Generalized Information Theory Meets Human Cognition: Introducing a Unified Framework to Model Uncertainty and Information Search

Vincenzo Crupi[1], Jonathan D. Nelson[2,3], Björn Meder[3], Gustavo Cevolani[4], and Katya Tentori[5]

[1] Center for Logic, Language, and Cognition, Department of Philosophy and Education, University of Turin, Italy

[2] School of Psychology, University of Surrey, Guildforfd, UK

[3] Center for Adaptive Behavior and Cognition, Max Planck Institute for Human Development, Berlin, Germany

[4] IMT School for Advanced Studies, Lucca, Italy

[5] Center for Mind / Brain Sciences, University of Trento, Italy

Author Note

Correspondence concerning this article should be addressed to Vincenzo Crupi, Department of Philosophy and Education, University of Turin, via Sant'Ottavio 20, 10124, Torino (Italy), vincenzo.crupi@unito.it. This research was supported by grants CR 409/1-2, NE 1713/1-2, and ME 3717/2-2 from the Deutsche Forschungsgemeinschaft as part of the priority program *New Frameworks of Rationality* (SPP 1516). We thank Nick Chater, Laura Martignon, Andrea Passerini, and Paul Pedersen for helpful comments and exchanges.

Abstract

Searching for information is critical in many situations. In medicine, for instance, careful choice of a diagnostic test can help narrow down the range of plausible diseases that the patient might have. In a probabilistic framework, test selection is often modeled by assuming that people's goal is to reduce uncertainty about possible states of the world. In cognitive science, psychology, and medical decision making, Shannon entropy is the most prominent and most widely used model to formalize probabilistic uncertainty and the reduction thereof. However, a variety of alternative entropy metrics (Hartley, Quadratic, Tsallis, Rényi, and more) are popular in the social and the natural sciences, computer science, and philosophy of science. Particular entropy measures have been predominant in particular research areas, and it is often an open issue whether these divergences emerge from different theoretical and practical goals or are merely due to historical accident. Cutting across disciplinary boundaries, we show that several entropy and entropy reduction measures arise as special cases in a unified formalism, the Sharma-Mittal framework. Using mathematical results, computer simulations, and analyses of published behavioral data, we discuss four key questions: How do various entropy models relate to each other? What insights can be obtained by considering diverse entropy models within a unified framework? What is the psychological plausibility of different entropy models? What new questions and insights for research on human information acquisition follow? Our work provides several new pathways for theoretical and empirical research, reconciling apparently conflicting approaches and empirical findings within a comprehensive and unified information-theoretic formalism.

KEYWORDS:
Entropy, Uncertainty, Value of information, Information search, Probabilistic models

Generalized Information Theory Meets Human Cognition: Introducing a Unified Framework to Model Uncertainty and Information Search

1. Introduction

A key topic in the study of rationality, cognition, and behavior is the effective search for relevant information or evidence. Information search is also closely connected to the notion of uncertainty. Typically, an agent will seek to acquire information to reduce uncertainty about an inference or decision problem. Physicians prescribe medical tests in order to handle arrays of possible diagnoses. Detectives seek witnesses in order to identify the culprit of a crime. And, of course, scientists gather data in order to discriminate among different hypotheses.

In psychology and cognitive science, most early work on information acquisition adopted a logical, deductive inference perspective. In the spirit of Popper's (1959) influential falsificationist philosophy of science, the idea was that learners should seek information that could help them falsify hypotheses (e.g., expressed as a conditional or a rule; Wason, 1960, 1966, 1968). However, many human reasoners did not seem to believe that information is useful if and only if it can potentially rule out (falsify) a hypothesis. From the 1980s, cognitive scientists started analyzing human information search with a closer look at inductive inference, using probabilistic models to quantify the value of information and endorsing them as normative benchmarks (e.g., Baron, 1985; Klayman & Ha, 1987; Skov & Sherman, 1986; Slowiaczek, Klayman, Sherman, & Skov, 1992; Trope & Bassok, 1982, 1983). This research was inspired by seminal work in philosophy of science (e.g. Good, 1950), statistics (e.g. Lindley, 1956), and decision theory (Savage, 1972). In this view, each outcome of a query could modify an agent's beliefs about the hypotheses being considered, thus providing some amount of information. For instance, the key theoretical point of Oaksford and Chater's (1994, 2003) analysis of Wason's selection task was to conceptualize information acquisition as a piece of probabilistic inductive reasoning, assuming that people's goal is to reduce uncertainty about whether a rule holds or not. In a similar vein, researchers in vision science have used measures of uncertainty reduction to predict visual queries for gathering

information (i.e., eye movements; Legge, Klitz, & Tjan, 1997; Najemnik & Geisler, 2005, 2009; Nelson & Cottrell, 2007; Renninger, Coughlan, Verghese, & Malik, 2005), or to guide a robot's eye movements (Denzler & Brown, 2002). Probabilistic models of uncertainty reduction have also been used to predict human query selection in causal reasoning (Bramley, Lagnado, & Speekenbrink, 2015), hypothesis testing (Austerweil & Griffiths, 2011; Navarro & Perfors, 2011; Nelson, Divjak, Gudmundsdottir, Martignon, & Meder, 2014; Nelson, Tenenbaum, & Movellan, 2001), and categorization (Meder & Nelson, 2012; Nelson, McKenzie, Cottrell, & Sejnowski, 2010).

