 60 kph.

In two alternative forms, the multiple answers appeared in different orders.

5.3. Results and discussion

Table 5 presents the distributions of the assessed mean speed in the three versions, Preference of equal weights was exhibited again in two opposite tasks. The prediction that inverse weighting will be intuitively less accessible than self-weighting—as found in the previous studies—was borne out in this study by comparing the two sampling methods in the same context with identical numbers.

Table 5
Percentages of assessed mean speed under self-weighting and inverse weighting, in three versions, in Study 4 ($N = 220$)

Participants' chosen mean (in kph)	The version		
	Mean recording,[a] given speeds, M_R	Mean speed, given recordings[a]	
		M_S	With priming, M'_S
70 or less	0.0	4.2	7.4
75 (=*H*)	0.0	0.0	11.8
80 (=*A*)	**65.4**	**89.6**	**69.1**
85 (=*SW*)	16.3	6.2	5.9
90 or more	18.3	0.0	5.9
n	104	48	68

Note. The correct answers are: $M_R = SW = 85$, and $M_S = M'_S = H = 75$.
[a]Recordings of the speeds of all cars that go past a radar device during 1 h.

In justifying $M_R = 80$, participants were oblivious of the SWS by the radar: "If equal amounts of cars pass each km, and half of them go at 100 kph and the other half at 60—altogether the average speed, should be the simple mean." Contrarily, the answer 90 (in the right direction) was explained: "The mean of 60 and 100 is indeed 80, however, during the hour more cars that travel at 100 kph go past the radar than slower cars going at 60 kph. Therefore the mean is greater than 80 and closer to 100." Counteracting the biased radar recordings for computing M_S was apparently too difficult. This resulted in a higher rate of choosing A than in assessing M_R. Choice of $A = 80$ seemed self-evident: "An equal number in each group. The ratio is 1:1 and one can compute $(100 + 60)/2$."

The moderate corrective effect of the priming was in the intended direction. The answers to the auxiliary question (not summarized in Table 5) showed that participants' evaluations of the relation between the frequencies of the two actual speeds largely determined their assessed mean, M'_S: "Because cars of 100 kph travel faster and an equal number of cars were registered by the radar, then there are more cars that travel at 60 kph." $H = 75$ was justified by: "The mean speed (considering the first question) should be closer to 60."

In the studies up to now, the weights have never been presented as such. Only the values were given. Respondents had to deduce from the described procedure that the values should be either directly or inversely weighted. Considering the difficulty of correcting the biases by different manipulations, we attempt in the next study to present the problem so that the values and the weights will be equally salient.

6. Study 5—Transparent weighting

6.1. Rationale

Whereas in the previous studies the weights had to be inferred from the problem's story, now the weights were visually displayed. We examined the effect of such an exhibit on tasks of comparing means.

Sums of money to be gained were written on sectors of a roulette dial. The size of a sector's central angle—which determines the probability that the rotated pointer will stop on that sector—was distinctly visible. In the case of SWS, the gains equaled the angles, so that the proportionality between value and weight could be seen. Participants were asked to compare expected gains between two roulettes with the same gains, one embodying SWS, and the other, representing UWS. This enabled pitting SW against A in a translucent way, and thus obtaining a measure of people's capability of distinguishing between the two under optimal conditions. The same method served also to compare uniform distributions with inversely weighted distributions (obtained by IWS), where H is the expected gain, and to compare SWS with IWS. In addition, to find out whether people sense that SW of a variable X with fixed A is directly related to the variance of X—see Eq. (3)—we presented for comparison two SWS roulettes with the same A, but with different variabilities between the gains. All the roulettes used in this study are portrayed in Fig. 1.

