Evidence Evaluation: Measure *Z* Corresponds to Human Utility Judgments Better than Measure *L* and Optimal-Experimental-Design Models

Patrice Rusconi

University of Surrey

Marco Marelli

University of Trento

Marco D'Addario

University of Milano-Bicocca

Selena Russo

The University of Sydney

Paolo Cherubini

University of Milano-Bicocca

Author Note

Patrice Rusconi, School of Psychology, University of Surrey; Marco Marelli, Center for Mind/Brain Sciences, University of Trento; Marco D'Addario, Department of Psychology, University of Milano-Bicocca; Selena Russo, School of Psychology, The University of Sydney; Paolo Cherubini, Department of Psychology, University of Milano-Bicocca.

The authors wish to thank Jonathan Nelson and two anonymous reviewers for their thorough comments and insightful suggestions on an earlier version of the manuscript.

Correspondence concerning this article should be addressed to Patrice Rusconi, School of Psychology, Faculty of Arts and Human Sciences, Room 27AC04, University of Surrey, Guildford, Surrey, United Kingdom, GU2 7XH. Phone: +44 (0)1483 689582. E-mail: p.rusconi@surrey.ac.uk

Abstract

Evidence evaluation is a crucial process in many human activities, spanning from medical diagnosis to impression formation. The present experiments investigated which, if any, normative model best conforms to people's intuition about the value of the obtained evidence. Psychologists, epistemologists, and philosophers of science have proposed several models to account for people's intuition about the utility of the obtained evidence with respect either to a focal hypothesis or to a constellation of hypotheses. We pitted against each other the so called optimal-experimental-design models (i.e., Bayesian diagnosticity, $\log_{10}$ diagnosticity, information gain, Kullback-Leibler distance, probability gain, and impact) and measures *L* and *Z* to compare their ability to describe humans' intuition about the value of the obtained evidence. Participants received words-and-numbers scenarios concerning two hypotheses and binary features. They were asked to evaluate the utility of "yes" and "no" answers to questions about some features possessed in different proportions (i.e., the likelihoods) by two types of extraterrestrial creatures (corresponding to two mutually exclusive and exhaustive hypotheses). Participants evaluated either how an answer was helpful or how an answer decreased/increased their beliefs with respect either to a single hypothesis or to both hypotheses. We fitted mixed-effects models and we used the Akaike information criterion (AIC) and the Bayesian information criterion (BIC) values to compare the competing models of the value of the obtained evidence. Overall, the experiments showed that measure *Z* was the best-fitting model of participants' judgments of the value of obtained answers. We discussed the implications for the human hypothesis-evaluation process.

Keywords: optimal-experimental-design models, measure *L*, measure *Z*, evidence evaluation, hypothesis testing.

Evidence Evaluation: Measure *Z* Corresponds to Human Utility Judgments Better than Measure *L* and Optimal-Experimental-Design Models

People are continually called upon to evaluate available information and to use it to determine which hypothesis under consideration is more appropriate. The evaluation of incoming information to confirm or revise prior beliefs is a ubiquitous phenomenon, that spans from the scrutiny of patients' symptoms when making a diagnosis (e.g., McKenzie, 2004) to interpreting information about an individual to infer her/his personality characteristics (e.g., Evett, Devine, Hirt, & Price, 1994). For example, a person might ask of a new acquaintance: "Do you often organize parties?" to learn about her/his extroversion. The target of the query might answer "yes, I do," and this response would convey a relatively different amount of information about her/his extroversion compared with the answer "no, I do not." In particular, the "yes" answer to the above question is relatively more diagnostic of the target's extroversion (a person who often organize parties is most likely extroverted) compared with the diagnosticity of the "no" answer regarding the target's introversion because a person who does not often organize parties might still be extroverted (e.g., Brambilla, Rusconi, Sacchi, & Cherubini, 2011, Study 2; Cameron & Trope, 2004; Cherubini, Rusconi, Russo, Di Bari, & Sacchi, 2010; Rusconi & McKenzie, in press; Rusconi, Sacchi, Toscano, & Cherubini, 2012; Sacchi, Rusconi, Bonomi, & Cherubini, in press; Sacchi, Rusconi, Russo, Bettiga, & Cherubini, 2012; Trope & Liberman, 1996; Trope & Thompson, 1997). Therefore, the tester should revise her/his prior confidence about the target's extroversion differently depending on whether she/he receives a "yes" or a "no" answer. A fair evaluation of the answers to a question (experiment results or test outcomes) is necessary to accurately revise initial beliefs with respect to a single hypothesis or to multiple hypotheses (Rusconi & McKenzie, in press; Slowiaczek, Klayman, Sherman, & Skov, 1992) and, eventually, for effective decision making. The present contribution investigated which, if any, of the different normative theories

advanced by statisticians, epistemologists, philosophers of science, and psychologists best described human intuition about the value of obtained evidence.

There is a large Bayesian reasoning literature on how people evaluate evidence to make judgments (e.g., Beach, 1968; Cherubini, Rusconi, Russo, & Crippa, 2013; Cosmides & Tooby, 1996; Fischhoff & Beyth-Marom, 1983; Gigerenzer & Hoffrage, 1995; Hammerton, 1973; McKenzie, 1994; Rusconi, Crippa, Russo, & Cherubini, 2012; Rusconi, Marelli, Russo, D'Addario, & Cherubini, 2013; Rusconi & McKenzie, in press; Slovic & Lichtenstein, 1971; Villejoubert & Mandel, 2002). The issue of whether people's intuitions about the value of obtained evidence align with theoretically optimal models has been the object of recent studies that noted the theoretical and empirical validity of two norms, namely, measures *L* and *Z* (e.g., Crupi, Tentori, & Gonzalez, 2007; Fitelson, 2001, 2006; Mastropasqua, Crupi, & Tentori, 2010; Tentori, Crupi, Bonini, & Osherson, 2007). Both norms quantify the changes in beliefs, as opposed to measures of the final, changed beliefs. However, it still remains to be clarified whether one of these two models should be preferred over the other in terms of descriptive adequacy. Indeed, although Crupi et al. (2007) found that *Z* was a slightly better predictor than *L*, Mastropasqua et al. (2010) found a reversed pattern whereby *L* performed slightly better than *Z*.

Furthermore, different normative standards, often called optimal-experimental-design (OED) models, have been proposed as models of human intuition about the value of information (e.g., Meder & Nelson, 2012; Nelson, 2005, 2008; Nelson, McKenzie, Cottrell, & Sejnowski, 2010). However, OED models have been investigated only with respect to their adequacy in describing human information gathering (e.g., Meder & Nelson, 2012; Nelson, 2005, 2008; Nelson et al., 2010). Therefore, further empirical investigation is needed to examine their relative descriptive power compared with measures *L* and *Z* as well as their absolute ability to capture human intuition in evidence evaluation, for example, when physicians scrutinize test outcomes,

when scientists interpret experiment results or, more generally, when one receives a "yes" or a "no" answer to a question.

Finally, an open question in the psychological literature on normative theories of evidence utility concerns whether a model of utility that yields negative values is preferable to a non-negative model or vice versa (e.g., Evans & Over, 1996; Nelson, 2008). Indeed, some metrics provide negative values for evidence utility even when the new evidence is informative (e.g., Evans & Over, 1996; Nelson, 2008). Some authors have argued that this property leads to counterintuitive predictions (e.g., Evans & Over, 1996; Nelson, 2008). The present experiments addressed these issues.

Normative models of the value of information

Several theories of the value of obtained information have been proposed in the philosophical and psychological literature as plausible models of human intuition. Among them are two inductive confirmation measures, namely, measures *L* and *Z* (e.g., Crupi, Tentori, & Gonzalez, 2007; Fitelson, 2001, 2006; Mastropasqua, Crupi, & Tentori, 2010; Tentori, Crupi, Bonini, & Osherson, 2007), and OED models (Nelson, 2005, 2008, 2009; Nelson et al., 2010), namely, Bayesian diagnosticity, $\log_{10}$ diagnosticity, information gain, Kullback-Leibler distance, probability gain, and impact. Although they differ in terms of how the utility of the obtained evidence is calculated, both of these classes of normative models are based on Bayes' rule, and thus, they involve prior probabilities (which express one's initial beliefs with respect to one or more hypotheses), likelihoods (which indicate the probability of the new evidence), and posterior probabilities (which express one's beliefs in light of the new evidence).

Two of the OED models, namely, Bayesian diagnosticity and $\log_{10}$ diagnosticity, are based solely on the likelihood ratio (LR), which is a constituent of Bayes' rule based on likelihoods and can be used to express the evidential strength of the new evidence (e.g., Good, 1950, 1983). Four other OED models often used in the psychological literature, namely, information gain, Kullback-

Leibler distance, probability gain, and impact, all entail the consideration of both prior and posterior probabilities (e.g., Nelson, 2005, 2008). Information Gain is a measure of uncertainty reduction (e.g., Evans & Over, 1996; Nelson, 2005, 2008; Nelson et al., 2010) that derives from Shannon's (1948) definition of entropy (or uncertainty) and was first suggested by Lindley (1956) as a method of quantifying the usefulness of an experiment before conducting it. Eimas (1970) was the first to use "a measure of the expected reduction in uncertainty in bits of information" (Eimas, 1970, p. 226) in an information-search task. The Kullback-Leibler distance is a measure of the distance between two distributions (e.g., Belov & Armstrong, 2011; Cover & Thomas, 1991; Kullback & Leibler, 1951). Probability Gain is a measure of error reduction, that is, it measures the extent to which an answer reduces the probability of preferring an erroneous hypothesis. Widely used by computer scientists, probability gain was first used as a normative and descriptive model in psychology by Baron (1985) (see also Baron, Beattie, & Hershey, 1988). Impact (also called absolute change, Nelson, 2005) is the absolute value of the difference between the posterior and the prior probability of a hypothesis. This metric has been used independently in different psychological studies (Klayman & Ha, 1987; Nelson, 2005; Nelson et al., 2010; Nickerson, 1996; Rusconi & McKenzie, in press; Wells and Lindsay, 1980). Whenever the prior probabilities of the hypotheses under exam are equal, the predictions of impact are identical to those of probability gain (e.g., Nelson, 2005, 2008; Rusconi & McKenzie, in press).

Measures *L* and *Z* are two confirmation measures recently advocated as normatively sound and descriptively adequate (e.g., Crupi et al., 2007; Mastropasqua et al., 2010; Tentori et al., 2007). Measure *Z* has been mostly discussed in the literature on expert systems and expert judgment (see Crupi & Tentori, 2010, footnote 9). It captures the notion of "*relative reduction of uncertainty*" (Crupi & Tentori, 2010, p. 7) because it quantifies how much a confirming or disconfirming piece of evidence reduces the distance between the prior probability of a hypothesis and the certainty that that hypothesis is true or false. Measure *L* is strictly connected to the log LR introduced as a

measure of the "weight of evidence" by Alan Turing (see Mastropasqua et al., 2010). Kemeny and Oppenheim (1952) proposed this measure as normatively sound, and Fitelson (2001, 2006) also argued in favor of it. We refer the reader to the Appendix for details on the calculation and the main properties of each of these metrics.

OED models and measures *L* and *Z* are formally and conceptually connected (Mastropasqua et al., 2010). However, most of the past research has described and used OED models as metrics that allow researchers to quantify the usefulness of the answers to a question (i.e., the new evidence) with respect to the *constellation* of two or more hypotheses under consideration. Going back to the example in the first paragraph, one could calculate the utility of a "yes" answer using one of the OED models and would thus consider the value of the "yes" answer with respect to both the introversion hypothesis and the extroversion hypothesis, that is, with respect to all of the hypotheses that could be considered in that example. In contrast, measures *L* and *Z* have been described as metrics designed to capture changes (i.e., increases or decreases) in the degrees of beliefs in a *particular* hypothesis brought about by a particular piece of evidence, that is, as inductive confirmation measures. Considering again the initial example, one could calculate the *L* or *Z* utility value of the "yes" answer to learn how that answer increases or decreases the plausibility of the extroversion hypothesis, that is, the tester considers the impact of the answer with respect to only one of the two possible hypotheses. In other words, whereas the values provided by the OED models are usually calculated with respect to both the hypothesis under consideration (i.e., the focal hypothesis) and its alternate/s, the values predicted by measures *L* and *Z* are usually calculated with respect to the focal hypothesis only. Previous studies have assumed this distinction between theories of the value of obtained evidence that refer to single vs. multiple hypotheses (e.g., Mastropasqua et al., 2010; Nelson, 2005). However, most of the metrics on which we shall focus in this article (with the exception of information gain and Kullback-Leibler distance, which cannot

refer to only one hypothesis) can be implemented in different fashions so that they can capture the different evaluation processes under examination (see Appendix).

Many readers might disagree that certain metrics can be used both as OED models and as confirmation measures: An example is probability gain, which, when it is treated as a confirmation measure, can no longer be defined as a measure of error reduction (see Appendix; the predictions of probability gain become identical to those of impact). However, the study by Wells and Lindsay (1980) represents a precedent. They showed that impact (which they called "informativeness" or "information gain") can be implemented with or without absolute values and thus convey different meanings. Consider again our initial example: The impact of a "no" answer ($\neg D$) to the question "Do you often organize parties?" on the hypothesis that the target person is extroverted (H) can be expressed as the absolute difference between the revised probability given the answer (i.e., the posterior probability) and the prior probability that the target person is extroverted, that is:

$$\left| p(H \mid \neg D) - p(H) \right|. \quad (1)$$

Suppose that your initial beliefs about the target were that she/he might be extroverted ($p(H) = .6$) and that the "no" answer induces you to slightly revise downward the prior probability, yielding a posterior probability of $p(H \mid \neg D) = .4$. In this case, Equation (1) yields an impact of .2.

Equation (1) can be used without absolute values, as follows:

$$p(H \mid \neg D) - p(H). \quad (2)$$

If we maintain the same values of prior and posterior probabilities given above, Equation (2) yields an Impact of -.2. Thus, Equation (1) allows researchers to gauge the absolute magnitude of changes in beliefs from prior to posterior probabilities engendered by the new evidence (in this example, the "no" answer). In contrast, Equation (2) provides the additional information of the direction (in this example, hypothesis-disconfirming) of that belief change, which is indicated by the algebraic operator + or - (Wells & Lindsay, 1980, pp. 778–779).

Under this light, OED models and measures *Z* and *L* have a kinship in that all of these metrics provide potentially optimal standards to interpret experimental or test results and in that all of them (with the exception of information gain and Kullback-Leibler distance) could potentially be used as inductive confirmation measures as well.

Overview

In the present experiments, we compared OED models with measures *L* and *Z* by testing the psychological plausibility of their main theoretical predictions. This investigation adds to the literature by deepening the study of which, if any, model best accounts for participants' intuitions in evidence evaluation. From one side, there is a lack of empirical investigation of the descriptive adequacy of OED models in evidence evaluation because past research has focused on which OED model best captures human information *acquisition* (e.g., Meder & Nelson, 2012; Nelson, 2005, 2008; Nelson et al., 2010). From the other side, further investigations are needed to determine which measure, *L* or *Z*, best describes human information evaluation. Furthermore, previous studies considered either OED models or measures *L* and *Z*, but none compared them. Therefore, there is a lack of empirical investigation on which of these two classes of competing models best corresponds to human intuition about the utility of obtained evidence.

To compare the relative strengths of the competing models, we tested them under different types of testing (single- vs. multiple-hypothesis testing) and for different types of evidence evaluation. Inductive confirmation measures such as *L* and *Z* typically quantify how much a piece of evidence confirms or disconfirms a single hypothesis, whereas OED models take into account the whole constellation of hypotheses (two or more). Accordingly, the type of evidence evaluation is also different. In the case of inductive confirmation measures, it is usually in terms of *decreases or increases in beliefs*, thus conveying information about the *direction* of the impact of a piece of evidence on one's beliefs with respect to a single hypothesis. In the case of the typical

implementation of OED models, the obtained evidence can be evaluated in terms of its *helpfulness*, thus reflecting the *magnitude* of its impact on one's beliefs with respect to all hypotheses.

Therefore, we devised different versions of the same task in which we took into account the type of testing and the type of evaluation required from participants. Across the experiments, we implemented the models that we compared in different ways (see Appendix).

Predictions

Based on previous implementations of OED models (e.g., Nelson, 2005, 2008; Nelson et al., 2010) and of measures *L* and *Z* (e.g., Crupi et al., 2007; Mastropasqua et al., 2010; Tentori et al., 2007), we might expect that OED models outperform measures *L* and *Z* in multiple-hypothesis testing even if *L* and *Z* are implemented so that they refer to more than one hypothesis (thus being potentially equivalent to OED models). Vice versa, we would expect measures *L* and *Z* to have an advantage over OED models in single-hypothesis testing even if OED models (except for information gain and Kullback-Leibler distance) are computed with respect to a single hypothesis (thus being treated as confirmation measures). This pattern of results would suggest that some norms are better suited to correspond to human intuition about the absolute change in beliefs engendered by obtained evidence, whereas others best describe human evaluations when the direction of that change is also considered.

However, it is also possible that the request to evaluate the obtained evidence in terms of either its helpfulness or its ability to increase or decrease one's beliefs might be captured differently by OED models and measures *L* and *Z*. This would reflect the effect of the type of evaluation. Finally, we could derive a third hypothesis: Neither the different implementations of a metric nor the different types of evaluation required would affect the descriptive adequacy of the metric. This result would indicate that there is a measure that corresponds to human intuition of the value of the obtained evidence in a consistent fashion. We conducted the following experiments to assess these hypotheses.

The experiments shared many methodological properties because, across experiments, we varied only the type of testing and the type of evaluation required from participants and, consequently, the models' implementations. Table 1 illustrates the differential properties of the experiments.

Experiment 1

Method

Participants

Ninety-five undergraduate students volunteered to participate in this study, 48 females and 47 males with a mean age of 22.33 years (*SD* = 2.14, range: 19–29). We refer the reader to Table 1 in Supplementary Materials for details on the sample characteristics of this and the other experiments we conducted.

Materials and procedure

We used a modified version of the planet Vuma scenario introduced by Skov and Sherman (1986) and thereafter widely used in the hypothesis-testing literature (Garcia-Marques, Sherman, & Palma-Oliveira, 2001; McKenzie, 2006; Nelson, 2005; Nelson et al., 2010; Rusconi & McKenzie, in press; Sacchi et al., 2012, Study 3; Slowiaczek et al., 1992; Villejoubert & Mandel, 2002). The prior probabilities of the hypotheses were set as not equiprobable to obtain divergent predictions from the OED models, which can lead to the same utility values when priors are equal (Nelson, 2005, 2008). Participants were given a booklet in which the following scenario was presented (the identifying information of the scenarios varied across experimental groups, which is reflected in the brackets):

> Imagine traveling to a planet called Vuma where there are only two types of creatures: Gloms and Fizos. Gloms comprise 25% [75%] of Vuma inhabitants, and 75% [25%] of inhabitants are Fizo. One cannot distinguish a Glom from a Fizo based

only on physical appearance. Your task is to identify the eight different creatures that you meet by chance.

