Journal of Behavioral Decision Making
J. Behav. Dec. Making, **13**: 133–139 (2000)

Was Bernoulli Wrong? On Intuitions about Sample Size

PETER SEDLMEIER[1]* and GERD GIGERENZER[2]
[1]*University of Paderborn, Germany*
[2]*Max Planck Institute for Human Development, Berlin, Germany*

ABSTRACT

Recently we proposed an explanation for the apparently inconsistent result that people sometimes take account of sample size and sometimes do not: Human intuitions conform to the 'empirical law of large numbers,' which helps to solve what we called 'frequency distribution tasks' but not 'sampling distribution tasks' (Sedlmeier and Gigerenzer, 1997). Keren and Lewis (2000) do not provide an alternative explanation but present a three-pronged criticism of ours: (1) the normative argument that a larger sample size will not invariably provide more reliable estimates, (2) the descriptive argument that under certain circumstances, people are insensitive to sample size, and (3) the claim that sampling distributions are essential for solving both frequency and sampling distribution tasks. We argue that (1) the normative argument is irrelevant for our descriptive hypothesis and, as a normative claim, only valid for a specific situation, (2) the descriptive argument is correct but consistent with our review, and (3) is incorrect. Bernoulli's assertion that the intuitions of 'even the stupidest man' follow the empirical law of large numbers may have been rather on the optimistic side, but in general the intuitions of the vast majority of people do. Copyright © 2000 John Wiley & Sons, Ltd.

KEY WORDS sample size; law of large numbers; statistical reasoning

Do people take into account information about sample size in statistical reasoning tasks? The literature exhibits seemingly contradictory results: in one group of studies people do and in another they do not. Recently we suggested an explanation for a substantial part of these inconsistent findings (Sedlmeier and Gigerenzer, 1997). We proposed the hypothesis that human intuition conforms to the *empirical law of large numbers*, which expresses the empirically observable fact that larger samples generally (but not always) lead to more accurate estimates of population means (Sedlmeier and Gigerenzer, 1997, p. 35; see also Sedlmeier, 1998, p. 281). We further argued that this intuition works only with what we termed 'frequency distribution tasks' and not with 'sampling distribution tasks.'

Consider, for example, Kahneman and Tversky's (1972) maternity ward task:

> A certain town is served by two hospitals. In the larger hospital about 45 babies are born each day, and in the smaller hospital about 15 babies are born each day. As you know, about 50% of all babies are boys. The exact percentage of boys, however, varies from day to day. Sometimes it may be higher than 50%, sometimes lower.

* Correspondence to: Peter Sedlmeier, University of Paderborn, FB2-Psychology, 33095 Paderborn, Germany.

CCC 0894–3257/2000/010133–07$17.50

In the original, that is, a *sampling distribution task*, it continues:

> For a period of 1 year, each hospital recorded the days on which more than 60% of the babies born were boys. Which hospital do you think recorded more such days?

The corresponding part in a *frequency distribution task* reads:

> Which hospital do you think is more likely to find on a randomly chosen day that more than 60% of babies born were boys?

The relevant difference between the two tasks lies in the kind of distribution involved. Whereas in the sampling distribution task the relative variance of sampling distributions (of proportions of babies who are boys) has to be judged, in the frequency distribution task two proportions based on frequency distributions (of numbers of babies who are boys and girls) have to be judged. The intuition that larger samples usually lead to more exact estimates of the population mean or proportion (here: 50% boys) helps in the solution of the frequency rather than the sampling distribution tasks.[1] Therefore we derived the hypothesis that researchers who reported that people attend to sample size used predominantly frequency distribution tasks whereas those who concluded that people largely ignore sample size used mainly sampling distribution tasks. This is what we found. The hypothesis that frequency distribution tasks yield higher solution rates than corresponding sampling distribution tasks was subsequently supported in independent studies (Sedlmeier, 1998).

Note that the hypothesis that people use an intuition corresponding to the law of large numbers when solving sample size tasks is descriptive, *not* normative. However, we also addressed the normative argument frequently found in sample size research: that the (mathematical) law of large numbers can be used as a yardstick to evaluate people's judgments. We explained why the law of large numbers cannot serve as such a normative yardstick and pointed out that there exists no general normative model for sample size tasks — only partial mathematical justifications for the superiority of larger samples, such as the variance of the sample mean, Chebychev's inequality, and the central limit theorem. None of these results justify a strictly monotonic relation between sample size and the confidence accruing therefrom.