6.2. Methods

Each of 228 participants responded to one of four different versions. Each version presented for comparison two of the five roulette games presented in Fig. 1. All the versions started with:

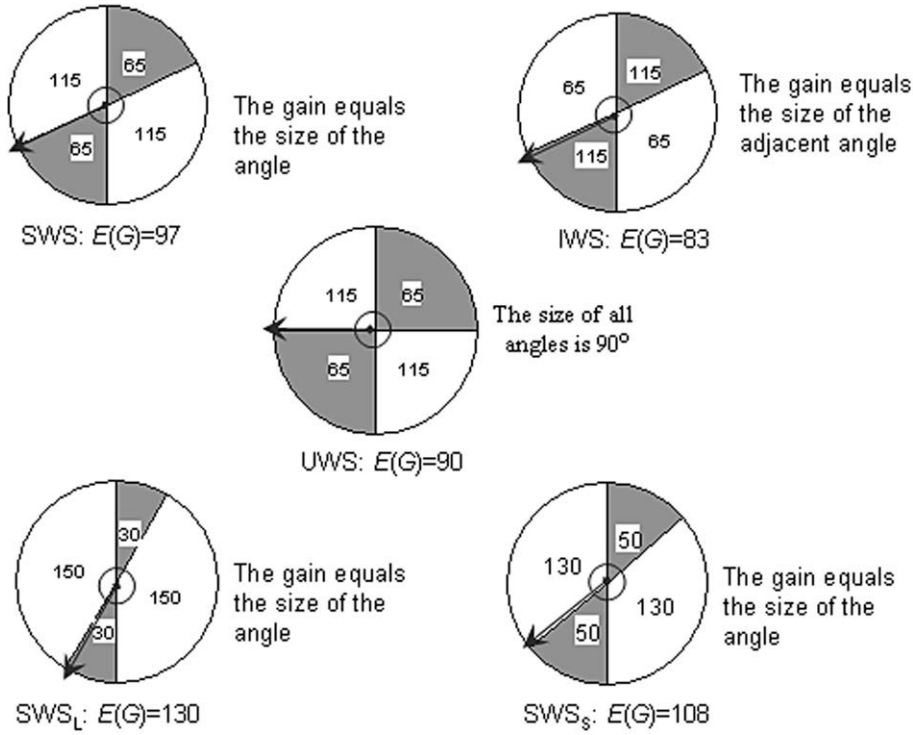

Fig. 1. The roulette games used as stimuli in four paired comparisons, in Study 5. The number written on each sector is the player's gain (G) in dollars for landing on that sector. Only the legend to the right of each roulette appeared in the forms given to the participants. The roulettes are labeled according to the *sampling method*: UWS, uniformly weighted sampling; SWS, self-weighted sampling; IWS, inversely weighted sampling; SWS_L, SWS—large variability; SWS_S, SWS—small variability.

> Two roulettes are presented below.
>
> Suppose you are offered to play a game in which you have first, to choose *one* of the roulettes, and then to turn the pointer 10 times.
>
> The number written on each sector is your gain (in $$), for each turning, if the pointer stops on that sector.

After that, the two roulettes for that version, numbered 1 and 2, were presented (including only the legend to the right of each roulette in Fig. 1), and the instructions continued:

Which roulette game is preferable (according to its expected average gain)?

The choices were:

> Roulette 1
> Roulette 2
> No difference

Each version appeared in two alternative forms with reversed order between the two roulettes. Note that the arithmetic mean of the potential gains in all the roulettes is the

same ($A = 90$), so that whoever relies on uniformity will opt for indifference between the two options. The versions below are identified by the two sampling methods (roulette games) to be compared:

Self-weighting versus uniform weighting (SWS vs. UWS).
Uniform weighting versus inverse weighting (UWS vs. IWS).
Self-weighting versus inverse weighting (SWS vs. IWS).
Self-weighting: Large versus small variance (SWS_L vs. SWS_S).

6.3. Results and discussion

The adage "seeing is believing" has been verified in this study. It turned out that displaying the probabilities paid off (Table 6). The results of the visual comparisons of the mean under uniform sampling with the mean under either self- or inverse-weighting were diametrically superior to those of assessing means in verbal problems both under SWS and under IWS, and they proved, for the most part, also superior to comparisons in the verbal problems.

Most explanations were correct, as in preferring SWS to UWS: "The greater the gain the greater its probability." Preferences of SWS to IWS were mostly justified by: "Because in roulette 2 (SWS), there is a greater area of winning more money, there are more chances of gaining 115." And, when circling indifference: "The mean is the same." Participants who had compared SWS_L versus SWS_S correctly justified their choice only by the greater probability of the greater gain: "Because I'll have a chance of $300/360 = 5/6$ to win 150. In the other roulette I'll have only a chance of $260/360 = 13/18$ to win 130." And those who were indifferent: "There is no difference between the roulettes because $50 + 130 = 180$ and $30 + 150 = 180$, therefore turning the pointer will show the same gain at the end of the day."