If reducing uncertainty is a major cognitive goal and motivation for information acquisition, a critical issue is how uncertainty and the reduction thereof can be represented in a rigorous manner. A fruitful approach to formalize uncertainty is using the mathematical notion of *entropy*, which in turn generates a corresponding model of the informational utility of an experiment as the *expected reduction* of entropy (uncertainty), sometimes called *expected information gain*.

In many disciplines, including psychology and neuroscience (Hasson, 2016), the most prominent model is Shannon (1948) entropy. However, a number of non-equivalent measures of entropy have been suggested, and are being used, in a variety of research domains. Examples include the application of Quadratic entropy in ecology (Lande, 1996), the family of Rényi (1961) entropies in computer science and image processing (Boztas, 2014; Sahoo & Arora, 2004), and Tsallis entropies in physics (Tsallis, 2011). It is currently unknown whether these other entropy models would have potential to address key theoretical and empirical questions in cognitive science. Here, we bring together these different models in a comprehensive theoretical framework, the Sharma-Mittal formalism (from Sharma & Mittal, 1975), which incorporates a large number of prominent entropy measures as special cases. Careful consideration of the formal properties of this family of entropy measures will reveal important implications for modeling uncertainty and information search behavior. Against this rich theoretical background, we will draw on existing behavioral data and novel simulations to explore how different models relate to each other, elucidate their psychological meaning and plausibility, and show how they can generate new testable predictions.

The remainder of this paper is organized as follows. We begin by spelling out what an entropy measure is and how it can be employed to represent uncertainty and the informational value of queries (questions, tests, experiments) (section 2.). Subsequently, we review four representative and influential definitions of entropy, namely Quadratic, Hartley, Shannon, and Error entropy (3.). These models have been, and continue to be, of importance in different areas of research. In the main theoretical section of the paper, we describe a unified formal framework generating a biparametric continuum of entropy measures. Drawing on work in generalized information theory, we show that many extant models of entropy and expected entropy reduction can be embedded in this comprehensive formalism (4.). We provide a number of new mathematical results in this section. We also address the theoretical meaning of the parameters involved when the target domain of application is human reasoning, with implications for both normative and descriptive approaches. We then further elaborate on the connection with experimental research in several ways. First, we present simulation results from an extensive exploration of information search decision problems in which alternative models provide strongly diverging, empirically testable predictions (5.). Second, we report and discuss an overarching analysis of the information-theoretic account of the most widely known experimental paradigm for the study of information gathering, i.e., Wason's (1966, 1968) abstract selection task (6.1.). Then we investigate which models perform better against data from a range of experience-based studies on human information search behavior (Meder & Nelson, 2012; Nelson et al., 2010) (6.2.). We also point out that some entropy models from this framework offer potential explanation of human information search behavior in experiments where probabilities are conveyed through words and numbers, which to date have been perplexing to account for theoretically (6.3). Finally, we show that new models offer a theoretically satisfying and descriptively adequate unification of disparate results across different kinds of tasks (6.4.). In the General Discussion (7.), we outline and assess the prospects of a generalized information-theoretic framework for guiding the study of human inference and decision making.

Part of our discussion relies and elaborates on mathematical analyses, including novel results. Moreover, although a number of the mathematical points in the paper can be found scattered through the mathematics and physics literature, here we bring them together

systematically. We provide Supplementary Materials where non-trivial derivations are given according to our unified notation. Throughout each section of the text, statements requiring a mathematical proof are flagged by square brackets [Suppl Mat], and the proof is then presented in the corresponding subsection of the Supplementary Materials file. Among the formal results provided that are novel to the best of our knowledge, the following we find especially important: the ordinal equivalence of Sharma-Mittal entropy measures of the same order (proof in Suppl Mat, section 4), the additivity of all Sharma-Mittal measures of expected entropy reduction for sequential tests (again Suppl Mat, 4), and the distinctive role of the degree parameter in information search tasks such as the Person Game (Suppl Mat ,5). Further novel results include the subsumption of diverse models such as the Arimoto (1971) and the Power entropies within the Sharma-Mittal framework (Suppl Mat, 3), and the specification of how a number of different entropy measures can be construed within the general theory of means (Table 4).

2. Entropies, uncertainty, and information search

According to a well-known anecdote, the origins of information theory were marked by a witty joke of John von Neumann. Claude Shannon was doubtful how to call the key concept of his groundbreaking work on the "mathematical theory of communication" (Shannon, 1948). "You should call it *entropy*," von Neumann suggested. Of course, von Neumann must have been aware of the close connections between Shannon's formula and Boltzmann's definition of entropy in classical statistical mechanics. But the most important reason for his suggestion, von Neumann quipped, was that "nobody knows what entropy really is, so in a debate you will always have the advantage" (see Tribus & McIrvine, 1971). Shannon accepted the advice. Several decades later, von Neumann's remark seems even more pointed, if anything. Influential observers have voiced caution and concern about the proliferation of mathematical analyses of entropy and related notions (Aczél, 1984, 1987). Meanwhile, many applications have been developed, for instance in physics and ecology (see, e.g., Beck, 2009; Keylock, 2005). But recurrent theoretical controversies have arisen, too, along with occasional complaints of conceptual confusion (see Cho, 2002, and Jost, 2006, respectively).