Gloms and Fizos each possess certain features, and you will be told which percentages of each have these features. You can ask questions of the creatures that you encounter to determine whether they have a certain feature. Neither Gloms nor Fizos ever lie in replying to a question.

In the subsequent pages of the booklet, four features possessed by Gloms and Fizos (i.e., having gills, playing the harmonica, exhaling fire, and drinking gasoline) were presented to participants along with their probability distributions in the two groups (i.e., the likelihoods). On each page, participants were reminded of prior probabilities. An example of how such information was conveyed is as follows:

Below, you are provided with the percentages of Gloms and Fizos on the planet Vuma:

Glom	25% [75%]
Fizo	75% [25%]

The participants were then told the percentage of Gloms and Fizos possessing each of the four features. In keeping with Cherubini et al. (2013, Study 2), Rusconi et al. (2012, Study 3), and Rusconi and McKenzie (in press, Experiment 2), we presented participants with both the percentages of feature occurrences and the percentages of feature nonoccurrences. With this method, we could equate the computational steps required by the evaluation of the values associated with featural presence and those required by featural absence. Thus, we used the standard probability format, in which the relevant information is in the form of percentages (i.e., relative frequencies, see Gigerenzer & Hoffrage, 1995, p. 688), plus complementary likelihoods. An example is as follows:

In the table below, you are provided with the percentages of Gloms and Fizos with and without gills:

	With gills	Without gills
Glom	90%	10%
Fizo	50%	50%

It should be noted that this standard probability format plus complementary likelihoods differs from the probability formats described by Gigerenzer and Hoffrage (1995) because it provides the additional information attached to featural absence (¬*D*). This aspect might also be an advantage compared with the frequency formats described by Gigerenzer and Hoffrage (1995), which focus on featural presence. Thus, although we used a words-and-numbers scenario, which is considered less effective in facilitating Bayesian reasoning than experience-based learning (e.g., Meder & Nelson, 2012; Nelson, 2009; Nelson et al., 2010), this format might still be more useful for participants than the traditional ones.

Participants were told to imagine meeting a creature and asking her a question (e.g., "Do you have gills?"). They were presented with the answer (either a "yes" or a "no"), and they were asked to provide an answer utility rating for each encountered creature[1]. In particular, participants were asked (variations are in brackets): "How do you deem that the received answer ("YES") [("NO")] decreases/increases the plausibility of the hypothesis that the encountered creature is a **Glom** [**Fizo**]? (mark a number from -3 to 3)." The labels under the endpoints of the scale were *definitely decreases* and *definitely increases*. The original instructions in Italian for all of the experiments presented in this article are available at the end of Supplementary Materials.

In terms of the two variables (i.e., type of testing and type of evaluation) that we considered to devise the experiments, this formulation reflects single-hypothesis testing and evidence

evaluation in terms of decreases and increases in beliefs (see Table 1). We calculated the answer utility values predicted by each competing model with respect to the focal hypothesis (either the Glom hypothesis or the Fizo hypothesis according to the experimental group) (see Table 2). Recall that the answer utility values calculated in terms of information gain and Kullback-Leibler distance cannot refer to a single hypothesis. For this reason, their predicted answer utility values were identical for the Glom and the Fizo hypotheses. Furthermore, with this implementation, probability gain and impact always produce identical answer utility values (see Table 2).

Each feature was presented twice to each participant: First, the feature was followed by a "yes" answer, and the second time, it was followed by a "no" answer, or vice versa. The order of the eight feature-answer combinations was randomized. Accordingly, we devised a 4 (feature percentage combination: 90%-50%, 75%-15%, 45%-85%, 22%-98%) × 2 (answer: yes vs. no) × 2 (focal hypothesis: Glom vs. Fizo) × 2 (prior probability associated with the focal hypothesis: .25 vs. .75) mixed design, with the first two factors as within-participant variables and the latter two factors as between-participant variables (Table 2 shows the formal properties of the 32 conditions). We chose the values of prior probabilities and feature combinations to present to participants randomly, with the following two general caveats: 1) that these values should not have properties that could influence participants' judgments in a systematic fashion (as would have been the case, for example, if all presented percentages had been very extreme); 2) that each feature had a different percentage combination (and a different complementary-likelihood combination) from every other feature.

We then asked participants to provide personal data (i.e., sex, age, nationality, and course of study). Individuals were approached in quiet places and study rooms within the different buildings of the University of Milano-Bicocca and were given the booklet if they agreed to participate. There were no time constraints. Upon completion of the booklet, they were thanked and debriefed.

Data analysis

Pearson and Spearman correlations are more sensitive to sample numerosity than other techniques; they do not inform researchers about the cause-effect direction, and they might have limited power because they cannot be performed on all data points but only on data collapsed across participants. Accordingly, we analyzed participants' ratings using linear mixed-effects models with the open source software environment R (Bates, Maechler, & Bolker, 2012; R Development Core Team, 2012; R version 2.15.0, 2012-03-30). This technique ensures statistical robustness and has the advantage of taking into account individual differences. This property is relevant to our study because previous research has shown that there may be variability in individuals' responses to judgment tasks (e.g., Gigerenzer & Hoffrage, 1995; Rusconi et al., 2013; Villejoubert & Mandel, 2002). Furthermore, this technique ensures a more reliable estimation of the fixed effects. We fit a different mixed-effects model for each of the theories of the value of information that we considered (instead of a single mixed-effects model that included all eight competing models as potentially significant predictors) because of the overall high correlations among their predicted values (see Table 2 in Supplementary Materials). Had we fit a single model, we would have had a problem of collinearity.

For each linear mixed-effects model that we fit, we included one of the eight competing models (i.e., one among OED models, measure *L*, and measure *Z*) as the fixed effect. As to the random-intercepts structure, we considered including the effects of the participants, their courses of study, and the 32 experimental conditions on the intercept. We removed the candidate random intercepts that did not significantly contribute to the goodness of fit of the mixed-effects model. In particular, the analysis started with a full factorial model including the fixed effect and all three potentially significant random intercepts. We then tested the candidate random intercepts one by one. We considered superfluous, and thus removed, the potential random intercept when the result of the LR test comparing the goodness of fit of the models before and after removing it was not significant. We refitted all final mixed-effects models after excluding the outliers with a

standardized residual at a distance greater than 2.5 standard deviations from zero (e.g., Baayen, 2008). For each mixed-effects model, we ascertained the statistical significance of the fixed effect using a Markov chain Monte Carlo (MCMC) sampling algorithm with 10,000 samples.

Beyond the analysis of significance, we wanted to acquire more insights into the relative strength of each competing model. Therefore, we used the Akaike information criterion (AIC, Akaike, 1974) and the Bayesian information criterion (BIC) to compare and rank the eight competing models (Wagenmakers & Farrell, 2004). Both the AIC and the BIC are criteria for model selection—that is, they allow researchers to identify the best model within a set of models that are a priori theoretically relevant. The model for which the AIC (or the BIC) is minimal is to be preferred and is considered the best among the candidate models for the data at hand. Once the best model is identified, it is possible to rank all of the candidate models by means of the Δ_i (e.g., Burnham & Anderson, 2001; Wagenmakers & Farrell, 2004), that is, the difference between a particular candidate model's AIC (or BIC) and the best model's AIC (or BIC). Accordingly, the best model's Δ_i (AIC) and Δ_i (BIC) are equal to zero. Burnham and Anderson (2001) provide some rules of thumb to gauge the relative strength of each model in the candidate set based on Δ_i (AIC) values. Models with $\Delta_i(\text{AIC}) \leq 2$ have substantial strength of evidence, and those with $4 \leq \Delta_i(\text{AIC}) \leq 7$ have considerably less strength, whereas models with $\Delta_i(\text{AIC}) > 10$ essentially have no strength.

A more refined method for scaling models is based on the conditional probabilities of the models given the data (e.g., Burnham & Anderson, 2001; Wagenmakers & Farrell, 2004). In particular, it is possible to calculate the relative probability of a model *i* versus all other models in the candidate set using the following equation:

$$w_i = \frac{\exp\left(-\frac{1}{2}\Delta_i\right)}{\sum_{k=1}^{K}\exp\left(-\frac{1}{2}\Delta_k\right)}. \tag{3}$$

w_i is a normalized ratio called an Akaike weight when Δ_i and Δ_k refer to the AIC values and a Schwarz weight when they refer to the BIC values. Akaike and Schwarz weights provide a measure of the strength of evidence of model *i* relative to the candidate set for the data at hand. It is then possible to compute the evidence ratio of both Akaike and Schwarz weights for model *i* over model *j*, that is, w_i / w_j , or its normalized probability, that is, $w_i / (w_i + w_j)$ (Wagenmakers & Farrell, 2004). This ratio provides a measure of how much the evidence favors model *i* over model *j*. In the present contribution, we shall report the results of both AIC-based and BIC-based model selection because both criteria have advantages and disadvantages (e.g., Wagenmakers & Farrell, 2004).

Results

Figure 1 shows eight scatterplots (one for each competing model) illustrating the relationships between theoretical and observed answer utility values (averaged across the 32 experimental conditions). Table 3 in Supplementary Materials shows participants' mean ratings and the standard errors of the mean. Table 3 reports the estimated parameters, the 95% highest posterior density (HPD) intervals (a Bayesian equivalent of the 95% confidence intervals), their statistical significance, and the AIC and BIC values for each of the eight models that we fit. In terms of statistical significance, all linear mixed-effects models had *p*MCMC ≤ .0002 except for those including as fixed effects information gain, *p*MCMC = .6574, and Kullback-Leibler distance, *p*MCMC = .4750. As to the random-intercepts structure, only the final model that included Kullback-Leibler distance as the fixed effect did not include the random intercept of participants because it did not significantly improve the goodness of fit of the mixed-effects model, $\chi^2(1) = 0$, $p = .9995$. The seven other final models included all three random intercepts that we considered (the effects of the participants, their courses of study, and the 32 experimental conditions on the intercept) because they were significant (or marginally so) on the LR tests, χ^2s(1) ≥ 3.11, *p*s ≤ .0779.

Clear-cut results emerged from the analysis of the AIC and BIC values. Measure *Z* had both Δ_i (AIC) and Δ_i (BIC) = 0 and both Akaike and Schwarz weights = 1. This finding indicates that measure *Z* was the best-fitting model within the candidate set and that none of the other competing models had support, with both Δ_i (AIC) and Δ_i (BIC) $\geq$ 28 and both Akaike and Schwarz weights = .00. The overwhelming advantage of measure *Z* over the other models in the candidate set in terms of descriptive adequacy is confirmed by the values of the evidence ratios and the normalized probabilities reported in Table 4. All contrasts favored measure *Z*, with evidence ratios of both Akaike and Schwarz weights yielding very high values, indicating that *Z* was at least one million times more likely to be the best model than any other model. The next-best model was measure *L*, whereas information gain, Kullback-Leibler distance, and Bayesian diagnosticity were the worst models in the candidate set (see the Δ_i (AIC) and Δ_i (BIC) values in Table 3). Together, measure *Z* and measure *L* outperformed all six OED models more than ninety-eight million times in terms of fit to participants' answer utility ratings (see Table 4).

Discussion

Experiment 1 provided clear-cut evidence for the ability of measure *Z* to capture participants' intuitions about the value of an obtained answer. Measure *L* was the second-best model, and Kullback-Leibler distance and Bayesian diagnosticity were among the worst models. The type of testing and the type of evaluation that we required from participants in Experiment 1 were those typically used in tasks that investigate inductive confirmation measures. In particular, the formulation of our request was in terms of how a piece of information decreased or increased the plausibility of a hypothesis; thus, we mimicked the labels of the impact scale used in Mastropasqua et al. (2010). Indeed, Mastropasqua et al. (2010) asked the participants to estimate inductive confirmation using an impact scale with labels such as "The information INCREASES the plausibility of the hypothesis" or "The information DECREASES the plausibility of the hypothesis." Furthermore, participants in those experiments had to express their judgments with

respect to a single hypothesis. However, Mastropasqua et al. (2010) found measure *L* to be superior both when using numbers (in Experiment 1) and when using pictures (in Experiment 2) to convey the uncertainty of the new evidence, whereas we used only a words-and-numbers scenario with the standard probability format plus complementary likelihoods. This difference in the materials used might partly account for the different results.

In contrast, our finding is similar to the one that emerged in Crupi et al. (2007). It is important to note that Tentori et al. (2007) and Crupi et al. (2007) used an urn problem, which is similar to our scenario: Participants were asked to judge the impact of 10 extractions without replacement on the hypothesis that urn A versus urn B was selected. Judgments were requested on a 7-point scale that ranged from "weakens my conviction extremely" to "strengthens my conviction extremely." Despite the similarity in the tasks, we found a clear-cut difference in the descriptive ability between measures *Z* and measure *L*. This difference was greater than that found in Crupi et al.'s (2007) study, in which the descriptive power of *L* was not so distant from that of *Z* (see Crupi et al., 2007, Table 4).

Beyond the different tasks used, a possible explanation of the differences between our experimental results and previous findings might be that the linear mixed-effects models that we used ensured greater statistical robustness than the correlations used in Tentori et al. (2007), Crupi et al. (2007), and Mastropasqua et al. (2007) (see the Data analysis section).

Another original aspect of Experiment 1, compared with previous studies in the literature, is that we computed the answer utility values predicted by the OED models with respect to a single hypothesis. That is, we calculated the OED models differently from their typical implementations, which refer to the whole constellation of hypotheses. We did this to allow a comparison between the OED models and measures *L* and *Z* as well as to compare the models' ability to capture the hypothesis-confirming vs. hypothesis-disconfirming direction of the belief change engendered by the obtained evidence. The exceptions were information gain and Kullback-Leibler distance, which

cannot refer to a single hypothesis. Indeed, they were introduced as fixed effects in the only two mixed-effects models that were not statistically significant. Although we attempted to make the other OED models more similar to inductive confirmation measures such as *L* and *Z*, the relative strengths of these OED models compared with those of measures *L* and *Z* were poor.

These results are the first in the literature that emerged from an empirical comparison of measures *L* and *Z* and OED models. Furthermore, these results are also the first to give insights in absolute terms because the analysis of significance suggested that two of the OED models, namely, information gain and Kullback-Leibler distance, might be poor predictors of human judgments of the value of the obtained evidence in this type of task and with this task requirement, whereas all other competing models might be good predictors.

Experiment 2

In Experiment 2, we used the same task as in Experiment 1, but we rephrased the question on answer utility so that we could assess the adequacy of the models when participants had to use another type of evaluation, that is, an evaluation in terms of the helpfulness of the received answer (see Table 1).

Method

Participants

Ninety-two undergraduate students of the University of Milano-Bicocca volunteered to participated in this study, 72 females and 20 males with a mean age of 20.6 years (*SD* = 2.02, range: 18-33).

Materials and procedure

The materials and procedure were identical to those used in Experiment 1. The only difference was in task requirement. The answer utility rating was required from participants in the following way (variations are in brackets): “How do you deem that the received answer (“YES”) [(“NO”)] is helpful for ascertaining the possibility that the encountered creature is a **Glom** [**Fizo**]?

(mark a number from -3 to 3)". The labels under the endpoints of the scale were *definitely useless* and *definitely useful*. As in Experiment 1, we calculated the theoretical values of answer utility so that they could convey the confirmatory or falsificatory utility of the received answer with respect to a *single* hypothesis (see Table 2). Recall that information gain and Kullback-Leibler distance cannot be implemented relative to a single hypothesis, and thus, their predicted values were identical for the Glom and the Fizo hypotheses (see Table 2). Furthermore, according to this implementation, probability gain and impact always yield the same answer utility values (see Table 2).

Results

We analyzed participants' ratings following the same statistical procedures used in Experiment 1. We did not fit a single linear mixed-effects model that included all of the competing models as fixed effects because their theoretical values were highly correlated (see Table 2 in Supplementary Materials). Table 3 in Supplementary Materials shows the mean participants' ratings and the standard errors of the mean. The eight scatterplots in Figure 2 illustrate the relationships between participants' ratings and the values predicted by each model averaged across the 32 experimental conditions. Table 5 shows the results of the eight linear mixed-effects models that we performed. The random-intercepts structure was identical for the eight mixed-effects models and differed with respect to Experiment 1. In particular, we excluded the random effect of participants' course of study on the intercept because it did not significantly contribute to the goodness of fit of the mixed-effects models, χ^2s(1) ≤ .84, *p*s ≥ .3605. In contrast, we included in all final models the effects of the participants and of the 32 experimental conditions on the intercept because they significantly improved the goodness of fit of the models, χ^2s(1) ≥ 19.02, *p*s ≤ .0001. The analysis of significance revealed that two linear mixed-effects models involving two OED models as fixed effects were not significant, namely, Kullback-Leibler distance, *p*MCMC = .2742,

and information gain, *p*MCMC = .6042. In contrast, all other mixed-effects models were statistically significant, *p*MCMC ≤ .0022.

The analysis of the AIC and BIC values revealed that measure *Z* was the best-fitting model within the candidate set, as shown by the values of both Δ_i (AIC) and Δ_i (BIC) being equal to 0. The next-best model was measure *L*, whose Δ_i (AIC) and Δ_i (BIC) values were both 8. The evidence ratios of both Akaike and Schwarz weights favored measure *Z* more than fifty-four times over measure *L*, giving normalized probabilities of .98 (see Table 6). All OED models had Δ_i (AIC) and Δ_i (BIC) > 10, indicating the poor support they received from the data. The superiority of measure *Z* over the OED models in fitting the data is clear from the evidence ratios and normalized probabilities in Table 6. Measure *Z* outperformed both probability gain and impact more than six hundred times and all other OED models more than ten thousand times. Together, measures *L* and *Z* outperformed all six OED models more than three hundred times. The normalized probabilities were 1 in all contrasts involving measure *Z* and OED models.

Discussion

Measure *Z* was the best-fitting model and measure *L* was the second-best model in Experiment 2, in which we asked participants to evaluate the received answer in terms of its helpfulness with respect to testing a single hypothesis (either the Glom or the Fizo hypothesis). As in Experiment 1, we implemented OED models so that they could be considered additional inductive confirmation measures (which refer to a single hypothesis). Again, the exceptions were information gain and Kullback-Leibler distance, which were introduced as fixed effects in the only two mixed-effects models that were not statistically significant. The lack of significance of these two OED models replicates that found in Experiment 1. The analysis of AIC and BIC values revealed that the different implementations of the other OED models did not achieve an adequate description of participants' answer utility ratings compared with measures *L* and *Z*.