The advantage of a presentation in which the values are given and their weights are visible is particularly noted when comparing weighted distributions (in either direction) with uniform distributions of the same values. The level of participants' performance is not a monotonous function of the gap between the expected gains of the two roulettes. It depends also on the ease of integrating the visual and numerical information. Comparing

Table 6
Percentages of the preferred roulette game, according to the relation of its expected gain, $E(G)$, to that of the other game, in four comparisons, in Study 5 ($N = 228$)

$E(G)$ of the preferred game in relation to $E(G)$ of the other game	The compared sampling methods			
	Self-weighting versus uniform weighting (SWS vs. UWS)	Uniform weighting versus inverse weighting (UWS vs. IWS)	Self-weighting versus inverse weighting (SWS vs. IWS)	Self-weighting: large versus small variance (SWS_L vs. SWS_S)
Lesser	6.5	11.5	3.8	25.8
Equal[a]	16.1	3.8	36.5	27.4
Greater	**77.4**	**84.6**	**59.6**	**46.8**
n	62	52	52	62

Note. The correct answer is always "Greater."
[a]"Equal" means choice of the answer "no difference" between the two games, which is equivalent to saying that their expected gains are equal.

the UWS game with either SWS or IWS is easier than comparing SWS with IWS, although the $E(G)$ gap is twice as big in the latter case. When the compared roulettes consist of different pairs of values, as well as different pairs of probabilities, the complexity of integrating all this information obliterates the edge of visual displays over verbal problems. The visual-display method is not without its limitations. Yet, seeing the values to be averaged on the background of sectors of various widths that are proportional to their probabilities can be expedient in alerting participants to the need to consider unequal weights. On the whole, the method panned out.

7. General discussion

Five studies of how people intuitively weight a set of given options have converged on one answer: They prefer *equal weights*. The epistemological conception of equiprobability—historically prominent in the foundations of probability theory—proved psychologically predominant. Over and over again, participants distributed their probabilities, or weights, equally over the available categories, even when unambiguous information indicating otherwise had been available. They failed in partitioning the outcome space into units that warrant equiprobability. The totality of the outcomes is definitely significant (whether statistically or otherwise). This is not to say that people invariably believe that all outcomes of an uncertain process are equally likely. Surely, despite wishful thinking, most people who buy a lottery ticket know that their chances of either winning or not winning the grand prize are not even. However, when the inequality of probabilities is not manifestly conspicuous, embracing uniformity is predominant.

7.1. A robust assumption

The problems in response to which participants relied on uniformity varied in many ways: different contexts (probabilistic experiment, family-size issues, mean speeds), different tasks (assessments, comparisons), and different underlying sampling procedures (with frequencies either directly or inversely proportional to the values). The uniformity fallacy was expressed in unjustified fifty–fifty answers in the binary case, and in adhering to the arithmetic mean when the weights should have been heavier for greater or smaller values. Recurrences of responses determined by uniformity throughout all the experimental varieties testify to the compellingness of the equality presupposition.

The rate of the uniformity fallacy in the unmodified (standard) assessment tasks in all the studies was at least 65%. There was a mere handful of powerful corrective measures that managed to reduce the rate of the fallacy below 20%. Most of the attempted debiasing manipulations had little or no effect. The preference of uniformity is persistent and hard to extinguish. Replications of similar results under multiplicity of situations—some resembling each other and some rather divergent–validate the robustness of the tendency and preclude alternative interpretations.

7.2. Two faces of the uniformity fallacy

Most outcomes of probability experiments can be minutely partitioned into a sample space comprising points that may justifiably be considered equiprobable. Hence, attribution of equal probabilities to units that should have been unequally weighted can be con-

strued as either an error of choosing the wrong probabilities or as a failure in properly refining the partition of the space.