Keren and Lewis now argue that Bernoulli (and we) may have been wrong on the following points: '(i) Normatively, an estimate from a larger sample is not necessarily closer to the corresponding population value that it attempts to estimate, and (ii) Descriptively, under certain circumstances, people are indeed insensitive to sample size' (p. 125). In addition, they argue that the concept of a sampling distribution is essential for solving frequency distribution tasks.

THE NORMATIVE ARGUMENT

Keren and Lewis mistakenly think that we argue that the empirical law of large numbers is (1) a normative model that (2) says that an estimate from a larger sample would be invariably closer to the population value. Based on this misunderstanding, they argue against these two points by presenting an example that shows a nonmonotonic relationship between sample size and the accuracy of estimates, using the binomial distribution. From that example, they also draw some implications about statistical power. But (1) the empirical law of large numbers is not a norm — we used it as a hypothesis about the intuitions of ordinary people, that is, a descriptive statement. Concerning point (2), Keren and Lewis criticize us (and Bernoulli) that 'it is inaccurate to assume that a larger sample size will *invariably* provide a more reliable estimate than the smaller one' (p. 128; italics ours). This implies incorrectly that we actually made this claim. What we said was 'This intuition — that larger samples *generally* lead to more accurate estimates of population means — is commonly referred to as the "empirical law of large numbers"' (Sedlmeier and Gigerenzer, 1998, p. 35; italics ours). There is a difference between generally and invariably.

Although Keren and Lewis's normative argument is rather peripheral to our hypothesis about people's intuitions, in the following we clarify it and the related comment on statistical power.

[1] To solve the frequency distribution task, the intuition conforming to the empirical law of large numbers would lead to the conclusion that the proportion of boys in the larger hospital can be expected to lie closer to 50% and that therefore, the proportion in the smaller hospital is more likely to have a deviation of 60% or more. Of course, one could apply the intuition repeatedly to solve the sampling distribution task but this implies additional reasoning processes and is apparently not what people commonly do.

Can smaller samples be closer to the population mean than larger ones?

There is no normative model that excludes the possibility that a smaller sample may be closer to the population value than a larger sample. The law of large numbers (here Keren and Lewis seem to be in full agreement with us) says nothing about such a comparison; it just says that as the sample size approaches infinity, the probability that the deviation of the mean of the sample from the population mean exceeds an arbitrarily small and fixed number approaches 0. As von Mises (1957/1981, p. 115) wrote: ' "The Law of Large Numbers" . . . says nothing about the course of a concrete sequence of observations.' Likewise, none of the partial justifications we mentioned allows one to conclude that as sample size increases, the sample mean *monotonically* approaches the population mean. How about the empirical law of large numbers? As the name implies, this is an *empirical* law that everybody can observe, not a normative yardstick. Keren and Lewis have incorrectly understood us as establishing a formal link between the central limit theorem and the empirical law of large numbers, possibly because we spoke of the central limit theorem as a partial mathematical justification for the superiority of larger samples. We reiterate that there is no mathematical connection between the two. Exhibit 1 shows a manifestation of the empirical law of large numbers where with increasing sample size the difference between the sample mean and the population mean generally, but not invariably, decreases (for similar examples, see Freedman *et al.*, 1991, p. 250, and Freudenthal, 1972, p. 484). Anybody can produce results similar to those shown in Exhibit 1 by tossing a coin 50 times and calculating the proportion of heads after every trial. Of course, one cannot expect that this proportion approaches the population value ($p = 0.5$) *monotonically*.