Luckily, these thorny issues will be tangential to our main concerns. Although a given formalization of entropy can be considered for the representation and measurement of different constructs in each of a variety of domains, we focus on one target concept for which entropies can be employed, namely the *uncertainty* concerning a variable X given a probability distribution P. In this regard, the key question is the following: *How much uncertainty is conveyed about variable X by a given probability distribution P*? This notion is central to the normative and descriptive study of human cognition.

Suppose, for instance, that an infection can be caused by three different types of virus, and label x_1, x_2, x_3 the corresponding possibilities. Consider two different probability assignments, such as, say:

$P(x_1) = 0.49$, $P(x_2) = 0.49$, $P(x_3) = 0.02$

and

$P^*(x_1) = 0.70$, $P^*(x_2) = 0.15$, $P^*(x_3) = 0.15$

Is the uncertainty about $X = \{x_1, x_2, x_3\}$ greater under P or under P^*? An entropy measure enables us to give precise quantitative values in both case, and hence a clear answer. Importantly, however, the answer will often be measure-dependent, for different entropy measures convey different ideas of uncertainty and exhibit distinct mathematical properties of theoretical interest. We will see this in detail later on.

Once uncertainty as our conceptual target has been outlined, we can turn to entropy as a mathematical object. Consider a finite set X of n mutually exclusive and jointly exhaustive possibilities $x_1, \ldots, x_n$ on which a probability distribution $P(X)$ is defined, so that $P(X) = \{P(x_1), \ldots, P(x_n)\}$, with $P(x_i) \geq 0$ for any i $(1 \leq i \leq n)$ and $\sum_{x_i \in X} P(x_i) = 1$. The n elements in $X = \{x_1, \ldots, x_n\}$ can be taken as representing different kinds of entities, such as events, categories, or propositions. For our purposes, *ent* is an entropy measure if it is a function f of the relevant probability values only, i.e.:

$$ent_P(X) = f[P(x_1), \ldots, P(x_n)]$$

and function f satisfies a small number of basic properties (see below). Notice that, in general, an entropy function can be readily extended to the case of a conditional probability

distribution given some datum y. In fact, under the conditional probability distribution $P(X|y)$, one has $ent_P(X|y) = f[P(x_1|y), \ldots, P(x_n|y)]$.

Shannon entropy has been so prominent in cognitive science that some readers will ask: why we do not just stick with it? More specific objections in this vein include that Shannon entropy is uniquely axiomatically motivated, that Shannon entropy is already central to psychological theory of the value of information, or that Shannon entropy is optimal in certain applied situations. Each objection can be addressed separately. First, a number of entropy metrics in our generalized framework (not only Shannon) have been or can be uniquely derived from specific sets of axioms (see Csiszár, 2008). Second, although Shannon entropy has a number intuitively desirable properties, it is not a serious competitive descriptive psychological model of the value of information in some tasks (e.g., Nelson et al., 2010). Third, several published papers in applied domains report superior performance when other entropy measures are used (e.g., Ramírez-Reyes et al., 2016). Indeed, Shannon's (1948) own view was that although axiomatic characterization can lend plausibility to measures of entropy and information, "the real justification" (p. 393) rests on the measures' operational relevance. A generalized mathematical framework can increase our theoretical understanding of the relationships among different measures, unify diverse psychological findings, and generate novel questions for future research.

Scholars have used different properties as defining an entropy measure (see, e.g., Csizsár, 2008). Besides some usual technical requirement (like non-negativity), a key idea is that entropy should be appropriately sensitive to how even or uneven a distribution is, at least with respect to the extreme cases of an uniform probability function, $U(X) = \{1/n, \ldots, 1/n\}$, or of a deterministic function $V(X)$ where $V(x_i) = 1$ for some i ($1 \leq i \leq n$) and 0 for all other xs. (In the latter case, the distribution actually reflects a truth-value assignment, in logical parlance.) In our setting, $U(X)$ represents the highest possible degree of uncertainty about X, while under $V(X)$ the true value of X is known for sure, and no uncertainty is left. Hence it must hold that, for any X and $P(X)$, $ent_U(X) \geq ent_P(X) \geq ent_V(X)$, with at least one inequality strict. This basic and minimal condition we label *evenness sensitivity*. It is conveyed by Shannon entropy as well as many others, as we shall see, and it guarantees, for instance, that entropy is strictly higher for, say, a distribution like $\{1/3, 1/3, 1/3\}$ than for $\{1,0,0\}$.