Judging that the two remaining cards are equiprobable, in Study 1, is mistaken because the probabilities should be 2/3 and 1/3 instead of 1/2 and 1/2, but also because the solver assumes that the remaining cards are equally likely instead of the relevant sides of these cards. Likewise, the false equal weighting of the four family-sizes, when asked about the mean per child in Study 2, reflects a double failure: that of missing the correct self-weights to be attached to the families, and that of focusing on families instead of relating to the uniform space of the children. These two ways of going wrong describe the same faulty reasoning. They are two sides of the same coin. Yet, a description of the results of averaging speeds (Experiments 3 and 4), as well as those of other studies cited above, in terms of a uniformity bias is more adequate than in terms of relating to the wrong sample space.

7.3. The effect of the sampling method

Erroneously applying uniformity is more pronounced under IWS than under SWS. This can be seen by comparing Study 3 with Studies 1 and 2. In the case of IWS, decomposing the outcome space into equally probable units is more complex. The extra difficulty of inversing the weights was conclusively demonstrated in Study 4, where the effect of the two sampling methods had been compared in identical setups. These results tally with the greater difficulty of apprehending inverse relations than direct relations, as found in cognitive developmental studies.

7.4. Remedial measures

The uniformity bias is not absolute. Though not easily accomplished—as evidenced by quite a few reasonable but futile attempts—some means did prove successful to different extents. These measures have instrumental implications for educational and practical ends. As a rule, the less abstract or formal, and the more vivid or personalized the presentation of the problem, the greater are people's chances of partitioning the space into equally likely elements. Thus, symbolic labeling of the cards' sides had only a small effect, whereas replacing sides with humans, and lending them individuality via names, had a greater corrective effect. In Study 2, drawing attention to individual children, by naming and sketching them, did overturn the pattern of the results in favor of the correct self-weighting. Extremizing the differential weighting, particularly when combined with a comparison task, reversed the pattern of the results. Above all, figural presentations of the weights (probabilities) in conjunction with the values to be weighted induce people to largely abandon uniformity and apply the proper weights both under self- and inverse-sampling.

7.5. The generality of the phenomenon

Uniformity is a construct that appears to play an important role in additional areas of human cognition. Some examples are the perception of randomness, problem solving, and social judgment. People's prototypical image of a random pattern is virtually an embodiment of *local uniformity*. Preference of uniformity appears in diverse problem-solving situations: Polya (1981) gave examples of geometric and algebraic problems in which uniformity is an expedient heuristic. Social and economical judgments are often motivated

by the notion of equity (Benartzi & Thaler, 2001; Roch et al., 2000). Despite the common knowledge that "some men are more equal than others", collocations such as "equal opportunity", "equal pay", and "equal rights" abound in public social discourse and betray human desiderata. Equality serves often as a starting point for diverse decisions. People anchor on uniformity and either stay there (Fox & Rottenstreich, 2003) or adjust their views insufficiently in the face of pertinent evidence (Fox & Clemen, 2005; Tversky & Kahneman, 1974).

7.6. The roots of the phenomenon

Considering the centrality of uniformity in probability theory, there is no wonder that the teaching of probability starts with the case of equally likely outcomes. The predisposition to uniformity and the teaching practices reinforce and perpetuate each other; it is hard to say which comes first. Hawkins et al. (1992) asserted, on the basis of vast statistical-education experience, that

> The equally likely approach seems to be a natural starting point for the study of probability, especially where young children are concerned...there is a (well-known) danger that a student reared on an 'equally-likely diet' will always attach a probability of 0.5 to each of two mutually exclusive and exhaustive events based on *any* probability experiment, irrespective of how different the events' probabilities really are. (pp. 65–66).

Moreover, the terms *average* and *mean* are usually introduced to children since early school years by instructing them to add up the given values and divide the result by their number. No mention of weights, let alone unequal weights, is ever made. Ubiquitous usage of the terms in that sense is commonplace in daily and professional discourse and in the media. This meaning has often been noticeable in our participants' verbalizations.