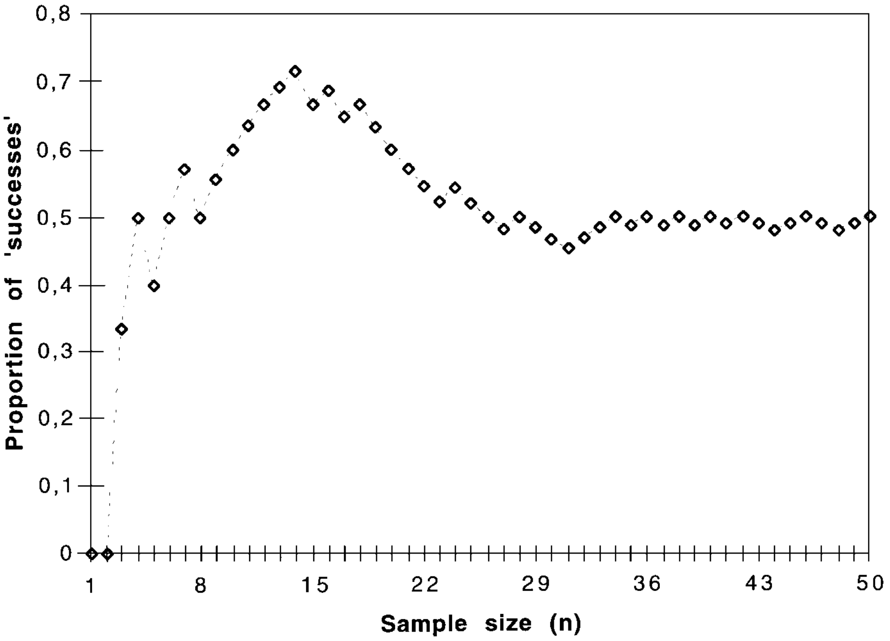

Exhibit 1. Proportions of 'successes,' calculated over 50 consecutive results from a Bernoulli distribution (i.e. possible outcomes of 0 and 1) with $p = 0.5$. For instance, in this special case, the outcomes of the first four trials were 0, 0, 1, and 1 (with 0 corresponding to 'failure' and 1 to 'success') yielding proportions of successes of 0, 0, 0.33, and 0.5 for trials 1 to 4, respectively

What is the criterion: arbitrarily fixed boundaries or mean absolute deviations?

Keren and Lewis claim that in the maternity ward task a larger sample does not necessarily provide an estimate that is closer to the population value. In their demonstration, they use a binomial distribution with $p = 0.5$ and sample sizes from $n = 2$ to $n = 50$. In analogy to one specific version of the maternity ward task, they examine the probability that the outcome of a sampling process lies between $0.4n$ and $0.6n$. They find that this probability does not increase monotonically with increasing sample size. Is this demonstration a good argument against the claim that, generally, the accuracy of an estimate increases with increasing sample size? We think not. Consider sample sizes of $n = 2$ and $n = 3$. The possible outcomes for $n = 2$ are 0, 1, and 2 successes out of 2. How many of these outcomes lie within 0.8 ($0.4n$) and 1.2 ($0.6n$) and what is their probability of occurrence? The only outcome that satisfies their criterion is 1 and its probability of occurrence is 0.5. Now consider the same for $n = 3$. Here, *none* of the possible outcomes lies within 1.2 ($0.4n$) and 1.8 ($0.6n$), and therefore, the probability that the result of a sampling process lies within the interval is 0. This is due to the fact that the size of the deviation from the expected mean or proportion of a discrete distribution is expressed in multiples of arbitrarily fixed proportions.[2] Moreover, the empirical law of large numbers is about the closeness between an estimate and the corresponding population value and not about a specific kind of deviation from it. So, we recommend looking at more general measures of deviation.

Exhibit 2 shows what happens when one examines just such a measure, the expected values for mean absolute deviations between proportion estimates and corresponding population values.[3] The example illustrates that the nonmonotonicity observed by Keren and Lewis disappears if one does not use arbitrarily fixed boundaries.