Once the idea of an entropy measure is characterized, one can study different measures of expected entropy reduction. This amounts to considering *two* variables *X* and *Y*, and defining the expected reduction of the initial entropy of *X* across the elements of *Y*. To illustrate, in the viral infection example mentioned above, *X* may concern the type of virus actually involved, while *Y* could be some clinically observable marker (like the result of a blood test) which is informationally relevant for *X*. Mathematically, given a joint probability distribution $P(X,Y)$ over the combination of two variables *X* and *Y* (i.e., their Cartesian product $X \times Y$), the actual change in entropy about *X* determined by an element *y* in *Y* can be represented as $\Delta ent_P(X, y) = ent_P(X) - ent_P(X|y)$. Accordingly, the expected reduction of the initial entropy of *X* across the elements of *Y* can be computed in a standard way, as follows:[1]

$$R_P(X, Y) = \sum_{y_j \in Y} \Delta ent_P\big(X, y_j\big)\, P(y_j)$$

The notation $R_P(X, Y)$ is adapted from work on the foundations of Bayesian statistics, where the expected reduction in entropy is seen as measuring the *dependence* of variable *X* on variable *Y*, or of the *relevance* of *Y* for *X* (see, e.g., Dawid & Musio, 2014).

Very much as for entropy itself, the expected reduction of entropy remains as general and neutral a notion as possible. *R* measures, too, can be given different interpretations in different domains. In many contexts, it is plausibly assumed that reduction of the uncertainty is a major dimension of the purely informational (or epistemic) value of the search for more data. We will thus consider a measure *R* as providing a formal approach to questions of the following kind: *Given X as a target of investigation, what is the expected usefulness of finding out about Y from a purely informational point of view*? Hence, the notion of uncertainty is tightly coupled to the rational assessment of the expected informational utility of pursuing a given *search* for additional evidence (performing a query, executing a test, running an experiment). (See Crupi & Tentori, 2014; Nelson, 2005, 2008. For more discussion, also see Evans & Over, 1996; Roche & Shogenji, 2016.)

[1] For technical reasons, we will assume $P(y_j) > 0$ for any *j*. This is largely a safe proviso for our current purposes. In fact, in our setting with both *X* and *Y* finite sets, any zero probability outcome in *Y* could just be omitted.

Formally, X and Y can just be seen as partitions of possibilities. In this interpretation, however, they play quite different roles in $R_P(X, Y)$. The first argument, X, represents the overall goal of the inquiry, while the second, Y, is supposed to be directly accessible to the information seeker. In a typical application, Y will be more or less useful a test to learn about target X, although unable to conclusively establish what the true hypothesis in X is.

Table 1. *Notation employed*.

Notation	**Description**
$H = \{h_1, \ldots, h_n\}$	A partition of n possibilities (or hypothesis space).
$P(H)$	Probability distribution P defined over the elements of H.
$P(H\|e)$	Probability distribution P defined over the elements of H conditional on e.
$U(H)$	Uniform probability distribution over the elements of H.
$V(H)$	A probability distribution such that $V(h_i) = 1$ for some i ($1 \leq i \leq n$) and 0 for all other hs.
$H{\times}E$	The variable obtained by the combination (Cartesian product) of variables H and E.
$P(H,E)$	Joint probability distribution over the combination of variables H and E.
$H \perp_P E$	Given $P(H,E)$, variables H and E are statistically independent.
$H \perp_P E \mid F$	Given $P(H,E,F)$, variables H and E are statistically independent conditional on each element in F.
$ent_P(H)$	Entropy of H given $P(H)$.
$ent_P(H\|e)$	Conditional entropy of H on e given $P(H\|e)$.
$\Delta ent_P(H, e)$	Reduction of the initial entropy of H provided by e, i.e., $ent_P(H) - ent_P(H\|e)$.
$R_P(H, E)$	Expected reduction of the entropy of H across the elements of E, given $P(H,E)$.
$R_P(H, E\|f)$	Expected reduction of the entropy of H across the elements of E, given $P(H,E\|f)$.
$R_P(H, E\|F)$	Expected value of $R_P(H, E\|f)$ across the elements of F, given $P(H,E,F)$.
$ln_t(x)$	The Tsallis generalization of the natural logarithm (with parameter t).
$e_t(x)$	The Tsallis generalization of the ordinary exponential (with parameter t).

In general, the occurrence of one particular element *y* of *Y* does not need to reduce the initial entropy about *X*; it might as much increase it, hence making $\Delta ent_P(X, y)$ negative. This quantity can be negative if (for instance) datum *y* changes probabilities from $P(X) = \{0.9, 0.1\}$ to $P(X|y) = \{0.6, 0.4\}$. But can $R_P(X, Y)$, i.e., the *expected* informational usefulness of *Y* for learning about *X*, be negative? Some *R* measures are strictly non-negative, but others can in fact be negative in the expectation; this depends on key properties of the underlying entropy measure, as we discuss later on.