These findings reveal a pattern consistent with the results from Experiment 1. However, the relative distances between the scaled models were more pronounced in Experiment 1 than in Experiment 2. This finding might reflect the overall greater ability of measures *Z* and *L* to capture human intuition about the value of obtained evidence in terms of decreases or increases in beliefs (Experiment 1) rather than in terms of evidence helpfulness (Experiment 2). Indeed, this was the only difference in task requirement between Experiment 1 and Experiment 2, with the type of testing (single-hypothesis testing in both experiments) having been left unchanged.

Experiment 3

In the first two experiments, we asked participants to judge the impact of a piece of evidence (i.e., the "yes" or "no" answer to a question) on a single hypothesis (i.e., the membership of an imaginary creature in a group). The crucial innovation in Experiment 3 was that we requested that participants respond with utility ratings with respect to two hypotheses (i.e., the membership of the encountered creature in one category *and* in the alternative category). In terms of the type of evaluation, we asked participants to evaluate the helpfulness of the received answer, as in Experiment 2 (see Table 1).

Method

Participants

One hundred and two undergraduate students of the University of Milano-Bicocca volunteered to participate in this study, 51 females and 50 males with a mean age of 21.63 years (*SD* = 1.99, range: 18–29).

Materials and procedure

The materials and procedure were identical to those used in the previous experiments. The only difference was in task requirement. The answer utility rating was required from participants in the following way (variations are in brackets): "How do you deem that the received answer ("YES") [("NO")] is helpful in distinguishing between the possibility that the encountered creature

is a Glom and the possibility that it is a Fizo? (mark a number from -3 to 3)". The labels under the endpoints of the scale were *definitely useless* and *definitely useful*.

In terms of the two variables that we took into account to devise the experiments, namely, the type of testing and the type of evaluation, this task requirement aimed to capture multiple-hypothesis testing (participants had to consider both the Glom and the Fizo hypotheses) and the evaluation of the helpfulness of the obtained evidence (i.e., the magnitude of its impact). Accordingly, the algebraic operator of the theoretical utility values did not indicate hypothesis confirmation or hypothesis falsification. The only models that also predicted negative utility values were information gain and probability gain, with a meaning of uncertainty increase and error increase, respectively (see Table 7).

Results

In Figure 3, there are eight scatterplots (one for each of the competing models that we considered) in which the mean ratings (averaged across the 32 cells) are compared with the normative predictions. Table 3 in Supplementary Materials shows the mean ratings (and the standard errors of the mean) provided by participants across the 32 cells of the experimental design. As in the previous experiments, we did not fit a single linear mixed-effects model that included all of the competing models as fixed effects because their theoretical values were overall highly correlated (see Table 4 in Supplementary Materials). Table 8 shows the estimates of the fixed effects resulting from the linear mixed-effects models that we fit, their statistical significance, and the AIC and BIC values. The inclusion of each of the random intercepts that we considered significantly improved the goodness of fit of all models, χ^2s(1) $\geq$ 12.59, *p*s $\leq$ .0004. The fixed effects were significant predictors in all models (except for impact, which fell slightly short of significance, *p*MCMC = .0522). Looking at the Δ_i (AIC) and Δ_i (BIC) values reported in Table 8, it emerges that measure *Z* was the only model with $\Delta_i \leq 2$ (considering either the AIC or the BIC values), that is, the only one that received substantial support and was thus the best fit to the data

according to Burnham and Anderson (2001)'s rule. Measure *L*, information gain, and probability gain, received less support, and Bayesian diagnosticity, impact, and $\log_{10}$ diagnosticity essentially had no support. These findings are corroborated by the scaling method based on conditional probabilities. Table 8 reports Akaike and Schwarz weight values greater than .75 for measure *Z*, equal to .17 for measure *L*, and less than .04 for the other candidate models. Recall that these values are normalized probabilities (i.e., they sum to 1) that can be interpreted as the likelihoods of the models to best fit the data at hand given a set of candidate models.

Table 9 shows the results of the "pair-wise" evidence ratios based on conditional probabilities. It emerged that measure *Z* was 4.48 times more likely to be the best model than was the second-best model, measure *L* (normalized probability of .82). Measure *Z*'s greater fit to the data was even more clear-cut compared with all other competing models. Measure *Z* was twenty to fifty-five times more likely to be the best predictor than information gain and probability gain. It outperformed Bayesian diagnosticity more than one hundred times, and it was a more than one thousand times better-fitting model than impact, Kullback-Leibler distance, and $\log_{10}$ diagnosticity (normalized probabilities of 1). Overall, measures *L* and *Z* were more than 10 times better than the 6 OED models at predicting participants' ratings (normalized probabilities $\geq .93$).

Discussion

Experiment 3 showed that measure *Z* was the best predictor of participants' ratings in the candidate set. This is the first empirical investigation that implements measure *Z* in a different way compared with the traditional implementation (see Appendix and Table 7) so that it can capture multiple-hypothesis testing and the evaluation of the helpfulness of the obtained evidence. Therefore, our findings extended measure *Z*'s field of application and showed its ability to correspond to participants' estimates of the impact of the acquired evidence on the constellation of two hypotheses.

As in the previous experiments, measure *Z* outperformed measure *L*, in keeping with Crupi et al. (2007) and contrary to Mastropasqua et al. (2010). Indeed, the evidence ratio of both Akaike weights and Schwarz weights for *Z* over *L* was 4.48, and the normalized probability was .82 (see Table 9). However, measure *L* was also a good predictor, in line with both Crupi et al. (2007) and Mastropasqua et al. (2010). Indeed, together, *Z* and *L* outperformed all 6 OED models (evidence ratios greater than 10 and normalized probability higher than .9). This finding is relevant because we asked participants to evaluate the received answers with respect to both hypotheses, and OED models have usually been conceived exactly as metrics that quantify answer utility with respect to two or more hypotheses rather than a single hypothesis. However, it should be noted that contrary to the two previous experiments, all OED models were statistically significant predictors (or marginally significant in the case of impact). This finding indicates that they were still good models in absolute terms and that the task requirement of Experiment 3 overall improved their ability to correspond to participants' ratings. It is interesting to note that among the OED models, information gain and probability gain were the best-fitting models and $\log_{10}$ diagnosticity was the worst. This finding, relative to the set of OED models, echoes the results from the studies on information gathering (Nelson, 2005; Nelson et al., 2010). Indeed, Nelson (2005), who used a planet Vuma scenario similar to ours, found that $\log_{10}$ diagnosticity, along with Bayesian diagnosticity, was empirically inadequate (beyond being theoretically flawed). He also found that information gain was the best predictor of participants' question utility ratings, followed by probability gain. Nelson et al. (2010) found that probability gain was the best model when participants experienced environmental probabilities in a simulated-plankton-categorization task. Furthermore, participants exhibited a tendency to prefer to inquire about the feature with higher Iinformation gain when the task used the planet Vuma scenario.

We conducted an additional experiment (which we shall refer henceforth to as the Additional Experiment). As shown in Table 1, we required from participants an evaluation in terms

of decreases or increases in beliefs with respect to two hypotheses in the Additional Experiment. This evaluation/testing combination produced a phrasing of the task requirement that might have appeared odd to participants for pragmatic-conversational reasons. This might have been the reason the results partially differed from those of the three previous experiments. In particular, three models outperformed the others in the candidate set: Probability gain, information gain, and measure *Z*. According to the BIC, all three of these models scored equally well, with probability gain preferred over information gain and measure *Z* according to the AIC. Both model selection criteria agreed in selecting measure *L* as the next-best model and impact, $\log_{10}$ diagnosticity, and Bayesian diagnosticity as the worst models in the candidate set.

We refer the reader to Tables 5–6 and Figure 2 in Supplementary Materials for the results of this Additional Experiment.

General Discussion

The experiments presented in this article provided consistent evidence for the ability of measure *Z* to predict participants' naïve judgments of answer utility better than 7 other obtained evidence value models. Measure *Z*, along with measure *L*, is an inductive confirmation metric recently advanced as a normatively appealing and empirically good approximation of human intuition about the impact of a datum on one's beliefs about the plausibility of a hypothesis (Crupi et al., 2007; Mastropasqua et al., 2010; Tentori et al., 2007). However, there is still debate on which, between *L* and *Z*, best accounts for human confirmation judgments (see Mastropasqua et al., 2010, p. 949). The 6 other models that we included in the candidate set were OED models: Bayesian diagnosticity, $\log_{10}$ diagnosticity, information gain, Kullback-Leibler distance, probability gain, and impact. Some of the OED models have been used in recent studies of human evidence evaluation. For example, Rusconi and McKenzie (in press) used impact and probability gain; Cherubini et al. (2013) used information gain and probability gain; Rusconi et al. (2013) used information gain and $\log_{10}$ diagnosticity (in *decibans*, see Appendix); Rusconi et al. (2012) used

information gain; and Cherubini, Russo, Rusconi, D'Addario, and Boccuti (2009) used information gain and $\log_{10}$ diagnosticity (the latter expressed in *decibans*). However, past research on human evidence evaluation lacked an empirical investigation of which, if any, of the OED models best accounted for people's judgments. Furthermore, there was a lack of empirical comparisons among measures *L* and *Z* and OED models. Our experiments provided the first answer to these issues.

Our experiments added to the extant literature by taking into account the type of testing (single-hypothesis vs. multiple-hypothesis testing) and the type of evaluation (assessment of evidence helpfulness vs. evidence's ability to increase or decrease one's beliefs). Furthermore, we implemented the models in different ways compared with their traditional implementations in the literature to pit them against one another in both single- and multiple-hypothesis testing. Whichever combination of testing (and thus of mathematical implementation) and evaluation type we used, measure *Z* was the best-fitting model, and measure *L* was the next-best model. In contrast, the OED models, in particular Bayesian diagnosticity, Kullback-Leibler distance, and $\log_{10}$ diagnosticity, were consistently poorer predictors of participants' answer utility ratings. Only one combination of testing/evaluation type yielded a more nuanced pattern of results, whereby probability gain, measure *Z*, and information gain scored almost equally well. This occurred in the Additional Experiment, for which we argued that participants might have encountered difficulties with the phrasing of the task requirement. In this case, the outperformance of measure *Z* might have been muted for pragmatic-conversational reasons.

The advantage of measure *Z* over the competitors was greater when the task required single-hypothesis testing, that is, in Experiments 1–2, in which the normalized probabilities of both Akaike and Schwarz weights were $\geq .98$, than in experiments with multiple-hypothesis testing. In particular, the descriptive adequacy of measure *Z* had its highest peak in Experiment 1, in which participants evaluated the obtained answer in terms of how much it decreased or increased the plausibility of the focal hypothesis. In contrast, the OED models, in particular probability gain and

information gain, performed better in tasks of multiple-hypothesis testing (Experiment 3 and the Additional Experiment) than in those of single-hypothesis testing.

This pattern of findings suggests that although measure *Z* fits well under different mathematical implementations (that is, both in single-hypothesis and multiple-hypothesis testing) and types of evaluation, its advantage is greater when used as a typical inductive confirmation metric, that is, when it measures the increases and decreases in one's beliefs with respect to a single hypothesis. In a similar way, although OED models are overall poorer predictors of human intuition in evidence evaluation than *Z*, they better fit the observed judgments when they are implemented in a way that captures changes in one's beliefs with respect to a constellation of two or more hypotheses.

Negative utility

An important issue concerning the theories of evidence utility is represented by negative vs. non-negative utility (e.g., Evans & Over, 1996; Nelson, 2008). Nelson (2008) discussed this issue with reference to OED models. For example, negative utility distinguishes Kullback-Leibler distance, which is a non-negative metric, from information gain, which assumes negative values whenever there is an increase in uncertainty about the hypotheses after the receipt of new evidence. Recall that both metrics cannot be implemented with respect to a focal hypothesis but only with respect to a constellation of two or more hypotheses. In this sense, Experiment 3 and the Additional Experiment, in which participants were required to evaluate evidence with respect to two hypotheses, provided the most useful data for comparing the psychological plausibility of these two models. The results of both experiments favored information gain over Kullback-Leibler distance: The evidence ratios of both Akaike and Schwarz weights were 90.02 (giving normalized probabilities of .99 in favor of information gain). Should one conclude that a model of the obtained evidence that allows for negative utility values, such as information gain, is more descriptively adequate than a non-negative metric, such as Kullback-Leibler distance? Although information gain

outperformed Kullback-Leibler distance in our multiple-hypothesis testing tasks, overall, its descriptive adequacy is called into question. Metrics such as *Z* and *L* consistently outperformed information gain in terms of describing participants' answer utility ratings, and this also held true when *Z* and *L* were implemented as non-negative measures (i.e., in Experiment 3 and the Additional Experiment; see Table 7. See the discussion above for the interpretation of the results that emerged from the Additional Experiment).

However, do people value some given evidence in terms of negative utility values? Nelson (2008) set the issue in these terms: "It is ultimately an empirical question whether people experience some information as having zero or negative utility" (Nelson, 2008, p. 154). The answer provided by our experimental results is affirmative. Looking at the descriptive values reported in Table 3 in Supplementary Materials and shown in Figures 1–2, it emerges that when people are given the possibility to provide negative ratings (our rating scale ranged from -3 to 3), they do sometimes conceive the utility of an obtained datum in terms of negative values. In particular, on average, negative utility ratings, when provided, appeared to be associated with single-hypothesis testing (Figures 1–2) and not with multiple-hypothesis testing (e.g., Figure 3). Accordingly, a model of the utility of obtained evidence that allows for negative utility values in single-hypothesis testing but not in multiple-hypothesis testing, such as *Z*, appears to be a descriptively useful criterion to gauge human intuition.

Implications for model selection in evidence evaluation and information gathering

The results of the present experiments bear implications for the selection of the normative criteria against which human judgments on these types of tasks ought to be compared. In an article on Turing's statistical work during World War II, Good (1979) stated "A deciban or half-deciban is about the smallest change in weight of evidence that is directly perceptible to human intuition" (Good, 1979, p. 394). Cherubini et al. (2009) found that non-experts engaged in hypothesis-testing tasks with abstract materials could perceive evidence if its impact was at least .12–.18 *bits*

(corresponding to 3.5–4.5 *decibans* in terms of $\log_{10}$ diagnosticity). However, experts (physicians) were able to perceive the clue informativeness (symptoms) even when its impact was only .03 *bits* or 1.8 *decibans* (see Cherubini et al., 2009, pp. 560–561). The results of our experiments raise the question of whether a more psychologically plausible model than $\log_{10}$ diagnosticity or information gain, namely, measure *Z*, might reveal an even more sophisticated ability of human intuition to perceive minimal changes in evidence informativeness. Indeed, it has been shown that the model selected by the experimenter as the normative criterion to gauge people's behavior can be decisive for the interpretation of experimental results. For example, in his revision of the literature on information gathering, Nelson (2005) noted that using either information gain or impact instead of probability gain in the study by Baron et al. (1988) would have caused the information bias (i.e., the tendency to choose a test as useful although it is normatively worthless) found by the authors to largely disappear (see Nelson, 2005, p. 985).

Our findings relative to human evidence *evaluation* might also suggest further lines of research on human information *gathering*. Previous literature has shown that OED theories are plausible models of human information acquisition (e.g., Nelson, 2005; Nelson et al., 2010). However, these studies did not include measures *L* and *Z* in the candidate set. Accordingly, further development of such work might include testing the psychological plausibility of metrics such as *L* and *Z* in the form of expected *L* and expected *Z*. More precisely, expected *L* and expected *Z* could be computed as the sum of the diagnosticities of the "yes" and "no" answers, each weighted for their probability of occurrence (e.g., Rusconi & McKenzie, in press, footnote 4).

Information format

In our experiments, we used a modified version of the planet Vuma scenario, a words-and-numbers vignette often used in the Bayesian reasoning literature (Garcia-Marques, Sherman, & Palma-Oliveira, 2001; McKenzie, 2006; Nelson, 2005; Nelson et al., 2010; Rusconi & McKenzie, in press; Sacchi et al., 2012, Study 3; Skov & Sherman, 1986; Slowiaczek et al., 1992; Villejoubert

& Mandel, 2002). In particular, we used the standard probability format (i.e., percentages) to convey the distributions of Gloms and Fizos on Vuma as well as their feature distributions. It has been shown that presenting probabilistic information in words-and-numbers formats is not meaningful for inductive inference (e.g., Cosmides & Tooby, 1996; Gigerenzer & Hoffrage, 1995; Meder & Nelson, 2012; Nelson, 2009). However, we added the complementary likelihoods to the standard probability format, and this addition might have enhanced participants' sensitivity to data informativeness (e.g., Rusconi & McKenzie, in press, Experiment 2). In any case, even in a judgment task with a suboptimal information format such as ours, the results of the analysis of significance suggested that participants were able to adhere to theoretically optimal models of the value of obtained answers. Indeed, the significant fixed effects indicated that the observed participants' responses were associated with the theoretical predictions of the models (with the exceptions of information gain in Experiments 1–2, Kullback-Leibler distance in Experiments 1–2 and the Additional Experiment, and impact in Experiment 3 and the Additional Experiment). Future studies should extend this investigation to tasks that use different information formats and, in particular, that focus on experience-based learning of environmental probabilities (e.g., Meder & Nelson, 2012; Nelson, 2009; Nelson et al., 2010).

The dichotomy between evidence evaluation and evidence use

The ability of participants to adhere to a normatively correct criterion, such as *Z*, provides evidence for optimal data evaluation. This finding is remarkable if we consider that previous studies using the differential answer diagnosticity task and requiring posterior probability estimates from participants found a relative insensitivity to Bayesian answer diagnosticity when unfamiliar scenarios, such as ours, were used (McKenzie, 2006; Rusconi & McKenzie, in press; Skov & Sherman, 1986, p. 118; Slowiaczek et al., 1992). This finding points to the difference between evaluating the impact of an obtained piece of evidence and using this evaluation to revise one's initial beliefs. Indeed, it has been shown that people's posterior probability estimates might be

prone to biases even though the evidence-evaluation process might be consistent with normatively appealing models (Tentori et al., 2007). This difference in performance when evaluating vs. using evidence might be explained in different ways. One reason might reside in the response mode. People might find it more difficult to express their likelihood or belief-revision estimates in terms of probabilities than to use an impact scale (e.g., Mastropasqua et al., 2010; Tentori et al., 2007) or a rating scale such as that used in our experiments. Another possible explanation rests on the repartition of the judgment process (e.g., Fischhoff & Beyth-Marom, 1983; Slovic & Lichtenstein, 1971). People might perceive the impact of obtained evidence in a Bayesian fashion (Rusconi et al. [2013] noted a case of *misperception*, but they used weight of evidence, that is, $\log_{10}$ diagnosticity, and not *Z* as the normative criterion). However, they might then fail in the combining process—that is, they might integrate their estimated impact of the obtained evidence with their initial beliefs (i.e., the priors) in a non-Bayesian fashion. One example of such *misaggregation* is the integration of priors and new evidence in an additive instead of a multiplicative, Bayesian way (e.g., Fischhoff & Beyth-Marom, 1983; Juslin, Nilsson, & Winman, 2009; Rusconi et al., 2013). If the latter were the case, it would be possible to devise algorithms that receive as inputs people's assessments of the components of Bayes' model. In this way, mechanical systems could help people avoid the computational errors that they might incur during the combining process (Fischhoff & Beyth-Marom, 1983).