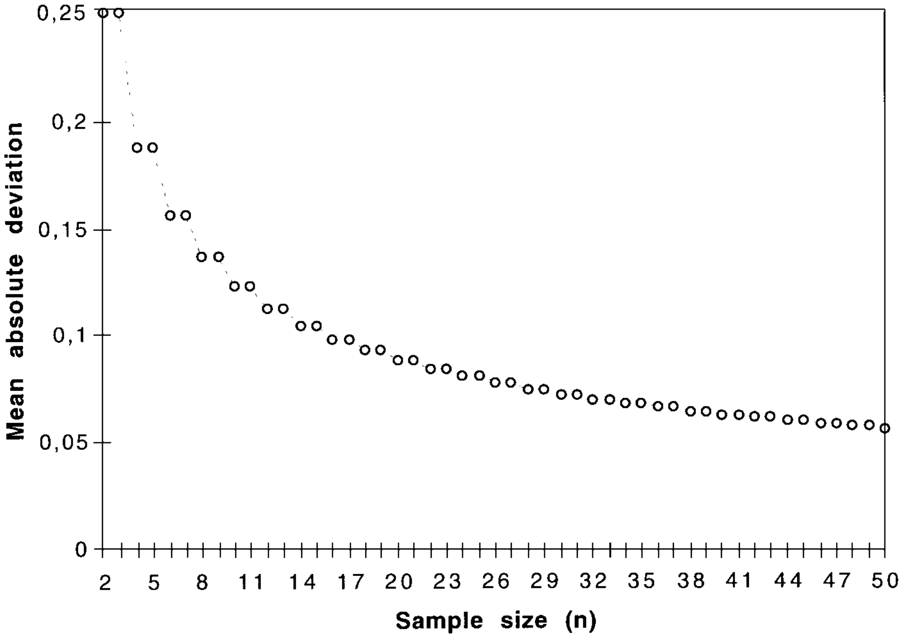

Exhibit 2. Expected values for mean absolute deviations between proportion estimates and corresponding population values for sample sizes from $n = 2$ to $n = 50$, using a binomial distribution with $p = 0.5$ (see also footnote 3)

[2] In the original maternity ward task, which also deals with a fixed deviation from the proportion of 0.5, this problem does not arise, because the size of the larger sample is a multiple of the size of the smaller one (15 versus 45), and the fixed boundaries coincide with possible outcomes in the smaller sample ($0.4 \times 15 = 6$, and $0.6 \times 15 = 9$).

[3] For any sample size n, the mean absolute deviation (MAD) is calculated as:

$$\mathrm{MAD}_n = \sum_{k=0}^{n}\left[\binom{n}{k}p^k(1-p)^{n-k}\left(\left|\frac{k}{n}-p\right|\right)\right]$$

For instance, to calculate the MAD for $n = 2$, one has possible values for k of 0, 1, and 2. The absolute deviations of the k/n from a $p = 0.5$ are 0.5, 0, and 0.5, and their respective probabilities according to the binomial distribution with $p = 0.5$ are 0.25, 0.5, and 0.25. Thus $\mathrm{MAD}_2 = 0.25 \times 0.5 + 0.5 \times 0 + 0.25 \times 0.5 = 0.25$.

Statistical power

Keren and Lewis argue that the nonmonotonicity they found has implications for statistical power. Again, this argument is only indirectly linked to our hypothesis that people's intuitions conform to the empirical law of large numbers. It deals with how formal statistical properties depend on the variation in sample sizes, not with people's intuitions. Here, the issue is the same as in the previous section: the restriction of possible outcome values when expressed in multiples of fixed proportions. Consider one of Keren and Lewis's examples in which, using binomial distributions, they find that for $H_0: \mu = 0.5$ and $H_1: \mu = 0.7$, and alpha $= 0.05$, the power for $n = 18$ is larger than that for $n = 25$. How can this be? The reason is that, because of their discrete values, the possible outcomes closest to and smaller than alpha differ from alpha by different amounts. Thus, the power calculation uses different 'true alphas.' For instance, for $n = 18$, the value closest to and smaller than alpha is $p(13 \leqslant X \leqslant 18) = 0.048$ with a corresponding power of 0.534. However, for $n = 25$, the corresponding value is only half the size: $p(18 \leqslant X \leqslant 25) = 0.022$; and not so surprisingly, the power is not larger than in the prior case — 0.512. If, for $n = 25$, one took an alpha slightly larger than 0.05, that is, $p(17 \leqslant X \leqslant 25) = 0.054$, then the corresponding power would jump up to 0.677. Keren and Lewis's argument concerning power shows that one should not slavishly stick to an alpha of 0.05 (or 0.01) because 'Good loves the 0.06 nearly as much as the 0.05' (Rosnow and Rosenthal, 1989, p. 1277; see also Gigerenzer, 1993). Anyway, the problem of differently spaced values in discrete distributions (depending on sample size) ceases to exist with continuous sampling distributions such as the *t* or the *F* distribution mostly used in behavioural research.

To sum up, the empirical law of large numbers is neither a mathematical theorem nor a norm for how people should solve sample size tasks. Instead, we used it as a hypothesis for the intuitions people use to answer frequency distribution tasks.