To summarize, in the domain of human cognition, probability distributions can be employed to represent an agent's degrees of belief (be they based on objective statistical information or subjective confidence), with entropy $ent_P(X)$ providing a formalization of the uncertainty about *X* (given *P*). Relying on the reduction of uncertainty as an informational utility, $R_P(X, Y)$ is then interpreted as a measure of the expected usefulness of a query (test, experiment) *Y* relative to a target hypothesis space *X*. From now on, to emphasize this interpretation, we will often use $H = \{h_1, \ldots, h_n\}$ to denote a hypothesis set of interest and $E = \{e_1, \ldots, e_m\}$ for a possible search for evidence. Table 1 summarizes our terminology in this respect as well as for the subsequent sections.

3. Four Influential Entropy Models

We will now briefly review four important models of entropy and the corresponding models of expected entropy reduction.

3.1. Quadratic entropy

Entropy / Uncertainty. Some interesting entropy measures were originally proposed long before the exchange between Shannon and von Neumann, when *entropy* was not yet a scientific term outside statistical thermodynamics. Here is one major instance:

$$ent_P^{Quad}(H) = 1 - \sum_{h_i \in H} P(h_i)^2$$

Labeled *Quadratic entropy* in Vajda and Zvárová (2007), this measure is widely known as the *Gini* (or *Gini-Simpson*) *index*, after Gini (1912) and Simpson (1949) (also see Gibbs & Martin,

1962). It is often employed as an index of *biological diversity* (see, e.g., Patil & Taille, 1982) and sometimes spelled out in the following equivalent formulation:

$$ent_P^{Quad}(H) = \sum_{h_i \in H} P(h_i)(1 - P(h_i))$$

The above formula suggests a meaningful interpretation with *H* amounting to a partition of hypotheses considered by an uncertain agent. In this reading, ent^{Quad} computes the average (expected) *surprise* that the agent would experience in finding out what the true element of *H* is, given 1 – *P*(*h*) as a measure of the surprise that arises in case *h* obtains (see Crupi & Tentori, 2014).[2]

Entropy reduction / Informational value of queries. Quadratic entropy reduction, namely, $\Delta ent_P^{Quad}(H, e) = ent_P^{Quad}(H) - ent_P^{Quad}(H|e)$, has been occasionally mentioned in philosophical analyses of scientific inference (Niiniluoto & Tuomela, 1973, p. 67). In turn, its associated expected reduction measure, $R_P^{Quad}(H, E) = \sum_{e_j \in E} \Delta ent_P^{Quad}\big(H, e_j\big)\, P(e_j)$, was applied by Horwich (1982, pp. 127-129), again in formal philosophy of science, and studied in computer science by Raileanu and Stoffel (2004).

3.2. Hartley entropy

Entropy / Uncertainty. Gini's work did not play any apparent role in the development of Shannon's (1948) theory. A seminal paper by Hartley (1928), however, was a starting point for Shannon's analysis. One lasting insight of Hartley was the introduction of logarithmic functions, which have become ubiquitous in information theory ever since. As Hartley also realized, the choice of a base for the logarithm is a matter of conventionally setting a unit of measurement (Hartley, 1928, pp. 539-541). Throughout our discussion, we will employ the natural logarithm, denoted as *ln*.

[2] ent^{Quad} also quantifies the overall expected *inaccuracy* of probability distribution *P*(*H*) as measured by the so-called Brier score (i.e., the squared Euclidean distance from the possible truth-value assignments over *H*; see Brier, 1950; Leitgeb & Pettigrew, 2010a,b; Pettigrew, 2013; Selten, 1998). Festa (1993, 137 ff.) also gives a useful discussion of Quadratic entropy in the philosophy of science, including Carnap's (1952) classical work in inductive logic.

Inspired by Hartley's (1928) original idea that the information provided by the observation of one among *n* possible values of a variable is increasingly informative the larger *n* is, and that it immediately reflects the entropy of that variable, one can define the Hartley entropy as follows (Aczél, Forte, and Ng, 1974):

$$ent_P^{Hartley}(H) = ln\left[\sum_{h_i \in H} P(h_i)^0\right]$$

Under the convention $0^0 = 0$ (which is standard in the entropy literature), and given that $P(h_i)^0 = 1$ whenever $P(h_i) > 0$, $ent^{Hartley}$ computes the logarithm of the number of all non-null probability elements in *H*.

Entropy reduction / Informational value of queries. When applied to the domain of reasoning and cognition, the implications of Hartley entropy reveal an interesting Popperian flavor. A piece of evidence *e* is useful, it turns out, only to the extent that it excludes ("falsifies") at least some of the hypotheses in *H*, for otherwise the reduction in Hartley entropy, $\Delta ent_P^{Hartley}(H, e) = ent_P^{Hartley}(H) - ent_P^{Hartley}(H|e)$, is just zero. An agent adopting such a measure of informational utility would then only value a test outcome, *e*, insofar as it conclusively rules out at least one hypothesis in *H*. If no possible outcome in *E* is potentially a "falsifier" for some hypothesis in *H*, then the expected reduction of Hartley entropy, $R^{Hartley}$, is also zero, implying that query *E* has no expected usefulness at all with respect to *H*.