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, AC-19*, 716–723. doi:10.1109/TAC.1974.1100705

Baayen, R. H. (2008). *Analyzing Linguistic Data. A Practical Introduction to Statistics Using R*. Cambridge: Cambridge University Press.

Baron, J. (1985). *Rationality and intelligence*. Cambridge: Cambridge University Press.

Baron, J., Beattie, J., & Hershey, J. C. (1988). Heuristics and biases in diagnostic reasoning: II. Congruence, information, and certainty. *Organizational Behavior and Human Decision Processes,42*, 88–110. doi:10.1016/0749-5978(88)90021-0

Bates, D., Maechler, M., & Bolker, B. (2012). *Linear mixed-effects models using S4 classes*. R package version 0.999999-0. Retrieved from http://cran.r-project.org/web/packages/lme4/index.html

Beach, L. R. (1968). Probability magnitudes and conservative revision of subjective probabilities. *Journal of Experimental Psychology, 77*, 57–63. doi:10.1037/h0025800

Belov, D. I., & Armstrong, R. D. (2011). Distributions of the Kullback-Leibler divergence with applications. *British Journal of Mathematical and Statistical Psychology, 64*, 291–309. doi: 10.1348/000711010X522227

Beyth-Marom, R. & Fischhoff, B. (1983). Diagnosticity and pseudodiagnosticity. *Journal of Personality and Social Psychology, 45*, 1185–1195. doi:10.1037//0022-3514.45.6.1185

Brambilla, M., Rusconi, P., Sacchi, S., & Cherubini, P. (2011). Looking for Honesty: The Primary Role of Morality (vs. Sociability and Competence) in Information Gathering. *European Journal of Social Psychology, 41*, 135-143. doi: 10.1002/ejsp.744.

Burnham, K. P., & Anderson, D. R. (2001). Kullback-Leibler information as a basis for strong inference in ecological studies. *Wildlife Research, 28*, 111–119.

Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods & Research, 33*,261–304. doi: 10.1177/0049124104268644

Cameron, J. A., & Trope, Y. (2004). Stereotype-biased search and processing of information about group members. *Social Cognition, 22*, 650-672. doi: 10.1521/soco.22.6.650.54818

Carnap, R. (1950). *Logical foundations of probability*. Chicago: The University of Chicago Press.

Cherubini, P., Rusconi, P., Russo, S., & Crippa, F. (2013). Missing the dog that failed to bark in the nighttime: on the overestimation of occurrences over non-occurrences in hypothesis testing. *Psychological Research, 77*, 348–370. doi: 10.1007/s00426-012-0430-3

Cherubini, P., Rusconi, P., Russo, S., Di Bari, S., & Sacchi, S. (2010). Preferences for different questions when testing hypotheses in an abstract task: Positivity does play a role, asymmetry does not. *Acta Psychologica*, *134,* 162-174. doi: 10.1016/j.actpsy.2010.01.007

Cherubini, P., Russo, S., Rusconi, P., D'Addario, M., & Boccuti, I. (2009). *Il ragionamento probabilistico nella diagnosi medica: sensibilità e insensibilità alle informazioni*. In P. Giaretta, A. Moretto, G. F. Gensini, & M. Trabucchi (Eds.), *Filosofia della medicina: Metodo, modelli, cura ed errori* (pp. 541-564). Bologna: Il Mulino.

Cosmides, L., & Tooby, J. (1996). Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgment under uncertainty. *Cognition, 58*, 1–73. doi:10.1016/0010-0277(95)00664-8

Cover, T. M., & Thomas, J. A. (1991). *Elements of information theory*. New York: John Wiley & Sons, Inc.

Crupi, V., & Tentori, K. (2010). Irrelevant conjunction: Statement and solution of a new paradox. *Philosophy of Science, 77*, 1–13. doi: 10.1086/650205

Crupi, V., Tentori, K., & Gonzalez, M. (2007). On Bayesian measures of evidential support: Theoretical and empirical issues. *Philosophy of Science, 74*, 229–252. doi:10.1086/520779

Edwards, W. (1965). Optimal strategies for seeking information: Models for statistics, choice reaction times, and human information processing. *Journal of Mathematical Psychology, 2*, 312–329. doi: 10.1016/0022-2496(65)90007-6

Eimas, P. D. (1970). Information processing in problem solving as a function of developmental level and stimulus saliency. *Developmental Psychology, 2*, 224-229. doi: 10.1037/h0028746

Evans, J. St. B. T., & Over, D. E. (1996). Rationality in the selection task: Epistemic utility versus uncertainty reduction. *Psychological Review, 103*, 356–363. http://dx.doi.org/10.1037//0033-295X.103.2.356

Evett, S. R., Devine, P. G., Hirt, E. R., & Price, J. (1994). The role of the hypothesis and the evidence in the trait hypothesis testing process. *Journal of Experimental Social Psychology, 30*, 456-481. doi: 10.1006/jesp.1994.1022

Fischhoff, B., & Beyth-Marom, R. (1983). Hypothesis evaluation from a Bayesian perspective. *Psychological Review, 90*, 239–260. doi:10.1037//0033-295X.90.3.239

Fitelson, B. (2001). A Bayesian account of independent evidence with applications. *Philosophy of Science, 68*, S123–S140. doi: 10.1086/392903

Fitelson, B. (2006). Logical foundations of evidential support. *Philosophy of Science, 73*, 500-512. doi: 10.1086/518320

Garcia-Marques, L., Sherman, S. J., & Palma-Oliveira, J. M. (2001). Hypothesis testing and the perception of diagnosticity. *Journal of Experimental Social Psychology, 37*, 183–200. doi:10.1006/jesp.2000.1441

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review, 102*, 684–704. doi:10.1037//0033-295X.102.4.684

Good, I. J. (1950). *Probability and the weighing of evidence*. London: Charles Griffin & Co. Ltd.

Good, I. J. (1979). Studies in the history of probability and statistics. XXXVII A. M. Turing's statistical work in World War II. *Biometrika, 66*, 393–396. doi: 10.1093/biomet/66.2.393

Good, I. J. (1983). *Good thinking: The foundations of probability and its applications*. Minneapolis, MN: University of Minnesota Press.

Good, I. J., & Card, W. I. (1971). The diagnostic process with special reference to errors. *Methods of Information in Medicine, 10*, 176–188.

Hammerton, M. (1973). A case of radical probability estimation. *Journal of Experimental Psychology, 101*, 252–254. doi:10.1037/h0035224

Juslin, P., Nilsson, H., & Winman, A. (2009). Probability theory, not the very guide of life. *Psychological Review, 116*, 856–874. doi:10.1037/a0016979

Kemeny, J. G. (1953). A logical measure function. *The Journal of Symbolic Logic, 18*, 289–308. doi:10.2307/2266553

Kemeny, J. G., & Oppenheim, P. (1952). Degree of factual support. *Philosophy of Science, 19*, 307–324. doi:10.1086/287214

Klayman, J., & Ha, Y. W. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review, 94*, 211-228. doi: 10.1037/0033-295X.94.2.211

Kruschke, J. K. (2001). Toward a unified model of attention in associative learning. *Journal of Mathematical Psychology, 45*, 812–863. http://dx.doi.org/10.1006%2Fjmps.2000.1354

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics, 22,* 79-86. doi: 10.1214/aoms/1177729694

Lindley, D. V. (1956). On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics, 27*, 986-1005. doi: 10.1214/aoms/1177728069

Mastropasqua, T., Crupi, V., & Tentori, K. (2010). Broadening the study of inductive reasoning: Confirmation judgments with uncertain evidence. *Memory & Cognition, 38*, 941-950. doi:10.3758/MC.38.7.941

McKenzie, C. R. M. (1994). The accuracy of intuitive judgment strategies: Covariation assessment and Bayesian inference. *Cognitive Psychology, 26*, 209-239. doi: 10.1006/cogp.1994.1007

McKenzie, C. R. M. (2004). Hypothesis testing and evaluation. In D. J. Koehler & N. Harvey (Eds.), *Blackwell handbook of judgment and decision making* (pp. 200-219). Oxford: Blackwell.

McKenzie, C. R. M. (2006). Increased sensitivity to differentially diagnostic answers using familiar materials: Implications for confirmation bias. *Memory & Cognition, 34*, 577–588. doi:10.3758/BF03193581

Meder, B., & Nelson, J. D. (2012). Information search with situation-specific reward functions. *Judgment and Decision Making, 7*, 119–148.

Nelson, J. D. (2005). Finding useful questions: On Bayesian diagnosticity, probability, impact, and information gain. *Psychological Review, 112*, 979–999. doi:10.1037/0033-295X.112.4.979

Nelson, J. D. (2008). Towards a rational theory of human information acquisition. In N. Chater & M. Oaksford (Eds.), *The probabilistic mind: Prospects for Bayesian cognitive science* (pp. 143-163). Oxford, UK: Oxford University Press.

Nelson, J. D. (2009). Naïve optimality: Subjects' heuristics can be better motivated than experimenters' optimal models. *Behavioral and Brain Sciences, 32*, 94–95. doi:10.1017/S0140525X09000405

Nelson, J. D., McKenzie, C. R. M., Cottrell, G. W., & Sejnowski, T. J. (2010). Experience matters: Information acquisition optimizes probability gain. *Psychological Science, 21*, 960–969. doi:10.1177/0956797610372637

Oaksford, M., & Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review, 101,* 608–631. doi:10.1037//0033-295X.101.4.608

Oaksford, M., & Chater, N. (2003). Optimal data selection: Revision, review, and reevaluation. *Psychonomic Bulletin & Review, 10*, 289–318. doi:10.3758/BF03196492

Peirce, C. S. (1878). The probability of induction. *Popular Science Monthly*, reprinted in *Philosophical writings of Peirce*, Justus Buchler, ed., New York: Dover Publications, Inc. (1955), pp. 174–189.

Popper, K. R. (1954). Degree of confirmation. *The British Journal for the Philosophy of Science, 5*, 143–149. doi: 10.1093/bjps/V.18.143

R Development Core Team (2012). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Rusconi, P., Crippa, F., Russo, S., & Cherubini, P. (2012). Moderators of the feature-positive effect in abstract hypothesis-evaluation tasks. *Canadian Journal of Experimental Psychology, 66*, 181–192. doi: 10.1037/a0028173

Rusconi, P., Marelli, M., Russo, S., D'Addario, M., & Cherubini, P. (2013). Integration of base rates and new information in an abstract hypothesis-testing task. *British Journal of Psychology, 104*, 193–211. doi: 10.1111/j.2044-8295.2012.02112.x

Rusconi, P., & McKenzie, C. R. M. (in press). Insensitivity and oversensitivity to answer diagnosticity in hypothesis testing. *The Quarterly Journal of Experimental Psychology*. doi: 10.1080/17470218.2013.793732

Rusconi, P., Sacchi, S., Toscano, A., & Cherubini, P. (2012). Confirming expectations in asymmetric and symmetric social hypothesis testing. *Experimental Psychology, 59*, 243–250. doi: 10.1027/1618-3169/a000149

Sacchi, S., Rusconi, P., Bonomi, M., & Cherubini, P. (in press). Effects of asymmetric questions on impression formation: A trade-off between evidence diagnosticity and frequency. *Social Psychology*. 10.1027/1864-9335/a000158

Sacchi, S., Rusconi, P., Russo, S., Bettiga, R., & Cherubini, P. (2012). New knowledge for old credences: Asymmetric information search about in-group and out-group members. *British Journal of Social Psychology, 51*, 606–625. doi: 10.1111/j.2044-8309.2011.02026.x

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*, 461–464. doi: 10.1214/aos/1176344136

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal, 27*, 379–423, 623–656.

Skov, R. B., & Sherman, S. J. (1986). Information-gathering processes: Diagnosticity, hypothesis-confirmatory strategies, and perceived hypothesis confirmation. *Journal of Experimental Social Psychology, 22*, 93–121. doi: 10.1016/0022-1031(86)90031-4

Slowiaczek, L. M., Klayman, J., Sherman, S. J., & Skov, R. B. (1992). Information selection and use in hypothesis testing: What is a good question, and what is a good answer? *Memory & Cognition, 20*, 392–405. doi: 10.3758/BF03210923

Tentori, K., Crupi, V., Bonini, N., & Osherson, D. (2007). Comparison of confirmation measures. *Cognition, 103*, 107-119. doi:10.1016/j.cognition.2005.09.006

Trope, Y., & Liberman, A. (1996). Social hypothesis-testing: Cognitive and motivational mechanisms. In E. T. Higgins & A. W. Kruglanski (Eds.), *Social psychology: Handbook of basic principles* (pp. 239-270). New York: Guilford Press.

Trope, Y., & Thompson, E. P. (1997). Looking for truth in all the wrong places? Asymmetric search of individuating information about stereotyped group members. *Journal of Personality and Social Psychology, 73*, 229-241. doi: 10.1037/0022-3514.73.2.229

Villejoubert, G., & Mandel, D. R. (2002). The inverse fallacy: An account of deviations from Bayes's theorem and the additivity principle. *Memory & Cognition, 30*, 171–178. doi: 10.3758/BF03195278

Wagenmakers, E.-J., & Farrell, S. (2004). AIC model selection using Akaike weights. *Psychonomic Bulletin & Review, 11*, 192–196. doi: 10.3758/BF03206482

Wason, P. C. (1966). Reasoning. In B. M. Foss (Ed.), *New horizons in psychology I*. Harmondsworth, UK: Penguin.

Wason, P.C. (1968). Reasoning about a rule. *Quarterly Journal of Experimental Psychology*, *20*, 273-281. http://dx.doi.org/10.1080/14640746808400161

Footnotes

[1] In all of the experiments in the present article, participants also estimated the probability (between 0% and 100%) that the encountered creature was a Glom [Fizo]. Figures illustrating means and standard errors of the mean are provided in Figure 1 of Supplementary Materials. Furthermore, at the end of the questionnaire, participants rated on a 7-point scale to what extent they considered the priors when (1) rating answer utility and (2) estimating the posterior probabilities ("1" = *little*, "7" = *a lot*). However, we do not discuss these results because they are out of the scope of the present contribution.

Appendix

All of the models described in this appendix are based on Bayes' rule, which can be expressed in terms of odds (e.g., Beyth-Marom & Fischhoff, 1983; Fischhoff & Beyth-Marom, 1983) as follows:

$$\frac{p(H \mid D)}{p(\neg H \mid D)} = \frac{p(H)}{p(\neg H)} \times \frac{p(D \mid H)}{p(D \mid \neg H)}, \tag{A1}$$

where *p*(denotes the "probability of," | should be read as "given that," *H* is the hypothesis being or to be tested (i.e., the focal hypothesis), ¬*H* stands for the alternate/s (¬ is the logical symbol for negation), and *D* represent the new evidence. From the left, there are three ratios in Equation (A1): The posterior odds, which defines the updated tester's beliefs after receiving the new evidence; the prior odds, which quantifies the tester's initial beliefs (prior to receiving the new evidence); and the likelihood ratio (LR), which expresses the probability of receiving the new evidence as a function of the truth or falsity of the focal hypothesis.

Bayesian diagnosticity and $\log_{10}$ diagnosticity

Bayesian diagnosticity and $\log_{10}$ diagnosticity are two OED models based solely on the LR expressed in Bayes' rule (also known as (Bayes) factor, e.g., Good, 1983, and equal to Jeffreys' *K*, e.g., Good, 1950; the third term from the left in Equation (A1)). Accordingly, both Bayesian diagnosticity and $\log_{10}$ diagnosticity provide a net value of the informativeness of a datum, and to compute these measures, it is not necessary to know either the prior or the posterior probabilities of the hypotheses. Both measures can be implemented so that they can define either the usefulness of a datum with respect to a single hypothesis or the usefulness of a datum with respect to the whole constellation of hypotheses.

Nelson (2005, 2008) used Bayesian diagnosticity and $\log_{10}$ diagnosticity without reference to a focal hypothesis by using the "maximum" formulation. Another example of implementation without reference to a focal hypothesis is the absolute log LR to which Klayman and Ha (1987) referred in the appendix of their article (see also Evans & Over, 1996, Equation (2)). In the present

article, we followed Nelson's implementation to define the utility values of Bayesian diagnosticity and $\log_{10}$ diagnosticity in the case of multiple-hypothesis testing. That is, the Bayesian diagnosticity of a "yes" answer is given by the following expression:

$$\max\left[p(D \mid H)/p(D \mid \neg H), p(D \mid \neg H)/p(D \mid H)\right]. \quad \text{(A2)}$$

The Bayesian diagnosticity of a "no" answer is expressed as follows:

$$\max\left[p(\neg D \mid H)/p(\neg D \mid \neg H), p(\neg D \mid \neg H)/p(\neg D \mid H)\right]. \quad \text{(A3)}$$

In a similar way, the $\log_{10}$ diagnosticity of a "yes" answer is given by the following expression (Nelson, 2005):

$$\log_{10} \max\left[p(D \mid H)/p(D \mid \neg H), p(D \mid \neg H)/p(D \mid H)\right]. \quad \text{(A4)}$$

The $\log_{10}$ diagnosticity of a "no" answer is expressed as follows:

$$\log_{10} \max\left[p(\neg D \mid H)/p(\neg D \mid \neg H), p(\neg D \mid \neg H)/p(\neg D \mid H)\right]. \quad \text{(A5)}$$

In contrast, Irving John Good (1916–2009), who was the statistical assistant of Alan Turing (1912–1954) during World War II, described how Turing used log diagnosticity with respect to a single hypothesis: "Turing introduced the expression '(Bayes) factor in favour of a hypothesis'" (Good, 1979, p. 393) and: "Dr. A. M. Turing suggested in a conversation in 1940 that the word "factor" should be regarded as (…) *the factor in favour of the hypothesis H in virtue of the result of the experiment*" (Good, 1950, p. 63). Taking these definitions literally, Bayesian diagnosticity reduces to the mere LR, whereas $\log_{10}$ diagnosticity can be expressed as follows for a "yes" answer:

$$\log_{10} p(D \mid H)/p(D \mid \neg H), \quad \text{(A6)}$$

and as follows for a "no" answer:

$$\log_{10} p(\neg D \mid H)/p(\neg D \mid \neg H), \quad \text{(A7)}$$

where *H* stands for the focal hypothesis. Turing called the logarithm in base ten of the LR "weight of evidence," an expression first used by Peirce (1878) with a similar meaning (e.g., Good, 1983). Turing defined the unit of measure of the weight of evidence as *ban* or, by analogy with the *decibel* scale in acoustics, *deciban*, one-tenth of a *ban*, when the base of the logarithm is 10 (e.g.,

Good, 1950, 1979, 1983; Rusconi et al., 2013). In the experiments described in the main text, we used *decibans*.