THE DESCRIPTIVE ARGUMENT

Keren and Lewis's descriptive conjecture is quite modest: 'under certain circumstances, people are indeed insensitive to sample size' (p. 125). In our article (Sedlmeier and Gigerenzer, 1997), we started out from exactly this observation — that people are sometimes sensitive and sometimes insensitive to sample size — and proposed an answer to what 'under certain circumstances' means: sampling distribution tasks. The challenging question is not whether or not people are 'sometimes' insensitive, but why and when (Gigerenzer, 1996). Our review of the literature indicated that insensitivity is particularly acute when the judgement is about the variability of a sampling distribution, that is, when people were asked to solve sampling distribution tasks.

Keren and Lewis mention a study of their own where they used the frequency distribution version of the maternity ward task. Although they do not provide exact values, they report (1) a relatively high proportion of 'equally likely' responses and (2) that the participants who did not choose the 'equally likely' category generally were sensitive to sample size. Given that they used arbitrarily fixed boundaries in their experiment, the choice of 'larger sample' was not always normatively correct, but consistent with our hypothesis that people use an intuition conforming to the empirical law of large numbers. The result of their study seems to be within the range of those which we have reviewed (Sedlmeier and Gigerenzer, 1997; Sedlmeier, 1998).

In sum, Keren and Lewis's conjecture that people are 'sometimes' insensitive to sample size is consistent with our previous review. But the challenging question is, rather, why and when?

FREQUENCY DISTRIBUTION TASKS VERSUS SAMPLING DISTRIBUTION TASKS

According to Keren and Lewis, both frequency and sampling distribution versions of sample size tasks (e.g. the maternity ward task cited above) require statistical inference for which the concept of sampling distribution is indispensable. Of course, both versions of the task can be solved by recourse to sampling distributions. But they need not. After a fair amount of statistical training, sampling distributions may indeed become a natural way to think about the two kinds of tasks. We agree that the concept of a sampling distribution is necessary to solve the sampling distribution version of the task, but is it also necessary to solve the frequency distribution version? A glance at Exhibit 1 shows that a frequency distribution task can be solved without sampling distributions: from intuitively knowing that generally, the estimate from a larger sample lies closer to the population value than that from a smaller sample. This solution is less elegant than the one relying on the concept of sampling distributions, but this seems to be the way laypeople solve the task. It is also evident that the empirically observable law illustrated in Exhibit 1 does not give much help in solving sampling distribution tasks.

To sum up, frequency distribution tasks can be solved without recourse to sampling distributions. How? By an intuition that conforms to the empirical law of large numbers, just as Bernoulli proposed. This might be difficult to conceive for statistically trained professionals, whose statistical education has replaced 'satisficing' statistical intuitions (Fischbein, 1975).

CONCLUSION

We have argued that Keren and Lewis's normative argument that 'an estimate from a larger sample is not necessarily closer to the corresponding population value' (p. 125) is of little, if any, relevance to the descriptive hypothesis we have put forward. Nevertheless, we tried to clarify the preconditions of situations in which this can happen. The descriptive argument that 'under certain circumstances, people are indeed insensitive to sample size' (p. 125) is entirely consistent with the studies we have reviewed. The argument that 'the concept of a sampling distribution is equally essential for answering either of the two versions [i.e. sampling and frequency distribution tasks]' (p. 128) is incorrect, if 'essential' means necessary for solving the task.

However, the major issue remains how to explain the discrepant results in research on intuitions on sample size. Keren and Lewis conclude that we are still lacking the theoretical knowledge that enables us to predict under which circumstances people will or will not be sensitive to sample size. We do not think that one has to be so pessimistic. There are several variables that have been shown to influence sample size judgements to a certain extent (Keren and Lewis, 2000; Sedlmeier and Gigerenzer, 1997) and these variables might also be used in prediction. However, the strongest variable so far is the type of task: one can predict that frequency distribution tasks will be solved more easily than sampling distribution tasks. We have offered an explanation for why this is so. Human intuition conforms to the empirical law of large numbers, which helps much more in the solution of frequency than sampling distribution tasks.