Some of the reasons for people's uniformity disposition are rather clear. Allotting equal probabilities or weights to all the available options is apparently the *simplest* of all decisions, and the first that comes to mind. It calls for *minimal mental effort*; any other distribution requires careful deliberation. Zabell (1988) highlighted "The Insidious Assumption of Symmetry" (pp. 159–165). He talked about the seductive attraction of symmetry arguments and explained the gist of the historical appeal of symmetry for philosophers and scientists as a *compromise* that had resolved conflicts between opposing possibilities. Symmetry or equilibrium apparently reflect also certain aesthetic expectations and a desire for "elegance," which characterizes the motivation of mathematicians. According to Polya (1981), regular polygons, whose sides and angles are all equal, are popular among problem solvers because they are nearest to perfection. All these factors conspire to make the choice of uniformity preeminent.

Polya (1981) suggested that frequently elements that play the same role in the givens may be expected to play the same role in the solution. People have learned that *equal conditions produce equal results*. This works often in mathematics. For example, in a triangle, the angles opposite equal sides are also equal, and in equilateral triangles—altitudes, medians, and angle-bisectors are all equal. In our research, the three cards appeared to be of equal status, as did the different family-sizes and speeds. This could explain why so many participants persisted in regarding these elements equiprobable also in the solution.

Polya's rule of thumb: "Symmetry should result from symmetry" (p. 161) apparently failed them because of overlooking the dissymmetry introduced by the procedures.

The preponderance of uniformity choices might well be a matter of *expediency*. Harris and Joyce (1980) reasoned that minimizing conceptual effort might have underlain their participants' recommendations to divide outcomes equally among all partners in a group project. Since people make inferences under constraints of limited time and computational capacity, they often resort to simple "fast and frugal" (Gigerenzer & Goldstein, 1996) algorithms. Yet they may perform adequately. Assuming uniformity, even when not realistically correct, simplifies the computation considerably while frequently allowing good-enough results. It is reinforced by the fact that a uniform distribution often constitutes an extreme case of the situation, as in the birthday problem. Possibly, people's spontaneous tendency to uniformity had evolved because of long-standing positive experience associated with relying on that assumption.

7.7. Conclusions

Decisions based on unjustifiably assuming uniformity are being made time and again in diversified circumstances. Equality is not just a heuristic, it is people's default assumption, which is rather stubborn. Even under conditions that maximally reduce the tendency to uniformity, there always remains a vestige of fallacious answers affected by that assumption. The primacy of opting for uniformity is accounted for by a cognitive quest for equity, impartiality, perhaps also harmony, as well as by pragmatic, utilitarian factors. Equality of all probabilities could sometimes be false as a belief, but expeditious as a guide for behavior. Whether uniformity is adaptive (Gigerenzer, 2000) by often providing satisfactory results without spending excessive efforts, and is being reinforced because of promoting one's goals, should be examined. This requires independent behavioral studies comparing the outcomes of decisions based on assumed uniformity and on real distributions, as done by Burns (2004) with respect to the "hot hand" belief (Gilovich, Vallone, & Tversky, 1985). This inquiry is outside the scope of the present basic research. It is worth pursuing.

Though the roots of people's penchant for equiprobability are not fully understood, the contribution of this research is documenting the centrality of uniformity as a construct that permeates many cognitions. The size of the deviations from truth caused by falsely applying uniformity might not be practically pernicious, nonetheless, such judgments are *wrong in principle*. We deem it important for psychologists, educators, and other scholars to be aware of the risks involved in the seductive appeal of uniformity. Echoing Zabell (1988), who echoed Freud, this work could have been entitled "Uniformity and its Discontents".

References

Albert, J. H. (2003). College students' conceptions of probability. *The American Statistician, 57*(1), 37–45.

Bar-Hillel, M., & Falk, R. (1982). Some teasers concerning conditional probabilities. *Cognition, 11*, 109–122.

Basano, L., & Ottonello, P. (1996). The ambiguity of random choices: Probability paradoxes in some physical processes. *American Journal of Physics, 64*(1), 34–39.

Benartzi, S., & Thaler, R. H. (2001). Naive diversification strategies in defined contribution saving plans. *The American Economic Review, 91*(1), 79–98.

Berresford, G. C. (1980). The uniformity assumption in the birthday problem. *Mathematics Magazine, 53*(5), 286–288.