Evans and Over (1996) argued that the measures based on LR are to be favored as normative criteria for epistemic utility over other standards (such as information gain) because they provide positive values of utility whenever data are informative. However, Nelson (2005, 2008, 2009) described some flaws of these measures. For example, Bayesian diagnosticity and $\log_{10}$ diagnosticity cannot be applied straightforwardly when there are more than two hypotheses under examination. Furthermore, they sometimes provide infinite utility values when one hypothesis is confirmed or disconfirmed with certainty by the evidence or when feature probabilities are extreme (Nelson, 2005, 2008, 2009).

Information gain

Information gain is a measure of uncertainty reduction (e.g., Evans & Over, 1996; Nelson, 2005, 2008; Nelson et al., 2010). Specifically, it is a theory of the value of information that derives from Shannon's (1948) definition of entropy (or uncertainty), which is as follows:

$$E_n(X) = -\sum_{i=1}^{n} p(x_i)\log_2 p(x_i), \tag{A8}$$

where X is a discrete random variable and x_i represents the possible values with probability $p(x_i)$. In turn, this equation is related to the Second Law of Thermodynamics and to Boltzmann's definition that connects entropy and probability by a logarithmic relationship. Information gain defines uncertainty reduction by subtracting posterior entropy from prior entropy, both when a feature is present (e.g., a "yes" answer), as in the following expression:

$$\begin{aligned}&\{[p(H)\times -\log_2 p(H)]+[p(\neg H)\times -\log_2 p(\neg H)]\}\\&-\{[p(H\mid D)\times -\log_2 p(H\mid D)]+[p(\neg H\mid D)\times -\log_2 p(\neg H\mid D)]\}\end{aligned}, \tag{A9}$$

and when a feature is absent (e.g., a "no" answer), as expressed in the following equation:

$$\begin{aligned}&\{[p(H)\times -\log_2 p(H)]+[p(\neg H)\times -\log_2 p(\neg H)]\}\\&-\{[p(H\mid \neg D)\times -\log_2 p(H\mid \neg D)]+[p(\neg H\mid \neg D)\times -\log_2 p(\neg H\mid \neg D)]\}\end{aligned}, \tag{A10}$$

where $p(H)$ and $p(\neg H)$ are the prior probabilities of the focal hypothesis and the alternate, respectively. $p(H \mid D)$ and $p(\neg H \mid D)$ are the posterior probabilities of the occurrence of the focal hypothesis given the receipt of new evidence *D* and of the alternate given the receipt of the same evidence, respectively. Finally, $p(H \mid \neg D)$ and $p(\neg H \mid \neg D)$ are the posterior probabilities of the respective hypotheses when evidence *D* is absent. Being the logarithm in base 2, the unit of measure of information gain is the *bit*.

In the psychological literature, Oaksford and Chater (1994, 2003) proposed an "optimal data selection model" in which expected information gain was the normative standard that aimed to account for people's performance on Wason's (1966, 1968) selection task.

An important property that distinguishes information gain from the diagnosticity measures illustrated above is that information gain allows negative utility values. Negative information gain does not mean that new evidence disconfirms a *particular* hypothesis. On the contrary, it means greater uncertainty about the hypotheses after the receipt of new evidence than before (e.g., Rusconi et al., 2013). That is, negative information gain means information loss, at least whenever information is conceived under a strictly logical conception (e.g., Evans & Over, 1996). As Evans and Over (1996) noted, this definition of information gain leads to some anomalies in terms of the psychological plausibility of this metric. For example, in a case in which there are two hypotheses and a datum shifts the probability of the focal hypothesis from $p = .3$ to $p = .5$, there is a loss of information caused by the increased uncertainty. Indeed, the highest peak of uncertainty occurs when the hypotheses are equiprobable (i.e., when $p = .5$). However, a tester interested in the truth value of the focal hypothesis would benefit from the receipt of the new datum and would likely judge it a gain of information (Evans & Over, 1996).

Accordingly, information gain does not provide utility values of obtained evidence with respect to a single hypothesis but only with respect to two or more hypotheses. It is thus not

possible to implement information gain so that positive and negative utility values systematically indicate hypothesis confirmation and hypothesis falsification, respectively.

Kullback-Leibler distance

The Kullback-Leibler distance (also known under a variety of other labels that include Kullback-Leibler divergence, Kullback-Leibler discrepancy, Kullback-Leibler information, Kullback-Leibler loss, cross-entropy, relative entropy, information divergence, and information for discrimination) is a measure of the distance between two distributions (e.g., Belov & Armstrong, 2011; Cover & Thomas, 1991). Given two probability distributions, $f(x)$ and $g(x)$, Kullback-Leibler distance can be expressed by the following equation (e.g., Belov & Armstrong, 2011; Burnham & Anderson, 2001):

$$KL(f,g)=\int_{-\infty}^{+\infty} f(x)\ln\frac{f(x)}{g(x)}dx. \qquad \text{(A11)}$$

$KL(f,g)$ represents the loss of information that occurs when the distribution g is used to approximate the distribution f. Accordingly, Kullback-Leibler distance is an information-theoretic measure of information loss or inefficiency (e.g., Burnham & Anderson, 2001, 2004; Cover & Thomas, 1991). It is a non-negative metric that yields a value of zero if and only if the two distributions are identical (e.g., Belov & Armstrong, 2011; Cover & Thomas, 1991; Nelson, 2008). The Akaike information criterion (AIC) for statistical model selection (Akaike, 1974) is based on Kullback-Leibler distance (e.g., Belov & Armstrong, 2011; Burnham & Anderson, 2001).

When one receives a "yes" answer to a question, Kullback-Leibler distance can be computed as:

$$\left[p(H\mid D)\times\log_2 p(H\mid D)/p(H)\right]+\left[p(\neg H\mid D)\times\log_2 p(\neg H\mid D)/p(\neg H)\right]. \text{(A12)}$$

For a "no" answer, it can be expressed as:

$$\left[p(H\mid \neg D)\times\log_2 p(H\mid \neg D)/p(H)\right]+\left[p(\neg H\mid \neg D)\times\log_2 p(\neg H\mid \neg D)/p(\neg H)\right]. \text{(A13)}$$

As with information gain, Kullback-Leibler distance also cannot be implemented with respect to a single hypothesis exactly because it implies a comparison between two distributions.

Probability gain

Whereas information gain is a measure of uncertainty reduction, probability gain is a measure of error reduction, that is, it measures the extent to which an answer reduces the probability of preferring an erroneous hypothesis. In other words, probability gain quantifies an answer's usefulness in terms of how much it increases the probability of the correct hypothesis. Previous studies have found that error reduction guides human information gathering when participants learn environmental probabilities (Nelson et al., 2010) as well as attention learning (e.g., Kruschke, 2001). The probability gain of a "yes" answer (the presence of a feature) when the tester evaluates the usefulness of the obtained evidence with respect to the whole constellation of hypotheses can be expressed as follows:

$$\max\left[p(H \mid D), p(\neg H \mid D)\right] - \max\left[p(H), p(\neg H)\right]. \quad \text{(A14)}$$

In the case of a "no" answer (i.e., of the absence of a feature), the probability gain is given by the following expression:

$$\max\left[p(H \mid \neg D), p(\neg H \mid \neg D)\right] - \max\left[p(H), p(\neg H)\right]. \quad \text{(A15)}$$

When the tester evaluates the usefulness of the obtained evidence with respect to her/his beliefs about a focal hypothesis, the probability gain of a "yes" answer is expressed as follows:

$$p(H \mid D) - p(H), \quad \text{(A16)}$$

where H is the focal hypothesis and D is the obtained evidence. The probability gain for a "no" answer is expressed as follows:

$$p(H \mid \neg D) - p(H), \quad \text{(A17)}$$

where $\neg D$ stands for the "no" answer.

In other words, the probability gain reduces to an inductive confirmation measure whenever it is calculated with respect to a single hypothesis. Indeed, by adding the algebraic operator to its

formula, it is possible to quantify the decreases or increases in the degrees of belief in a focal hypothesis. In this form, probability gain is no longer a measure of error reduction.

Impact

This metric has been advanced independently in different studies. Wells and Lindsay (1980) called it "informativeness" (or "information gain") in their Bayesian analysis of eyewitness lineup identifications and nonidentifications. They defined it as the absolute value of the difference between the prior probability of the focal hypothesis (e.g., the suspect is the criminal) and the posterior probability of the same hypothesis (e.g., the suspect is the criminal given an identification/a nonidentification). Accordingly, it measures the degree of revision of one's beliefs required by new data. Klayman and Ha (1987) called it "impact" and "expected change in belief (EΔP)" (Klayman & Ha, 1987, p. 219) in their work on confirmation and disconfirmation in hypothesis testing. They used it as a measure for assessing the most informative test. Indeed, EΔP measures the absolute magnitude of changes in beliefs by taking into account both confirming and disconfirming evidence. Along the same lines, Nickerson (1996) used the term "impact" to indicate "the *absolute value* of an observation's effect" (Nickerson, 1996, p. 20), whereby the effect of an observation is the difference between the posterior and the prior probability of a hypothesis.

In the present contribution, we shall use the formula described in Nelson (2005). In particular, the impact of a "yes" answer can be expressed as follows:

$$1/2 \times \left\{ \left| p(H \mid D) - p(H) \right| + \left| p(\neg H \mid D) - p(\neg H) \right| \right\}, \tag{A18}$$

whereas the impact of a "no" answer can be computed as:

$$1/2 \times \left\{ \left| p(H \mid \neg D) - p(H) \right| + \left| p(\neg H \mid \neg D) - p(\neg H) \right| \right\}. \tag{A19}$$

In the case of mutually exclusive and exhaustive hypotheses the formula reduces to:

$$\left| p(H \mid D) - p(H) \right|, \tag{A20}$$

for a "yes" answer and to:

$$\left| p(H \mid \neg D) - p(H) \right|, \tag{A21}$$

in the case of a "no" answer. Absolute values are not used when a tester is inquiring about a focal hypothesis so that the obtained values inform the tester about the direction of the change engendered by the evidence (i.e., toward confirmation of the focal hypothesis or toward its disconfirmation).

Whenever the prior probabilities of the hypotheses under examination are equal the predictions of impact are identical to those of probability gain (e.g., Nelson, 2005, 2008; Rusconi & McKenzie, in press).

Measures *L* and *Z*

Parallel to the investigation of the psychological plausibility of OED models in information seeking (e.g., Edwards, 1965; Good & Card, 1971; Nelson, 2005, 2008; Nelson et al., 2010), philosophers of science and psychologists debated on the normative and descriptive adequacy of measures of confirmation or of evidential support (e.g., Carnap, 1950; Crupi et al., 2007; Fitelson, 2001, 2006; Kemeny, 1953; Kemeny & Oppenheim, 1952; Mastropasqua et al., 2010; Popper, 1954; Tentori et al., 2007). It should be noted that, within a probabilistic framework, measures of confirmation are distinct from posterior probabilities because they capture the degrees of change (i.e., increases or decreases) that occur in one's initial beliefs and leads to the final confidence, that is, to posterior probabilities (e.g., Mastropasqua et al., 2010).

Recent studies have shown that two metrics, namely, measures *L* and *Z* adequately describe human confirmation judgments with either ascertained or uncertain evidence (e.g., Crupi et al., 2007; Mastropasqua et al., 2010; Tentori et al., 2007). Mathematically, measure *L* is strictly related to the log LR and best approximates intuitive judgments of confirmation with uncertain evidence (Mastropasqua et al., 2010).

When the tester evaluates the absolute change of beliefs engendered by the obtained evidence with respect to the whole constellation of hypotheses, the utility of a "yes" answer according to measure *L* can be expressed as:

$$\max\begin{Bmatrix}\left[p(H \mid D)/\, p(\neg H \mid D) - p(H)/\, p(\neg H)\right]/\left[p(H \mid D)/\, p(\neg H \mid D) + p(H)/\, p(\neg H)\right], \\ \left[p(\neg H \mid D)/\, p(H \mid D) - p(\neg H)/\, p(H)\right]/\left[p(\neg H \mid D)/\, p(H \mid D) + p(\neg H)/\, p(H)\right]\end{Bmatrix}. \quad \text{(A22)}$$

In a similar way, the utility of a "no" answer can be computed as:

$$\max\begin{Bmatrix}\left[p(H \mid \neg D)/\, p(\neg H \mid \neg D) - p(H)/\, p(\neg H)\right]/\left[p(H \mid \neg D)/\, p(\neg H \mid \neg D) + p(H)/\, p(\neg H)\right], \\ \left[p(\neg H \mid \neg D)/\, p(H \mid \neg D) - p(\neg H)/\, p(H)\right]/\left[p(\neg H \mid \neg D)/\, p(H \mid \neg D) + p(\neg H)/\, p(H)\right]\end{Bmatrix}. \quad \text{(A23)}$$

Whenever the tester considers the plausibility of a focal hypothesis, measure *L* can be implemented with the algebraic operator so that the value indicates the direction of belief change (i.e., toward hypothesis confirmation vs. hypothesis disconfirmation). In particular, the value of a "yes" answer can be expressed as follows:

$$\left[p(H \mid D)/\, p(\neg H \mid D) - p(H)/\, p(\neg H)\right]/\left[p(H \mid D)/\, p(\neg H \mid D) + p(H)/\, p(\neg H)\right]. \quad \text{(A24)}$$

The value of a "no" answer can be computed as:

$$\left[p(H \mid \neg D)/\, p(\neg H \mid \neg D) - p(H)/\, p(\neg H)\right]/\left[p(H \mid \neg D)/\, p(\neg H \mid \neg D) + p(H)/\, p(\neg H)\right]. \quad \text{(A25)}$$

Measure *Z* has been advocated as theoretically appealing by Crupi et al. (2007) who found that it performed slightly but significantly better than *L* in describing human intuitions about confirmation with ascertained evidence. Measure *Z* defines the value of a "yes" answer with respect to the whole constellation of hypotheses as follows:

If either $p(H \mid D) \geq p(H)$ or $p(\neg H \mid D) \geq p(\neg H)$:

$$\max\left\{\left[p(H \mid D) - p(H)\right]/\, p(\neg H), \left[p(\neg H \mid D) - p(\neg H)\right]/\, p(H)\right\}$$

otherwise: (A26)

$$\max\left\{\left[p(H \mid D) - p(H)\right]/\, p(H), \left[p(\neg H \mid D) - p(\neg H)\right]/\, p(\neg H)\right\}.$$

The utility value of a "no" answer can be expressed as:

If either $p(H \mid \neg D) \geq p(H)$ or $p(\neg H \mid \neg D) \geq p(\neg H)$:

$$\max\{[p(H \mid \neg D) - p(H)]/\, p(\neg H), [p(\neg H \mid \neg D) - p(\neg H)]/\, p(H)\}$$

otherwise: (A27)

$$\max\{[p(H \mid \neg D) - p(H)]/\, p(H), [p(\neg H \mid \neg D) - p(\neg H)]/\, p(\neg H)\}.$$

Whenever the tester evaluates a focal hypothesis, measure *Z* for a "yes" answer can be computed as:

$$\textit{If } p(H \mid D) \geq p(H): [p(H \mid D) - p(H)]/\, p(\neg H) \quad \text{(A28)}$$

$$\textit{If } p(H \mid D) < p(H): [p(H \mid D) - p(H)]/\, p(H),$$

while the value of a "no" answer can be expressed as follows:

$$\textit{If } p(H \mid \neg D) \geq p(H): [p(H \mid \neg D) - p(H)]/\, p(\neg H) \quad \text{(A29)}$$

$$\textit{If } p(H \mid \neg D) < p(H): [p(H \mid \neg D) - p(H)]/\, p(H).$$

Table 1

Main Properties of the Experiments

Experiment	Type of testing	Type of evaluation	Phrasing of the task requirement
Experiment 1	single hypothesis	propensity of evidence to decrease/increase beliefs	How do you deem that the received answer ("YES") [("NO")] decreases/increases the plausibility of the hypothesis that the encountered creature is a **Glom** [**Fizo**]? (mark a number from -3 to 3)
Experiment 2	single hypothesis	helpfulness of evidence	How do you deem that the received answer ("YES") [("NO")] is helpful for ascertaining the possibility that the encountered creature is a **Glom** [**Fizo**]? (mark a number from -3 to 3)
Experiment 3	multiple hypotheses	helpfulness of evidence	How do you deem that the received answer ("YES") [("NO")] is helpful in distinguishing between the possibility that the encountered creature is a Glom and the possibility that it is a Fizo? (mark a number from -3 to 3)
Additional Experiment	multiple hypotheses	propensity of evidence to decrease/increase beliefs	How do you deem that the received answer ("YES") [("NO")] decreases/increases the plausibility of the hypothesis that the encountered creature is a Glom or that the encountered creature is a Fizo? (mark a number from -3 to 3)