3.3. Shannon entropy

Entropy / Uncertainty. In many contexts, the notion of entropy is simply and immediately equated to Shannon's formalism. Overall, such special consideration is well-deserved and motivated by countless applications spread over virtually all branches of science. The form of Shannon entropy is fairly well-known:

$$ent_P^{Shannon}(H) = \sum_{h_i \in H} P(h_i) ln\left(\frac{1}{P(h_i)}\right)$$

Concerning the interpretation of the formula, many points made earlier for quadratic entropy apply to Shannon entropy too, given relevant adjustments. In fact, $ln(1/P(h))$ is another

measure of the surprise in finding out that a state of affairs h obtains, and thus $ent^{Shannon}$ is its overall expected value relative to H.[3]

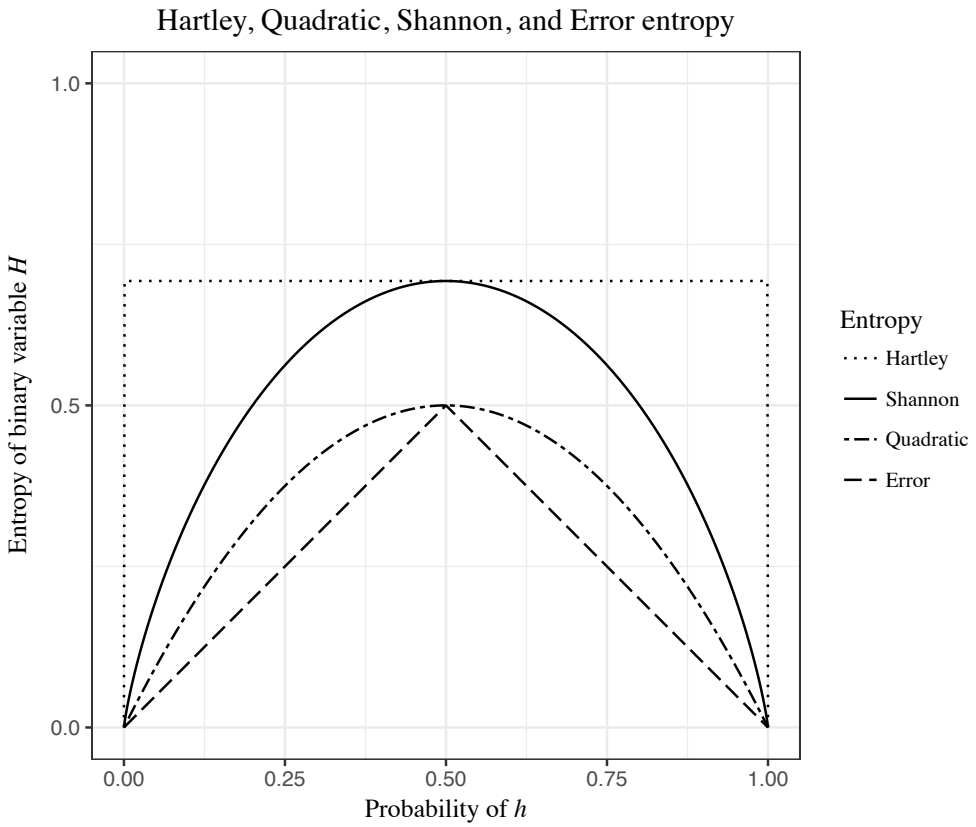

Figure 1. A graphical illustration of Quadratic, Hartley, Shannon, and Error entropy as distinct measures of uncertainty over a binary hypothesis set $H = \{h,\overline{h}\}$ as a function of the probability of h.

Entropy reduction / Informational value of queries. The reduction of Shannon entropy, $\Delta ent_P^{Shannon}(H,e) = ent_P^{Shannon}(H) - ent_P^{Shannon}(H|e)$, is sometimes called *information gain* and it is often considered as a measure of the informational utility of a datum e. Its expected value, also called *expected information gain*, $R_P^{Shannon}(H,E) = \sum_{e_j \in E} \Delta ent_P^{Shannon}\big(H,e_j\big)\, P(e_j)$, is then viewed as a measure of usefulness of query E for learning about H. (See, e.g.,

[3] The quantity $ln(1/P(h))$ also characterizes a popular approach to the measurement of the inaccuracy of probability distribution $P(H)$ when h is the true element in H (so-called logarithmic score), and $ent^{Shannon}$ can be seen as computing the expected inaccuracy of $P(H)$ accordingly (see Good, 1952; also see Gneiting & Raftery, 2007).

Austerweil & Griffiths, 2011; Bar-Hillel & Carnap, 1953; Lindley, 1956; Oaksford & Chater, 1994, 2003, and Ruggeri & Lombrozo, 2015; also see Benish, 1999, and Nelson, 2005, 2008, for more discussion.)