The challenge is now either to find more evidence for this explanation or to come up with an alternative. Such an alternative does not seem to exist, so far, nor have Keren and Lewis proposed one. Meanwhile, we have elaborated our initial explanation into a computational model based on associative learning mechanisms, the *PASS* model (Sedlmeier, 1998, 1999). PASS models the cognitive processes underlying human probability judgement for the case of serially encoded events. Consistent with the empirical law of large numbers, PASS's confidence judgements increase with increasing sample size. In a single run, that is, in the process of being presented successive events, PASS's confidence estimates do not always increase strictly monotonically, similar to the example shown in Exhibit 1. The average result over many runs, however, conforms to the results shown in Exhibit 2. Because the input to PASS consists of serially presented events, it can only make judgements about frequency distributions and not sampling distributions. This model provides a possible explanation for why people's intuitions about sample size conform to the empirical law of large numbers.

Several hundred years ago, Jacob Bernoulli argued that 'even the stupidest man knows by some instinct of nature *per se* and by no previous instruction' that the greater the number of confirming observations, the surer the conjecture (Gigerenzer *et al.*, 1989, p. 29). Bernoulli may have been rather on the optimistic side, but in general the vast majority of people reason in accord with this intuition.

ACKNOWLEDGEMENTS

We thank Anita Todd for her comments. Special thanks go again to Berna Eden for her valuable contribution.

REFERENCES

Fischbein, E. *The Intuitive Sources of Probabilistic Thinking in Children*, Dordrecht: Reidel, 1975.

Freedman, D., Pisani, R., Purves, R. and Adhikari, A. *Statistics*, 2nd edn, New York: Norton, 1991.

Freudenthal, H. 'The "empirical law of large numbers" or "the stability of frequencies", *Educational Studies in Mathematics*, **4** (1972), 484–90.

Gigerenzer, G. 'The superego, the ego, and the id in statistical reasoning'. In Keren, G. and Lewis, C. (eds), *A Handbook for Data Analysis in the Behavioral Sciences: Methodological issues* (pp. 311–339), Hillsdale, NJ: Erlbaum, 1993.

Gigerenzer, G. 'On narrow norms and vague heuristics: A rebuttal to Kahneman and Tversky', *Psychological Review*, **103** (1996), 592–6.

Gigerenzer, G., Swijtink, Z., Porter, T., Daston, L., Beatty, J. and Krüger, L. *The Empire of Chance: How probability changed science and everyday life*, Cambridge: Cambridge University Press, 1989.

Kahneman, D. and Tversky, A. 'Subjective probability: A judgement of representativeness', *Cognitive Psychology*, **3** (1972), 430–54.

Keren, G. and Lewis, C. 'Even Bernoulli might have been wrong: A comment on intuitions about sample size', *Journal of Behavioral Decision Making*, **13** (2000), 125–132.

Rosnow, R. L. and Rosenthal, R. 'Statistical procedures and the justification of knowledge in psychological science', *American Psychologist*, **44** (1989), 1276–84.

Sedlmeier, P. 'The distribution matters: Two types of sample-size tasks', *Journal of Behavioral Decision Making*, **11** (1998), 281–301.

Sedlmeier, P. *Improving Statistical Reasoning: Theoretical models and practical implications*, Mahwah, NJ: Erlbaum, 1999.

Sedlmeier, P. and Gigerenzer, G. 'Intuitions about sample size: The empirical law of large numbers', *Journal of Behavioral Decision Making*, **10** (1997), 33–51.

von Mises, R. *Probability, Statistics and Truth*, New York: Dover, 1981. (Original work published 1957.)

Authors' biographies:
Peter Sedlmeier is currently a stand-in professor at the University of Münster and a Privatdozent at the University of Paderborn. His research includes models of frequency processing, tutorial systems for judgement under uncertainty, and probabilistic models of language processing. He is author of *Improving Statistical Reasoning: Theoretical models and practical implications* (Erlbaum, 1999).

Gerd Gigerenzer is director of the Center for Adaptive Behavior and Cognition at the Max Planck Institute for Human Development. He is the recipient of many prizes, including the AAAS Prize for Behavioral Science Research 1992. He studies bounded rationality, ecological rationality, and social rationality. His most recent book is *Simple Heuristics that Make us Smart* (with P. M. Todd and the ABC Research Group; Oxford University Press, 1999).

Authors' addresses:
Peter Sedlmeier, University of Paderborn, FB-2 Psychology, 33095 Paderborn, Germany.

Gerd Gigerenzer, Max Planck Institute for Human Development, Center for Adaptive Behavior and Cognition, Lentzeallee 94, 14195 Berlin, Germany.