Table 2

Formal Properties of Experiments 1–2

Answer	Hypothesis	Priors (%)	Likelihoods (%)	Bayesian Diagnosticity	Log_{10} Diagnosticity (*decibans*)	Information Gain	Kullback-Leibler distance	Probability Gain	Impact	*L*	*Z*
yes	Glom	25	90	1.8	2.55	-.14	.07	.13	.13	.29	.17
	Fizo	75	50	.56	-2.55	-.14	.07	-.12	-.12	-.29	-.16
no	Glom	25	10	.2	-6.99	.48	.18	-.19	-.19	-.68	-.76
	Fizo	75	50	5	6.99	.48	.18	.19	.19	.68	.76
yes	Glom	25	75	5	6.99	-.14	.47	.38	.38	.67	.51
	Fizo	75	15	.2	-6.99	-.14	.47	-.37	-.37	-.67	-.49
no	Glom	25	25	.29	-5.31	.37	.12	-.16	-.16	-.54	-.64
	Fizo	75	85	3.4	5.31	.37	.12	.16	.16	.54	.64
yes	Glom	25	45	.53	-2.76	.2	.04	-.1	-.1	-.31	-.4
	Fizo	75	85	1.89	2.76	.2	.04	.1	.1	.31	.4
no	Glom	25	55	3.67	5.64	-.18	.29	.3	.3	.57	.4
	Fizo	75	15	.27	-5.64	-.18	.29	-.3	-.3	-.57	-.4
yes	Glom	25	22	.22	-6.49	.45	.16	-.18	-.18	-.63	-.72
	Fizo	75	98	4.45	6.49	.45	.16	.18	.18	.63	.72
no	Glom	25	78	39	15.91	.45	1.52	.68	.68	.95	.91
	Fizo	75	2	.03	-15.91	.45	1.52	-.68	-.68	-.95	-.91
yes	Glom	75	90	1.8	2.55	.18	.03	.09	.09	.27	.36
	Fizo	25	50	.56	-2.55	.18	.03	-.09	-.09	-.27	-.36
no	Glom	75	10	.2	-6.99	-.14	.47	-.37	-.37	-.67	-.49
	Fizo	25	50	5	6.99	-.14	.47	.38	.38	.67	.51
yes	Glom	75	75	5	6.99	.48	.18	.19	.19	.68	.76
	Fizo	25	15	.2	-6.99	.48	.18	-.19	-.19	-.68	-.76
no	Glom	75	25	.29	-5.31	-.19	.26	-.28	-.28	-.54	-.37
	Fizo	25	85	3.4	5.31	-.19	.26	.28	.28	.54	.37
yes	Glom	75	45	.53	-2.76	-.15	.07	-.14	-.14	-.31	-.19
	Fizo	25	85	1.89	2.76	-.15	.07	.14	.14	.31	.19
no	Glom	75	55	3.67	5.64	.41	.14	.17	.17	.59	.68
	Fizo	25	15	.27	-5.64	.41	.14	-.17	-.17	-.59	-.68
yes	Glom	75	22	.22	-6.49	-.16	.4	-.35	-.35	-.64	-.47
	Fizo	25	98	4.45	6.49	-.16	.4	.35	.35	.64	.47
no	Glom	75	78	39	15.91	.73	.35	.24	.24	.94	.96
	Fizo	25	2	.03	-15.91	.73	.35	-.24	-.24	-.94	-.96

Table 3

Experiment 1: Parameters of the Fixed Effects Estimated by Means of Linear Mixed-Effects Models, their Statistical Significance, and the AIC and BIC Values of the Eight Competing Models

Fixed Effects	Estimate	Std. Error	*t* value	MCMCmean	95% Highest Posterior Density Interval	pMCMC	AIC_i	Δ_i (AIC)	Akaike weights	BIC_i	Δ_i (BIC)	Schwarz weights
Bayesian Diagnosticity	.08	.03	2.99	.08	[.04, .12]	.0002	2689	76	.0000	2716	76	.0000
Log_{10} Diagnosticity	.18	.02	11.35	.18	[.15, .21]	.0001	2650	37	.0000	2677	37	.0000
Information Gain	-.27	.90	-.30	-.28	[-1.50, 1.00]	.6574	2687	74	.0000	2714	74	.0000
Kullback-Leibler distance	-.40	.79	-.50	-.40	[-1.52, .70]	.4750	2690	77	.0000	2713	73	.0000
Probability Gain	4.55	.49	9.32	4.54	[3.68, 5.42]	.0001	2656	43	.0000	2684	44	.0000
Impact	4.55	.49	9.32	4.53	[3.65, 5.43]	.0001	2656	43	.0000	2684	44	.0000
Measure *L*	2.30	.16	14.64	2.30	[2.00, 2.61]	.0001	2641	28	.0000	2668	28	.0000
Measure *Z*	2.39	.16	14.74	2.39	[2.08, 2.71]	.0001	2613	0	1.0000	2640	0	1.0000

Note. $\Delta_i(AIC) = AIC_i - \min AIC$; $\Delta_i(BIC) = BIC_i - \min BIC$.

Table 4

Experiment 1: Normalized Probabilities and Evidence Ratios of Akaike and Schwarz Weights for the Competing Models

Contrasts	Evidence ratio of Akaike weights	Normalized probability of Akaike weights	Evidence ratio of Schwarz weights	Normalized probability of Schwarz weights
Measure *Z* vs. Measure *L*	1202604.28	1.00	1202604.28	1.00
Measure *Z* vs. Bayesian Diagnosticity	31855931757113800.00	1.00	31855931757113800.00	1.00
Measure *Z* vs. Log_{10} Diagnosticity	108254987.75	1.00	108254987.75	1.00
Measure *Z* vs. Information Gain	11719142372802600.00	1.00	11719142372802600.00	1.00
Measure *Z* vs. Kullback-Leibler distance	52521552285925200.00	1.00	7108019154642240.00	1.00
Measure *Z* vs. Probability Gain	2174359553.58	1.00	3584912846.13	1.00
Measure *Z* vs. Impact	2174359553.58	1.00	3584912846.13	1.00
Measures *L* and *Z* vs. OED models	98451821.21	1.00	102089408.88	1.00

Table 5

Experiment 2: Parameters of the Fixed Effects Estimated by Means of Linear Mixed-Effects Models, their Statistical Significance, and the AIC and BIC Values of the Eight Competing Models

Fixed Effects	Estimate	Std. Error	*t* value	MCMCmean	95% Highest Posterior Density Interval	pMCMC	AIC_i	Δ_i (AIC)	Akaike weights	BIC_i	Δ_i (BIC)	Schwarz weights
Bayesian Diagnosticity	.04	.01	3.01	.04	[.01, .06]	.0022	2732	22	.0000	2754	21	.0000
Log_{10} Diagnosticity	.06	.01	4.83	.06	[.04, .09]	.0001	2729	19	.0001	2752	19	.0001
Information Gain	.21	.43	.50	.20	[-.56, 1.01]	.6042	2730	20	.0000	2753	20	.0000
Kullback-Leibler distance	.38	.38	1.01	.38	[-.32, 1.07]	.2742	2729	19	.0001	2752	19	.0001
Probability Gain	1.67	.35	4.70	1.64	[.97, 2.31]	.0001	2723	13	.0015	2746	13	.0015
Impact	1.67	.35	4.70	1.64	[.99, 2.35]	.0001	2723	13	.0015	2746	13	.0015
Measure *L*	.87	.15	5.94	.87	[.60, 1.15]	.0001	2718	8	.0179	2741	8	.0179
Measure *Z*	.91	.15	5.88	.90	[.62, 1.22]	.0001	2710	0	.9789	2733	0	.9789

Note. $\Delta_i(AIC) = AIC_i - \min AIC$; $\Delta_i(BIC) = BIC_i - \min BIC$.

Table 6

Experiment 2: Normalized Probabilities and Evidence Ratios of Akaike and Schwarz Weights for the Competing Models

Contrasts	Evidence ratio of Akaike weights	Normalized probability of Akaike weights	Evidence ratio of Schwarz weights	Normalized probability of Schwarz weights
Measure *Z* vs. Measure *L*	54.60	.98	54.60	.98
Measure *Z* vs. Bayesian Diagnosticity	59874.14	1.00	36315.50	1.00
Measure *Z* vs. Log_{10} Diagnosticity	13359.73	1.00	13359.73	1.00
Measure *Z* vs. Information Gain	22026.47	1.00	22026.47	1.00
Measure *Z* vs. Kullback-Leibler distance	13359.73	1.00	13359.73	1.00
Measure *Z* vs. Probability Gain	665.14	1.00	665.14	1.00
Measure *Z* vs. Impact	665.14	1.00	665.14	1.00
Measures *Z* and *L* vs. OED models	316.38	1.00	315.32	1.00

Table 7

Formal Properties of Experiment 3 and of the Additional Experiment

Answer	Hypothesis	Priors (%)	Likelihoods (%)	Bayesian Diagnosticity	Log_{10} Diagnosticity (*decibans*)	Information Gain	Kullback-Leibler distance	Probability Gain	Impact	*L*	*Z*
yes	Glom	25	90	1.8	2.55	-.14	.07	-.12	.13	.29	.17
	Fizo	75	50								
no	Glom	25	10	5	6.99	.48	.18	.19	.19	.68	.76
	Fizo	75	50								
yes	Glom	25	75	5	6.99	-.14	.47	-.12	.38	.67	.51
	Fizo	75	15								
no	Glom	25	25	3.4	5.31	.37	.12	.16	.16	.54	.64
	Fizo	75	85								
yes	Glom	25	45	1.89	2.76	.2	.04	.1	.1	.31	.4
	Fizo	75	85								
no	Glom	25	55	3.67	5.64	-.18	.29	-.2	.3	.57	.4
	Fizo	75	15								
yes	Glom	25	22	4.45	6.49	.45	.16	.18	.18	.63	.72
	Fizo	75	98								
no	Glom	25	78	39	15.91	.45	1.52	.18	.68	.95	.91
	Fizo	75	2								
yes	Glom	75	90	1.8	2.55	.18	.03	.09	.09	.27	.36
	Fizo	25	50								
no	Glom	75	10	5	6.99	-.14	.47	-.12	.38	.67	.51
	Fizo	25	50								
yes	Glom	75	75	5	6.99	.48	.18	.19	.19	.68	.76
	Fizo	25	15								
no	Glom	75	25	3.4	5.31	-.19	.26	-.22	.28	.54	.37
	Fizo	25	85								
yes	Glom	75	45	1.89	2.76	-.15	.07	-.14	.14	.31	.19
	Fizo	25	85								
no	Glom	75	55	3.67	5.64	.41	.14	.17	.17	.59	.68
	Fizo	25	15								
yes	Glom	75	22	4.45	6.49	-.16	.40	-.15	.35	.64	.47
	Fizo	25	98								
no	Glom	75	78	39	15.91	.73	.35	.24	.24	.94	.96
	Fizo	25	2								

Table 8

Experiment 3: Parameters of the Fixed Effects Estimated by Means of Linear Mixed-Effects Models, their Statistical Significance, and the AIC and BIC Values of the Eight Competing Models

Fixed Effects	Estimate	Std. Error	*t* value	MCMCmean	95% Highest Posterior Density Interval	pMCMC	AIC_i	Δ_i (AIC)	Akaike weights	BIC_i	Δ_i (BIC)	Schwarz weights
Bayesian Diagnosticity	.03	.01	4.19	.03	[.01, .04]	.0001	2725	10	.0051	2753	11	.0032
Log_{10} Diagnosticity	.08	.02	4.39	.08	[.04, .11]	.0001	2733	18	.0001	2761	19	.0001
Information Gain	.85	.24	3.49	.85	[.39, 1.33]	.0008	2721	6	.0380	2749	7	.0237
Kullback-Leibler distance	.61	.23	2.63	.61	[.19, 1.09]	.0088	2730	15	.0004	2758	16	.0003
Probability Gain	1.45	.48	3.05	1.46	[.50, 2.32]	.0018	2722	7	.0230	2750	8	.0143
Impact	1.09	.58	1.87	1.09	[.00, 2.24]	.0522	2729	14	.0007	2757	15	.0004
Measure *L*	1.31	.37	3.52	1.32	[.56, 2.03]	.0008	2718	3	.1701	2745	3	.1748
Measure *Z*	1.29	.31	4.17	1.28	[.70, 1.89]	.0001	2715	0	.7625	2742	0	.7833

Note. $\Delta_i(AIC) = AIC_i - \min AIC$; $\Delta_i(BIC) = BIC_i - \min BIC$.

Table 9

Experiment 3: Normalized Probabilities and Evidence Ratios of Akaike and Schwarz Weights for the Eight Competing Models

Contrasts	Evidence ratio of Akaike weights	Normalized probability of Akaike weights	Evidence ratio of Schwarz weights	Normalized probability of Schwarz weights
Measure *Z* vs. Measure *L*	4.48	.82	4.48	.82
Measure *Z* vs. Bayesian Diagnosticity	148.41	.99	244.69	1.00
Measure *Z* vs. Log_{10} Diagnosticity	8103.08	1.00	13359.73	1.00
Measure *Z* vs. Information Gain	20.09	.95	33.12	.97
Measure *Z* vs. Kullback-Leibler distance	1808.04	1.00	2980.96	1.00
Measure *Z* vs. Probability Gain	33.12	.97	54.60	.98
Measure *Z* vs. Impact	1096.63	1.00	1808.04	1.00
Measures *L* and *Z* vs. OED models	13.85	.93	22.84	.96

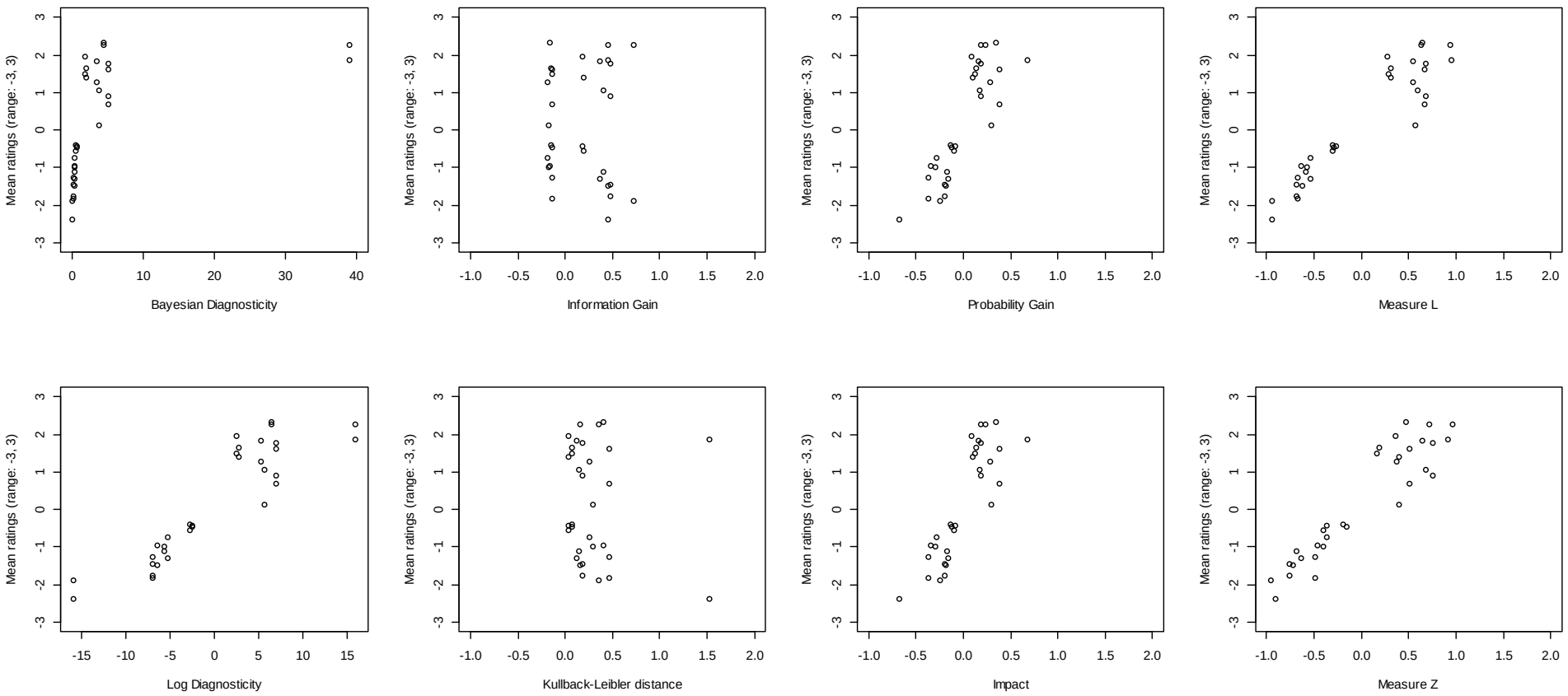


Figure 1. Experiment 1: Scatterplots illustrating the relationships between participants' mean ratings (averaged across the 32 cells of the experimental design) and the values predicted by the competing models (i.e., the six OED models and *L* and *Z*). The mean ratings are plotted along the Y-axis of each scatter plot, whereas the values predicted by the models are plotted along the X-axis. Each point in the scatterplots represents one of the 32 cells of the experimental design.

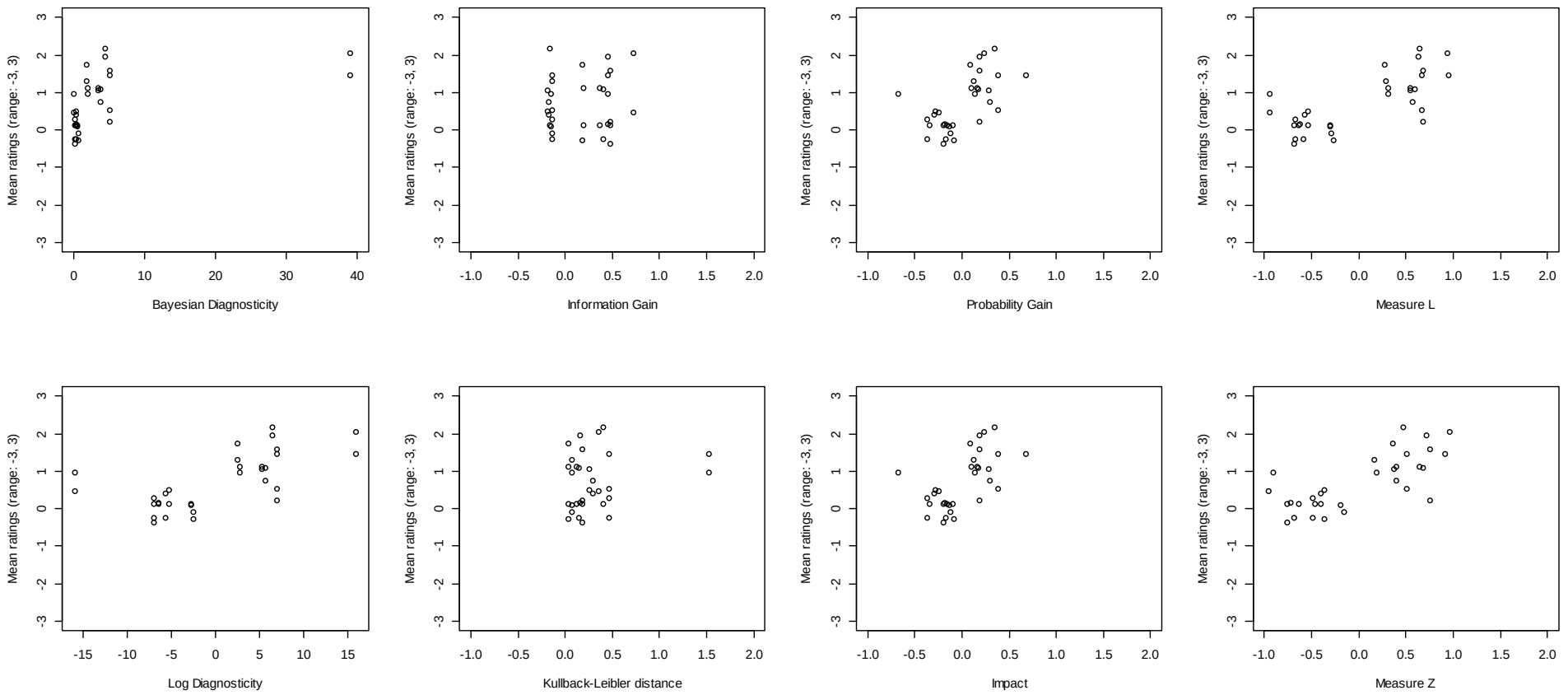


Figure 2. Experiment 2: Scatterplots illustrating the relationships between participants' mean ratings (averaged across the 32 cells of the experimental design) and the values predicted by the competing models (i.e., the six OED models and *L* and *Z*). The mean ratings are plotted along the Y-axis of each scatter plot, whereas the values predicted by the models are plotted along the X-axis. Each point in the scatterplots represents one of the 32 cells of the experimental design.