3.4. Error entropy

Entropy / Uncertainty. Given a distribution $P(H)$ and the goal of predicting the true element of H, a rational agent would plausibly select h^* such that $P(h^*) = \max_{h_i \in H}[P(h_i)]$, and $1 - \max_{h_i \in H}[P(h_i)]$ would then be the *probability of error*. Since Fano's (1961) seminal work, this quantity has received considerable attention in information theory. Also known as Bayes's error, we will call this quantity *Error entropy*:

$$ent_P^{Error}(H) = 1 - \max_{h_i \in H}[P(h_i)]$$

Note that ent^{Error} is only concerned with the largest value in the distribution $P(H)$, namely $\max_{h_i \in H}[P(h_i)]$. The lower that value, the higher the chance of error were a guess to be made, thus the higher the uncertainty about H.

Entropy reduction / Informational value of queries. Unlike the other models above, Error entropy has seldom been considered in the natural or social sciences. However, it can be taken as a sound basis for the analysis of rational behavior. In the latter domain, it is quite natural to rely on the reduction of the expected probability of error $\Delta ent_P^{Error}(H, e) = ent_P^{Error}(H) - ent_P^{Error}(H|e)$ as the utility of a datum (often labelled *probability gain*; see Baron, 1985; Nelson, 2005, 2008) and on its expected value, $R_P^{Error}(H, E) = \sum_{e_j \in E} \Delta ent_P^{Error}\left(H, e_j\right) P(e_j)$, as the usefulness of a query or test. Indeed, there are important occurrences of this model in the study of human cognition.[4]

[4] An early example is Baron's (1985, ch, 4) presentation of R^{Error}, following Savage (1972, ch. 6). Experimental investigations on whether R^{Error} can account for actual patterns of reasoning include Baron, Beattie, and Hershey (1988), Bramley, Lagnado, and Speekenbrink (2015), Meder and Nelson (2012), Nelson, McKenzie, Cottrell, and Sejnowski (2010), and Rusconi, Marelli, D'Addario, Russo, and Cherubini (2014), while Crupi, Tentori, and Lombardi (2009) relied on R^{Error} in their critical analysis of so-called pseudodiagnosticity (also see Crupi & Girotto, 2014; Tweeney, Doherty, & Kleiter, 2010).

4. A Unified Framework for Uncertainty and Information Search

The set of models introduced above represents a diverse sample in historical, theoretical, and mathematical terms (see Figure 1 for a graphical illustration). Is the prominence of particular models due to fundamental distinctive properties, or largely due to historical accident? What are the relationships among these models? In this section we show how all of these models can be embedded in a unified mathematical formalism, providing new insight.

4.1. Sharma-Mittal entropies

Let us take Shannon entropy again as a convenient starting point. As noted above, Shannon entropy is an average, more precisely a *self-weighted* average, displaying the following structure:

$$\sum_{h_i \in H} P(h_i)\ inf[P(h_i)]$$

The label *self-weighted* indicates that each probability $P(h)$ serves as a weight for the value of function *inf* having that same probability as its argument, namely, $inf[P(h)]$. The function *inf* can be seen as capturing a notion of *atomic information* (or *surprise*), assigning a value to each distinct element of H on the basis of its own probability (and nothing else). An obvious requirement here is that *inf* should be a decreasing function, because a finding that was antecedently highly probable (improbable) provides little (much) new information (an idea that Floridi, 2013, calls "inverse relationship principle" after Barwise, 1997, p. 491). In Shannon entropy, one has $inf(x) = ln(1/x)$. Given $inf(x) = 1 - x$, instead, Quadratic entropy arises from the very same scheme above.

A self-weighted average is a special case of a generalized (self-weighted) mean, which can be characterized as follows:

$$g^{-1}\left\{\sum_{h_i \in H} P(h_i) g\{inf[P(h_i)]\}\right\}$$

where g is a differentiable and strictly increasing function (see Wang & Jiang, 2005; also see Muliere & Parmigiani, 1993, for the fascinating history of these ideas). For different choices of g, different kinds of (self-weighted) means are instantiated. With $g(x) = x$, the weighted average above obtains once again. For another standard instance, $g(x) = 1/x$ gives rise to the

harmonic mean. Let us now consider the form of generalized (self-weighted) means above and focus on the following setting:

$g(x) = ln_r[e_t(x)]$

$inf(x) = ln_t(1/x)$

where

$ln_t(x) = \frac{x^{(1-t)}-1}{1-t}$

$e_t^x = [1 + (1-t)x]^{\frac{1}{1-t}}$

are generalized versions of the natural logarithm and exponential functions, respectively, often associated with Tsallis's (1988) work. Importantly, the ln_t function recovers the ordinary natural logarithm ln in the limit for $t \to 1$, so that one can safely equate $ln_t(x) = ln(x)$ for $t = 1$ and have a nice and smooth generalized logarithmic function.[5] Similarly, it is assumed that $e_t^x = e^x$ for $t = 1$, as this is the limit for $t \to 1$ [Suppl Mat, section 1]. Negative values of parameters r and t will not need concern us here: we'll be assuming $r,t \geq 0$ throughout.