Beyth-Marom, R. (1977). *Aspects of the perception of association between classes, events, and variables* (Hebrew with English Abstract). Unpublished Doctoral Dissertation, The Hebrew University of Jerusalem, Israel.

Bickel, P. J., Hammel, E. A., & O'Connell, J. W. (1975). Sex bias in graduate admissions: Data from Berkeley. *Science, 187*, 398–404.

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*(1), 3–21.

Bruine de Bruin, W., Fischhoff, B., Millstein, S. G., & Halpern-Felsher, B. L. (2000). Verbal and numerical expressions of probability: "It's a fifty–fifty chance". *Organizational Behavior and Human Decision Processes, 81*(1), 115–131.

Burns, B. D. (2004). Heuristics as beliefs and as behaviors: The adaptiveness of the "hot hand. *Cognitive Psychology, 48*, 295–331.

Bytheway, B. (1974). A statistical trap associated with family size. *Journal of Biosocial Science, 6*, 67–72.

Carnap, R. (1953). What is probability?. *Scientific American 189*(3), 128–138.

Carroll, L. (1958). *Pillow problems and a tangled tale*. New York: Dover (*Pillow problems* was originally published separately in 1895 by Macmillan).

Christensen, R., & Utts, J. (1992). Bayesian resolution of the "exchange paradox". *The American Statistician, 46*(4), 274–276.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 49*(12), 997–1003.

Cosmides, L., & Tooby, J. (1996). Are humans good intuitive statisticians after all? Rethinking some conclusions from literature on judgment under uncertainty. *Cognition, 58*(1), 1–73.

Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist, 34*, 571–582.

Falk, R. (1982). Do men have more sisters than women? *Teaching Statistics, 4*, 60–62.

Falk, R. (1992). A closer look at the probabilities of the notorious three prisoners. *Cognition, 43*, 197–223.

Falk, R., & Greenbaum, C. W. (1995). Significance tests die hard: The amazing persistence of a probabilistic misconception. *Theory & Psychology, 5*(1), 75–98.

Falk, R., & Konold, C. (1997). Making sense of randomness: Implicit encoding as a basis for judgment. *Psychological Review, 104*(2), 301–318.

Falk, R., Lann, A., & Zamir, S. (2005). Average speed bumps: Four perspectives on averaging speeds. *Chance, 18*(1), 25–32.

Falk, R., Lipson, A., & Konold, C. (1994). The ups and downs of the hope function in a fruitless search. In G. Wright & P. Ayton (Eds.), *Subjective probability* (pp. 353–377). Chichester, England: Wiley.

Falk, R., & Samuel-Cahn, E. (2001). Lewis Carroll's obtuse problem. *Teaching Statistics, 23*(3), 72–75.

Falk, R., & Wilkening, F. (1998). Children's construction of fair chances: Adjusting probabilities. *Developmental Psychology, 34*(6), 1340–1357.

Feller, W. (1957). *An introduction to probability theory and its applications* (2nd ed., Vol. 1). New York: Wiley.

Fiedler, K. (2000). Beware of samples! A cognitive-ecological sampling approach to judgment biases. *Psychological Review, 107*(4), 659–676.

Fine, T. L. (1973). *Theories of probability: An examination of foundations*. New York: Academic Press.

Fischhoff, B., & Bruine de Bruin, W. (1999). Fifty–fifty = 50%? *Journal of Behavioral Decision Making, 12*, 149–163.

Fisher, R. A. (1934). The effect of methods of ascertainment upon the estimation of frequencies. *Annals of Eugenics, 6*, 13–25.

Fox, C. R., & Clemen, R. T. (2005). Subjective probability assessment in decision analysis: Partition dependence and bias toward the ignorance prior. *Management Science, 51*(9), 1417–1432.

Fox, C. R., & Levav, J. (2004). Partition-edit-count: Naive extensional reasoning in judgment of conditional probability. *Journal of Experimental Psychology: General, 133*(4), 626–642.

Fox, C. R., & Rottenstreich, Y. (2003). Partition priming in judgment under uncertainty. *Psychological Science, 14*(3), 195–200.

Gardner, M. (1978). *Aha! Insight*. New York: Freeman.