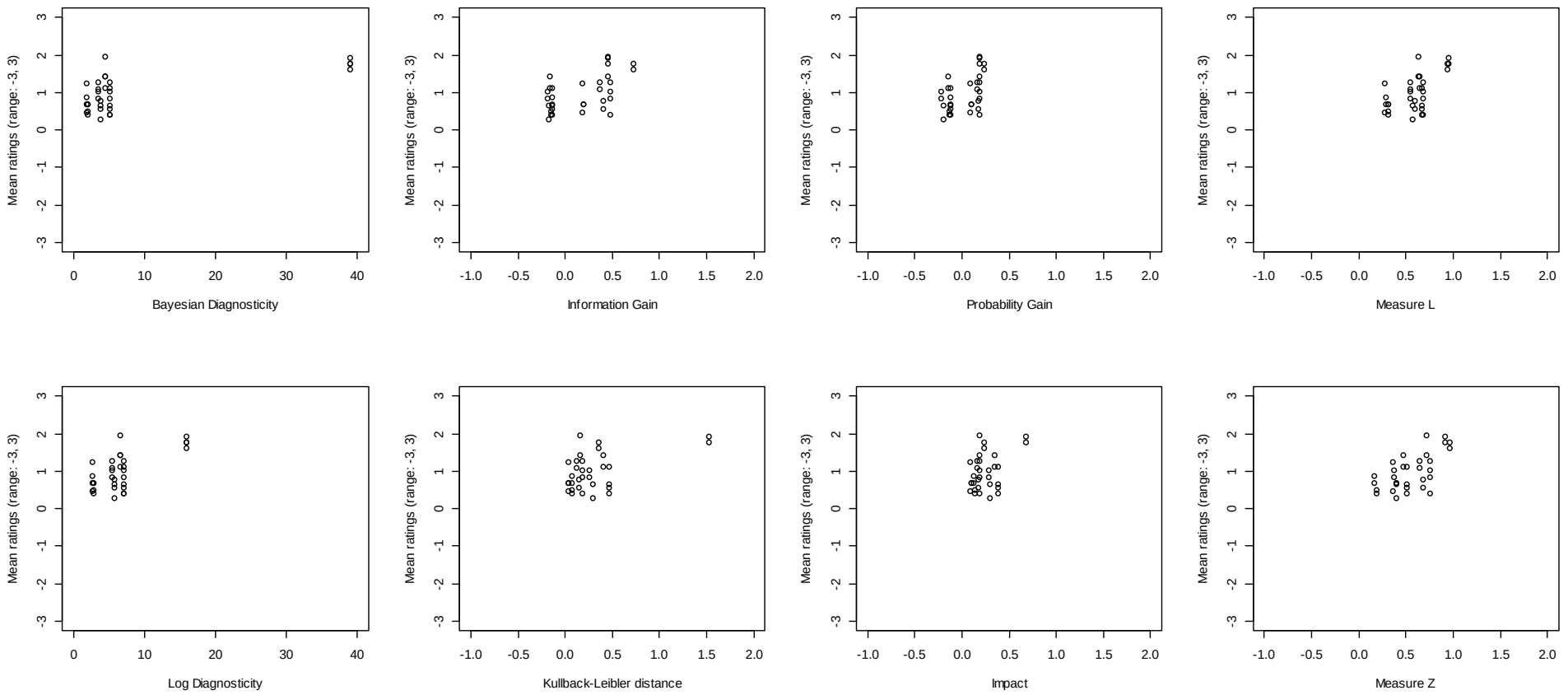


Figure 3. Experiment 3: Scatterplots illustrating the relationships between participants' mean ratings (averaged across the 32 cells of the experimental design) and the values predicted by the competing models (i.e., the six OED models and *L* and *Z*). The mean ratings are plotted along the Y-axis of each scatter plot, whereas the values predicted by the models are plotted along the X-axis. Each point in the scatterplots represents one of the 32 cells of the experimental design.

Supplementary materials

Table 1

The Sample Characteristics

	Experiment 1	Experiment 2	Experiment 3	Additional Experiment
N	95	92	102	92
Females	48	72	51	68
Males	47	20	50	23
NA			1	1
Mean age (*SD*, range)	22.33 (2.14, 19-29)	20.6 (2.02, 18-33)	21.63 (1.99, 18-29)	20.93 (1.95, 18-28)
Nationality: Italian	89	85	96	89
Italian and French				1
Albanian	1		1	
American	1			
Bulgarian		2		
Burmese			1	
Egyptian			1	
German		1	1	
Japanese	1			
Peruvian	1		1	
Polish		1		
Romanian		2		
Spanish		1		
NA	2		1	2
Course of study: Biology	5		1	1
Biostatistics	2			
Biotechnology		2	2	3
Chemistry		1	3	
Communication and society	1			
Communication and psychology	1	24	6	26
Computer science	7	3	7	1
Economics	31	3	26	1
Engineering	2			
Environmental sciences	1	3	3	
Geology			2	2
Goldsmith sciences	1			
Human sciences and education	2	13	10	15
Intercultural communication		2	5	3

Law	2	2	4	1
Materials engineering	1			
Materials science	4	1	1	1
Mathematics		2	4	
Medicine				1
Motor sciences			1	
Mathematics, physics and natural sciences	2		1	
Nursing	1			
Optics and optometry	2	3		
Psychology	12	25	17	25
Physics	5	4	6	2
Physiotherapy				1
Science of tourism		1		
Sociology	4	2		3
Social services	2	1		1
Statistical sciences	2			2
Theory and technology of communication				1
Tourism, territory and local development	3			
NA			3	2

Table 2

Correlations among the Theoretical Utility Values across the 32 Conditions of Experiments 1–2

	1	2	3	4	5	6	7
1. Bayesian Diagnosticity							
2. Log_{10}Diagnosticity	.68**						
3. Information Gain	.34	.00					
4. Kullback-Leibler distance	.46**	.00	.12				
5. Probability Gain	.56**	.92**	-.00	.00			
6. Impact	.56**	.92**	-.00	.00	1**		
7. Measure *L*	.56**	.97**	.00	.00	.93**	.93**	
8. Measure *Z*	.56**	.96**	-.00	.00	.87**	.87**	.98**

Note. ** the correlation is significant at the level .01 (two-tailed).

Table 3

The Mean Ratings (the Standard Errors of the Mean, SEM, in Parentheses) Provided by Participants in the Experiments

Answer	Hypothesis	Priors (%)	Likelihoods (%)	Mean ratings (SEM) – Experiment 1	Mean ratings (SEM) – Experiment 2	Mean ratings (SEM) – Experiment 3	Mean ratings (SEM) – Additional Experiment
yes	Glom	25	90	1.48 (.31)	1.29 (.32)	.88 (.36)	.84 (.24)
	Fizo	75	50	-.46 (.32)	-.08 (.32)	.69 (.29)	.61 (.30)
no	Glom	25	10	-1.43 (.41)	.13 (.46)	.42 (.4)	.45 (.35)
	Fizo	75	50	.92 (.31)	.22 (.29)	.84 (.24)	.57 (.39)
yes	Glom	25	75	1.61 (.19)	1.46 (.28)	1.12 (.25)	.59 (.28)
	Fizo	75	15	-1.83 (.17)	.29 (.33)	.65 (.35)	.61 (.31)
no	Glom	25	25	-1.30 (.30)	.13 (.41)	1.08 (.35)	.68 (.35)
	Fizo	75	85	1.83 (.26)	1.13 (.28)	1.28 (.33)	1.26 (.32)
yes	Glom	25	45	-.57 (.34)	.13 (.36)	.68 (.29)	.77 (.25)
	Fizo	75	85	1.41 (.27)	1.13 (.29)	.69 (.28)	1.13 (.28)
no	Glom	25	55	.13 (.33)	.75 (.32)	.28 (.37)	.36 (.27)
	Fizo	75	15	-1.00 (.34)	.42 (.35)	.64 (.35)	.57 (.29)
yes	Glom	25	22	-1.48 (.44)	.17 (.45)	1.44 (.35)	.86 (.43)
	Fizo	75	98	2.26 (.33)	1.96 (.29)	1.96 (.22)	2.00 (.33)
no	Glom	25	78	1.87 (.32)	1.46 (.34)	1.76 (.27)	1.23 (.29)
	Fizo	75	2	-2.38 (.24)	.96 (.48)	1.92 (.26)	1.64 (.35)
yes	Glom	75	90	1.96 (.24)	1.73 (.21)	1.24 (.33)	1.38 (.29)
	Fizo	25	50	-.44 (.27)	-.29 (.33)	.48 (.33)	1.39 (.19)
no	Glom	75	10	-1.26 (.32)	-.23 (.37)	.4 (.33)	.79 (.28)
	Fizo	25	50	.68 (.31)	.55 (.32)	.58 (.27)	.87 (.30)
yes	Glom	75	75	1.78 (.21)	1.59 (.24)	1.28 (.25)	1.58 (.22)
	Fizo	25	15	-1.76 (.28)	-.38 (.38)	1.04 (.24)	1.30 (.32)
no	Glom	75	25	-.74 (.35)	.50 (.33)	.84 (.26)	.83 (.32)
	Fizo	25	85	1.28 (.37)	1.05 (.35)	1.04 (.26)	1.57 (.23)
yes	Glom	75	45	-.39 (.26)	.10 (.28)	.42 (.26)	.38 (.30)
	Fizo	25	85	1.64 (.22)	.95 (.33)	.5 (.26)	1.00 (.23)
no	Glom	75	55	1.04 (.26)	1.10 (.26)	.56 (.31)	1.29 (.27)
	Fizo	25	15	-1.12 (.39)	-.23 (.39)	.77 (.26)	.43 (.31)
yes	Glom	75	22	-.96 (.37)	.14 (.42)	1.12 (.36)	1.00 (.38)
	Fizo	25	98	2.32 (.16)	2.18 (.26)	1.42 (.3)	1.57 (.18)
no	Glom	75	78	2.26 (.27)	2.05 (.29)	1.76 (.31)	1.71 (.39)
	Fizo	25	2	-1.88 (.38)	.45 (.51)	1.62 (.29)	1.70 (.36)

Table 4

Correlations among the Theoretical Utility Values across the 32 Conditions of Experiment 3 and the Additional Experiment

	1	2	3	4	5	6	7
1. Bayesian Diagnosticity							
2. Log_{10}Diagnosticity	.95**						
3. Information Gain	.54*	.55*					
4. Kullback-Leibler distance	.72**	.76**	.12				
5. Probability Gain	.43	.43	.98**	.08			
6. Impact	.60*	.69**	-.08	.96**	-.13		
7. Measure *L*	.76**	.93**	.49	.69**	.37	.71**	
8. Measure *Z*	.69**	.82**	.85**	.47	.78**	.38	.87**

Note. ** the correlation is significant at the level .01 (two-tailed) * the correlation is significant at the level .05 (two-tailed).

Table 5

Additional Experiment: Parameters of the Fixed Effects Estimated by Means of Linear Mixed-Effects Models, their Statistical Significance, and the AIC and BIC Values of the Eight Competing Models

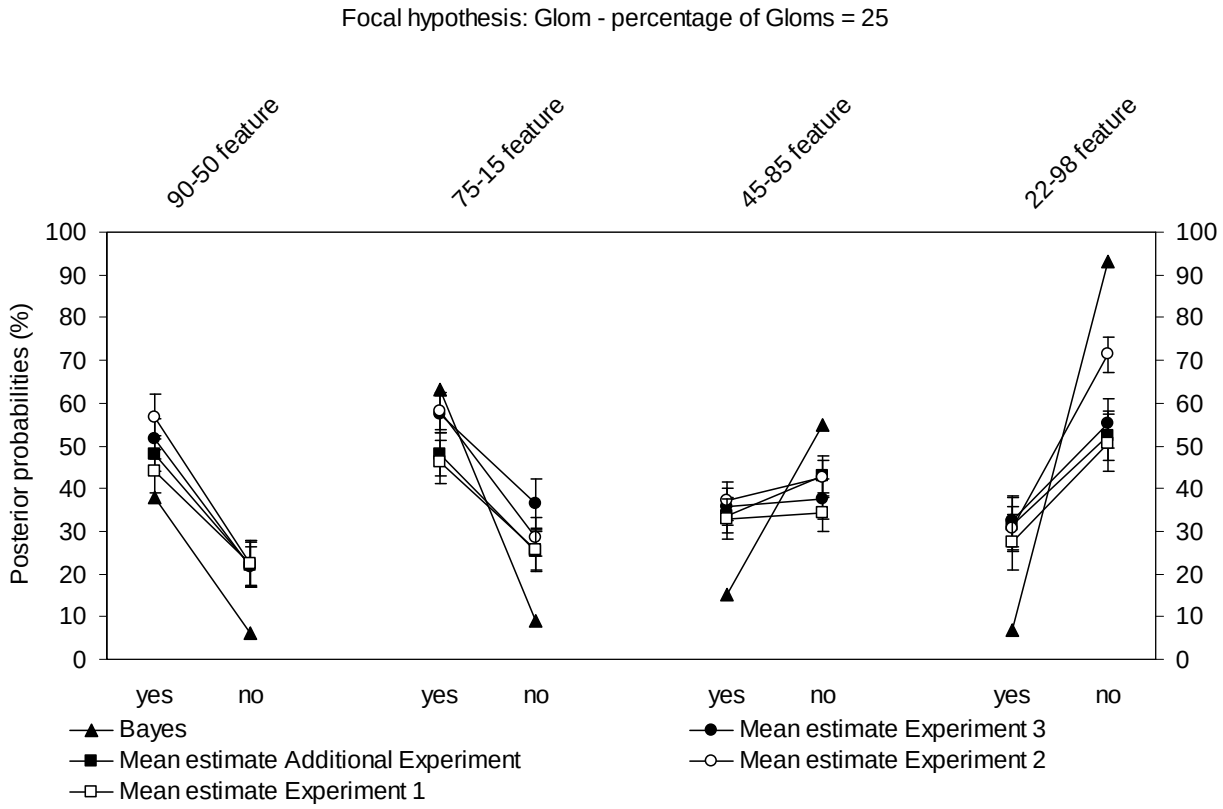
Fixed Effects	Estimate	Std. Error	*t* value	MCMCmean	95% Highest Posterior Density Interval	pMCMC	AIC_i	Δ_i (AIC)	Akaike weights	BIC_i	Δ_i (BIC)	Schwarz weights
Bayesian Diagnosticity	.02	.01	2.69	.02	[.00, .03]	.0080	2545	15	.0002	2572	14	.0003
Log_{10} Diagnosticity	.05	.02	2.43	.05	[.01, .08]	.0140	2543	13	.0007	2571	13	.0005
Information Gain	.62	.23	2.70	.62	[.17, 1.06]	.0070	2531	1	.2690	2558	0	.3281
Kullback-Leibler distance	.27	.23	1.19	.26	[-.18, .68]	.2310	2540	10	.0030	2567	9	.0036
Probability Gain	1.08	.44	2.45	1.07	[.22, 1.93]	.0166	2530	0	.4436	2558	0	.3281
Impact	.34	.55	.62	.34	[-.76, 1.37]	.5242	2542	12	.0011	2569	11	.0013
Measure *L*	.66	.38	1.74	.65	[-.07, 1.38]	.0770	2537	7	.0134	2565	7	.0099
Measure *Z*	.78	.31	2.49	.78	[.18, 1.39]	.0146	2531	1	.2690	2558	0	.3281

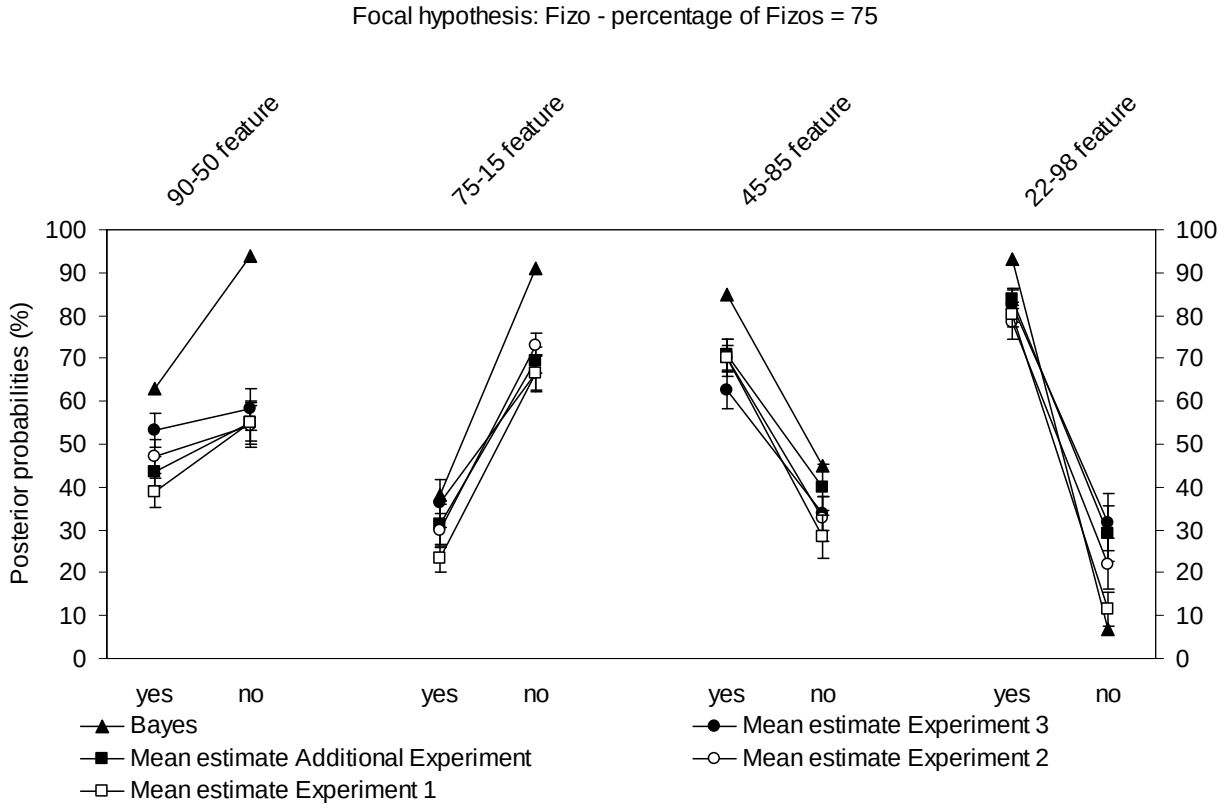
Note. $\Delta_i(AIC) = AIC_i - \min AIC$; $\Delta_i(BIC) = BIC_i - \min BIC$.