Gardner, M. (1982). *Aha! Gotcha: Paradoxes to puzzle and delight*. New York: Freeman.

Gavanski, I., & Hui, C. (1992). Natural sample spaces and uncertain belief. *Journal of Personality and Social Psychology, 63*(5), 766–780.

Gigerenzer, G. (2000). *Adaptive thinking: Rationality in the real world*. New York: Oxford University Press.

Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review, 103*, 650–669.

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review, 102*(4), 684–704.

Gigerenzer, G., & Murray, D. J. (1987). *Cognition as intuitive statistics*. Hillsdale, NJ: Lawrence Erlbaum.

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: Cambridge University Press.

Gilovich, T., Vallone, R., & Tversky, A. (1985). The hot hand in basketball: On the misperception of random sequences. *Cognitive Psychology, 17*, 295–314.

Gorodetsky, M., Hoz, R., & Vinner, S. (1986). Hierarchical solution models of speed problems. *Science Education, 70*(5), 565–582.

Hacking, I. (1975). *The emergence of probability*. Cambridge: Cambridge University Press.

Haight, F. A. (1963). *Mathematical theories of traffic flow*. New York: Academic Press.

Harris, R. J., & Joyce, M. A. (1980). What's fair? It depends on how you phrase the question. *Journal of Personality and Social Psychology, 38*(1), 165–179.

Hawkins, A., Jolliffe, F., & Glickman, L. (1992). *Teaching statistical concepts*. London: Longman.

Hemenway, D. (1982). Why your classes are larger than "average". *Mathematics Magazine, 55*(3), 162–164.

Huck, S. W., & Sandler, H. M. (1984). *Statistical illusions: Problems/Solutions*. New York: Harper & Row.

Jenkins, J. J., & Tuten, J. T. (1992). Why isn't the average child from the average family?—and similar puzzles. *American Journal of Psychology, 105*(4), 517–526.

Johnson-Laird, P. N., Legrenzi, P., Girotto, V., Legrenzi, M. S., & Caverni, J. P. (1999). Naive probability: A mental model theory of extensional reasoning. *Psychological Review, 106*(1), 62–88.

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology, 3*(3), 430–454.

Kareev, Y. (1992). Not that bad after all: Generation of random sequences. *Journal of Experimental Psychology: Human Perception and Performance, 18*(4), 1189–1194.

Keren, G. (1984). On the importance of identifying the correct 'problem space'. *Cognition, 16*, 121–128.

Keyfitz, N. (1968). *Introduction to the mathematics of population*. Reading, MA: Addison-Wesley.

Keyfitz, N. (1985). *Applied mathematical demography* (2nd ed.). New York: Springer.

Keynes, J. M. (1943). *A Treatise on probability*. London: Macmillan.

Konold, C. (1989). Informal conceptions of probability. *Cognition and Instruction, 6*(1), 59–98.

Kotz, S., & Stroup, D. F. (1983). *Educated guessing: How to cope in an uncertain world*. New York: Marcel Dekker.

Krauss, S., & Wang, X. T. (2003). The psychology of the Monty Hall problem: Discovering psychological mechanisms for solving a tenacious brain teaser. *Journal of Experimental Psychology: General, 132*(1), 3–22.

Lann, A. (2008). Three averages—ordered by comparing areas. *Teaching Statistics, 30*(1), 13.

Lann, A., & Falk, R. (2005). A closer look at a relatively neglected mean. *Teaching Statistics, 27*(3), 76–80.

Lann, A., & Falk, R. (2006). Tell me the method, I'll give you the mean. *The American Statistician, 60*(4), 322–327.

Levin, I. P. (1974). Averaging processes and intuitive statistical judgments. *Organizational Behavior and Human Performance, 12*, 83–91.

Nickerson, R. S. (1996). Ambiguities and unstated assumptions in probabilistic reasoning. *Psychological Bulletin, 120*(3), 410–433.

Nickerson, R. S. (2002). The production and perception of randomness. *Psychological Review, 109*(2), 330–357.

Nickerson, R. S. (2004). *Cognition and chance: The psychology of probabilistic reasoning*. Mahwah, NJ: Lawrence Erlbaum.

Nickerson, R. S., & Falk, R. (2006). The exchange paradox: Probabilistic and cognitive analysis of a psychological conundrum. *Thinking & Reasoning, 12*(2), 181–213.

Nisbett, R., & Ross, L. (1980). *Human inference: Strategies and shortcomings of social judgment*. Englewood Cliffs, NJ: Prentice-Hall.

Patil, G. P., & Rao, C. R. (1978). Weighted distributions and size-biased sampling with applications to wildlife populations and human families. *Biometrics, 34*, 179–189.

Patil, G. P., Rao, C. R., & Zelen, M. (1988). Weighted distributions. In S. Kotz & N. L. Johnson (Eds.), Encyclopedia of statistical sciences (Vol. 9, pp. 565–571). New York: Wiley.

Pollatsek, A., Lima, S., & Well, A. (1981). Concept or computation: Students' understanding of the mean. *Educational Studies in Mathematics, 12*, 191–204.

Polya, G. (1981). *Mathematical discovery: On understanding, learning, and teaching problem solving (combined ed.)*. New York: Wiley.

Portnoy, S. (1994). A Lewis Carroll pillow problem: Probability of an obtuse triangle. *Statistical Science, 9*(2), 279–284.

Rao, C. R. (1977). A natural example of a weighted binomial distribution. *The American Statistician, 31*(1), 24–26.

Roch, S. G., Lane, J. A. S., Samuelson, C. D., Allison, S. T., & Dent, J. L. (2000). Cognitive load and the equality heuristic: A two-stage model of resource overconsumption in small groups. *Organizational Behavior and Human Decision Processes, 83*(2), 185–212.

Ross, J. (1985). Misuse of statistics in social sciences. *Nature, 318*, 514.

See, K. E., Fox, C. R., & Rottenstreich, Y. S. (2006). Between ignorance and truth: Partition dependence and learning in judgment under uncertainty. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32*(6), 1385–1402.

Shimojo, S., & Ichikawa, S. (1989). Intuitive reasoning about probability: Theoretical and experimental analyses of the "problem of three prisoners". *Cognition, 32*, 1–24.

Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society (Series B), 13*(2), 238–241.

Smith, D. S. (1979). Averages for units and averages for individuals within units: A note. *Journal of Family History, 4*, 84–86.

Stein, W. E., & Dattero, R. (1985). Sampling bias and the inspection paradox. *Mathematics Magazine, 58*(2), 96–99.

Stigler, S. M. (1986). *The history of statistics: The measurement of uncertainty before 1900*. Cambridge, MA: Harvard University Press.

Teigen, K. H., & Keren, G. (2007). Waiting for the bus: When base-rates refuse to be neglected. *Cognition, 103*(3), 337–357.

Todhunter, I. (2001). *A history of the mathematical theory of probability from the time of Pascal to that of Laplace*. Bristol, UK: Thoemmes Press (Original work published 1865).

Tune, G. S. (1964). Response preferences: A review of some relevant literature. *Psychological Bulletin, 61*(4), 286–302.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*, 1124–1131.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science, 211*, 453–458.

Tversky, A., & Kahneman, D. (1982). Evidential impact of base rates. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 153–160). Cambridge: Cambridge University Press.

van Dijk, N. M. (1997). To wait or not to wait: That is the question. *Chance, 10*(1), 26–30.

von Mises, R. (1957). *Probability, statistics and truth* (J. Neyman, D. Scholl & E. Rabinowitsch, Trans., 2nd rev. English ed.). New York: Dover. (Original work published 1928).

vos Savant, M. (1990). Ask Marilyn. *Parade Magazine* (December 2), 25.

vos Savant, M. (1991). Ask Marilyn. *Parade Magazine* (February 17), 12.

Wilkening, F. (1981). Integrating velocity, time, and distance information: A developmental study. *Cognitive Psychology, 13*, 231–247.

Zabell, S. L. (1988). Symmetry and its discontents. In B. Skyrms & W. L. Harper (Eds.), *Causation, chance, and credence* (Vol. 1, pp. 155–190). Dordrecht, Holland: Kluwer.

Zelen, M., & Feinleib, M. (1969). On the theory of screening for chronic diseases. *Biometrika, 56*(3), 601–614.