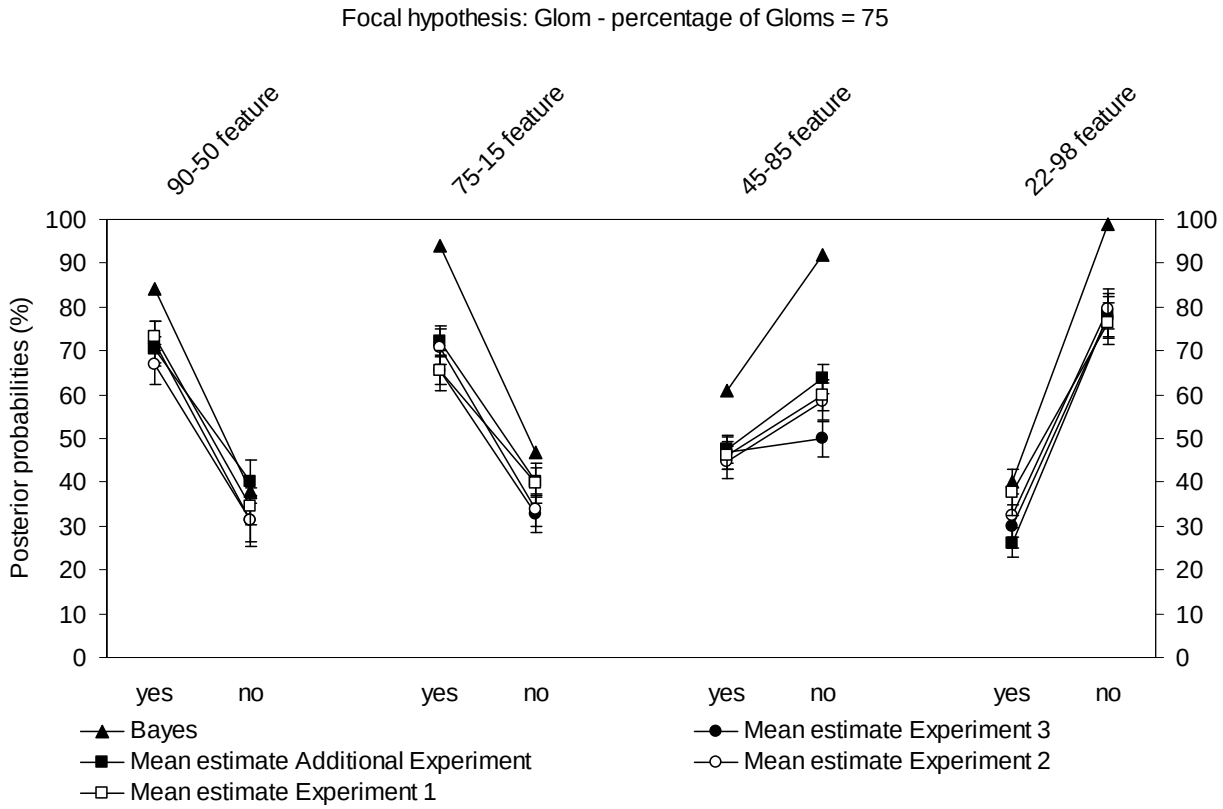
Table 6

Additional Experiment: Normalized Probabilities and Evidence Ratios of Akaike and Schwarz Weights for the Competing Models

Contrasts	Evidence ratio of Akaike weights	Normalized probability of Akaike weights	Evidence ratio of Schwarz weights	Normalized probability of Schwarz weights
Measure *Z* vs. Measure *L*	20.09	.95	33.12	.97
Measure *Z* vs. Bayesian Diagnosticity	1096.63	1.00	1096.63	1.00
Measure *Z* vs. Log_{10} Diagnosticity	403.43	1.00	665.14	1.00
Measure *Z* vs. Information Gain	1.00	.50	1.00	.50
Measure *Z* vs. Kullback-Leibler distance	90.02	.99	90.02	.99
Measure *Z* vs. Probability Gain	.61	.38	1.00	.50
Measure *Z* vs. Impact	244.69	1.00	244.69	1.00
Measures *L* and *Z* vs. OED models	.39	.28	.51	.34


Focal hypothesis: Glom - percentage of Gloms = 25
90-50 feature
75-15 feature
45-85 feature
22-98 feature
Posterior probabilities (%)
100
90
80
70
60
50
40
30
20
10
0
yes
no
Bayes
Mean estimate Additional Experiment
Mean estimate Experiment 1
Mean estimate Experiment 3
Mean estimate Experiment 2


Focal hypothesis: Fizo - percentage of Fizos = 75
90-50 feature
75-15 feature
45-85 feature
22-98 feature
Posterior probabilities (%)
100
90
80
70
60
50
40
30
20
10
0
yes
no
Bayes
Mean estimate Additional Experiment
Mean estimate Experiment 1
Mean estimate Experiment 3
Mean estimate Experiment 2


Focal hypothesis: Glom - percentage of Gloms = 75
90-50 feature
75-15 feature
45-85 feature
22-98 feature
Posterior probabilities (%)
100
90
80
70
60
50
40
30
20
10
0
yes
no
Bayes
Mean estimate Additional Experiment
Mean estimate Experiment 1
Mean estimate Experiment 3
Mean estimate Experiment 2

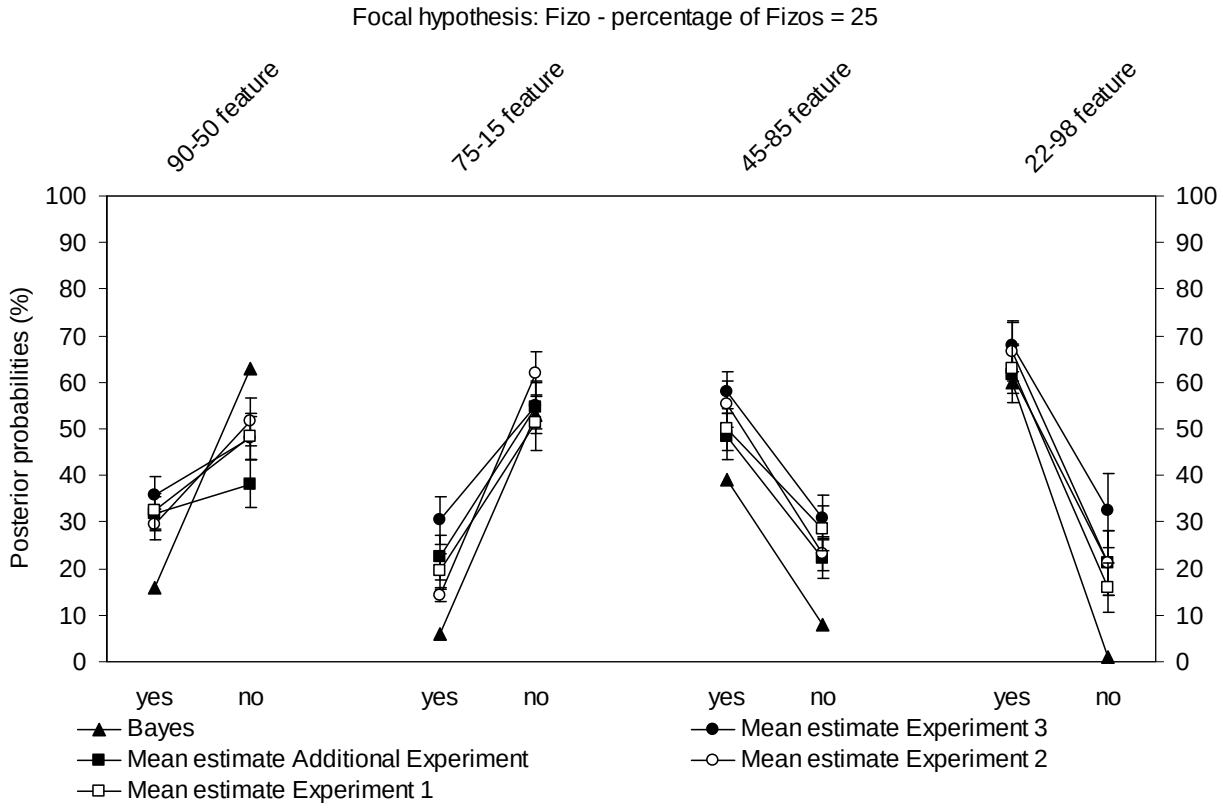

Focal hypothesis: Fizo - percentage of Fizos = 25
90-50 feature
75-15 feature
45-85 feature
22-98 feature
Posterior probabilities (%)
100
90
80
70
60
50
40
30
20
10
0
yes
no
Bayes
Mean estimate Additional Experiment
Mean estimate Experiment 1
Mean estimate Experiment 3
Mean estimate Experiment 2

Figure 1.

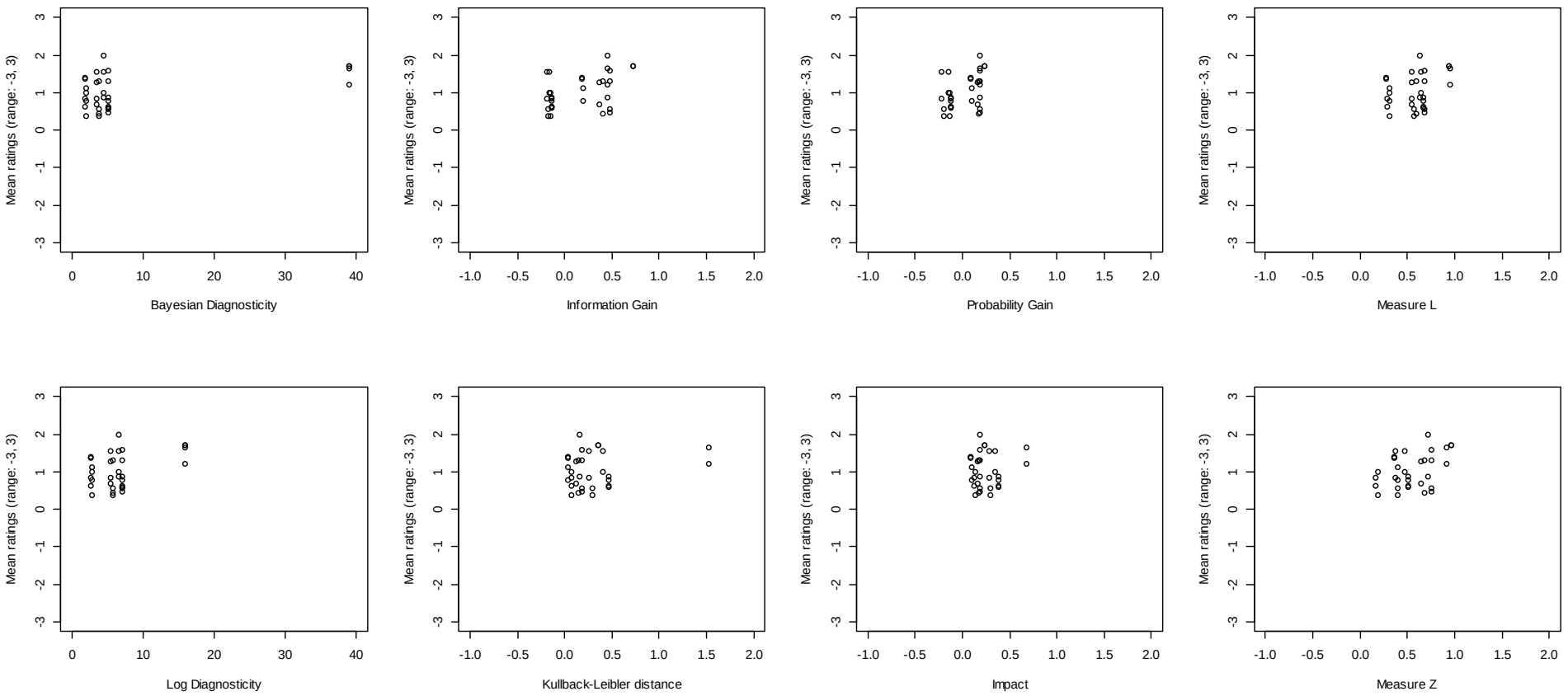


Figure 2. Additional Experiment: Scatterplots illustrating the relationships between participants' mean ratings (averaged across the 32 cells of the experimental design) and the values predicted by the competing models (i.e., the six OED models and *L* and *Z*). The mean ratings are plotted along the Y-axis of each scatter plot, whereas the values predicted by the models are plotted along the X-axis. Each point in the scatterplots represents one of the 32 cells of the experimental design.

The Original Italian Instructions Given to Participants in Experiment 1. Variations Between Experiments are in Italics Within Round Brackets, While Variations Across the Experimental Groups (Within the Same Experiment) are Within Square Brackets. We Provide Here Only One of the Eight Feature-Answer Combinations that We Used.

Ti chiediamo di leggere attentamente lo scenario e le istruzioni che troverai nelle prossime pagine. Volta pagina.

Immagina di viaggiare verso un pianeta, chiamato Vuma. Su questo pianeta esistono due e soltanto due tipi di creature: i Glom e i Fizo. In particolare, il 25% [75%] degli abitanti di Vuma è Glom, mentre il 75% [25%] è Fizo. Dal solo aspetto esterno non è possibile distinguere un Glom da un Fizo. Ti viene chiesto di identificare otto creature diverse che incontri per caso sul pianeta. Glom e Fizo posseggono alcune caratteristiche. Ti verrà riferito (in percentuale) quanti Glom e Fizo hanno queste caratteristiche. Sai di poter porre delle domande alle creature che incontri per stabilire se abbiano o meno una certa caratteristica. Inoltre, sai che entrambi i tipi di creature non mentono mai in risposta a una domanda.

Il tuo compito è, per ogni creatura incontrata, di:

- indicare quanto ritieni che la risposta ricevuta a una domanda su una determinata caratteristica *diminuisca/aumenti la plausibilità dell'ipotesi che la creatura incontrata sia un* ***Glom*** *[****Fizo****] usando una scala da -3 a 3, in cui -3 = diminuisce decisamente e 3 = aumenta decisamente* (Experiment 2: *ti aiuti ad accertarti della possibilità che la creatura incontrata sia un* ***Glom*** *[****Fizo****] usando una scala da -3 a 3, in cui -3 = decisamente inutile e 3 = decisamente utile* / Experiment 3: *ti aiuti a distinguere tra la possibilità che la creatura incontrata sia un Glom e la possibilità che sia un Fizo usando una scala da -3 a 3, in cui -3 = decisamente inutile e 3 = decisamente utile* / Additional Experiment: *diminuisca/aumenti la plausibilità dell'ipotesi che la creatura incontrata sia un Glom o che la creatura incontrata sia un Fizo usando una scala da -3 a 3, in cui -3 = diminuisce decisamente e 3 = aumenta decisamente*);

- indicare, su una scala da 0 a 100, quanto ritieni probabile che la creatura incontrata sia un **Glom** [**Fizo**].

Volta pagina.
Di seguito ti vengono fornite le percentuali di Glom e Fizo sul pianeta Vuma:

Glom 25% [75%]

Fizo 75% [25%]

Nella tabella sottostante ti vengono fornite le percentuali di Glom e Fizo che hanno e che non hanno le branchie:

	Hanno le branchie	Non hanno le branchie
Glom	90%	10%
Fizo	50%	50%

Immagina di trovarti di fronte una creatura e di rivolgerle una domanda. Alla creatura che incontri chiedi: *Hai le branchie?*

La creatura ti risponde: NO.

Quanto ritieni che la risposta ricevuta ("NO") *diminuisca/aumenti la plausibilità dell'ipotesi che la creatura incontrata sia un **Glom** [**Fizo**]* (Experiment 2: *sia d'aiuto per accertarsi della possibilità che la creatura incontrata sia un **Glom** [**Fizo**]* / Experiment 3: *sia d'aiuto per distinguere tra la possibilità che la creatura incontrata sia un Glom e la possibilità che sia un Fizo* / Additional Experiment: *diminuisca/aumenti la plausibilità dell'ipotesi che la creatura incontrata sia un Glom o che la creatura incontrata sia un Fizo*)? (segna un numero da -3 a 3)

-3	-2	-1	0	1	2	3
Experiment 1 and Additional Experiment: *diminuisce decisamente* (Experiment 2 and 3: *decisamente inutile*)						Experiment 1 and Additional Experiment: *aumenta decisamente* (Experiment 2 and 3: *decisamente utile*)

Quanto è probabile che la creatura incontrata sia un **Glom** [**Fizo**]? (scrivi nello spazio sottostante un numero da 0 a 100)

_____

Volta pagina.

Nel giudicare, negli otto casi presentati, *quanto la risposta ricevuta diminuisse/aumentasse la plausibilità dell'ipotesi che la creatura incontrata fosse un Glom [Fizo]* (Experiment 2: *quanto la risposta ricevuta fosse d'aiuto per accertarsi della possibilità che la creatura incontrata fosse un Glom [Fizo]* / Experiment 3: *quanto la risposta ricevuta fosse d'aiuto per distinguere tra la possibilità che la creatura incontrata fosse un Glom e la possibilità che fosse un Fizo* / Additional Experiment: *quanto la risposta ricevuta diminuisse/aumentasse la plausibilità dell'ipotesi che la creatura incontrata fosse un Glom o che la creatura incontrata fosse un Fizo*) (primo giudizio), quanto hai considerato l'informazione relativa alle percentuali di Glom e Fizo sul pianeta Vuma (rispettivamente 25% [75%] e 75% [25%])? (segna un numero da 1 a 7)

1	2	3	4	5	6	7
poco						molto

Nello stimare, negli otto casi presentati, la probabilità che la creatura incontrata fosse un Glom [Fizo] (secondo giudizio), quanto hai considerato l'informazione relativa alle percentuali di Glom e Fizo sul pianeta Vuma (rispettivamente 25% [75%] e 75% [25%])? (segna un numero da 1 a 7)

1	2	3	4	5	6	7
poco						molto

Di seguito puoi scrivere eventuali commenti riguardanti questo studio.

Ti chiediamo ora qualche informazione personale.

Genere: M ☐ F ☐

Età: _____

Nazionalità: _____

Corso di studi: _____

GRAZIE PER AVER PARTECIPATO!