Once fed into the generalized means equation, these specifications of $inf(x)$ and $g(x)$ yield a two-parameter family of entropy measures of *order r* and *degree t* [Suppl Mat, 2]:

$$ent_P^{SM(r,t)}(H) = \frac{1}{t-1}\left[1 - \left(\sum_{h_i \in H} P(h_i)^r\right)^{\frac{t-1}{r-1}}\right]$$

The label *SM* refers to Sharma and Mittal (1975), where this formalism was originally proposed (also see Masi, 2005, and Hoffmann, 2008). All functions in the Sharma-Mittal family are evenness sensitive (see 2. above), thus in line with a basic characterization of entropies [Suppl Mat, 2]. Also, with $ent^{SM(r,t)}$ one can embed the whole set of four classic measures in our initial list. More precisely [Suppl Mat, 3]:

[5] The idea of ln_t is often credited to Tsallis for his work in generalized thermodynamics (see Tsallis, 1988, and 2011). The mathematical point may well go back to Euler, however (see Hoffmann, 2008, p. 7). For more theory, also see Havrda and Charvát (1967), Daróczy (1970), Naudts (2002), Kaniadakis, Lissia, and Scarfone (2004).

- Quadratic entropy can be derived from the Sharma-Mittal family for $r = t = 2$, that is, $ent_P^{SM_{(2,2)}}(H) = ent_P^{Quad}(H)$;
- Hartley entropy can be derived from the Sharma-Mittal family for $r = 0$ and $t = 1$, that is, $ent_P^{SM_{(0,1)}}(H) = ent_P^{Hartley}(H)$;
- Shannon entropy can be derived from the Sharma-Mittal family for $r = t = 1$, that is, $ent_P^{SM_{(1,1)}}(H) = ent_P^{Shannon}(H)$;
- Error entropy is recovered from the Sharma-Mittal family in the limit for $r \to \infty$ when $t = 2$, so that we have $ent_P^{SM_{(\infty,2)}}(H) = ent_P^{Error}(H)$.

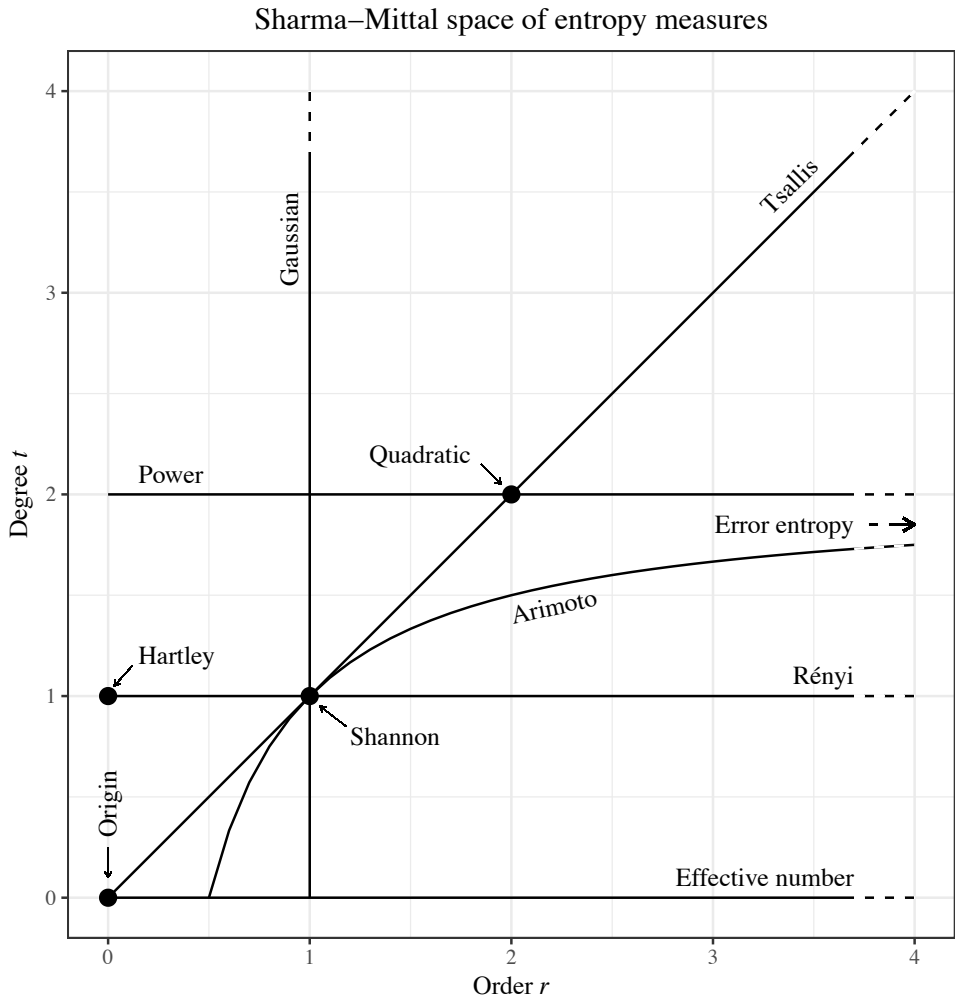

Figure 2. The Sharma-Mittal family of entropy measures is represented in a Cartesian quadrant with values of the order parameter *r* and of the degree parameter *t* lying on the *x*– and *y*–axis, respectively. Each point in the quadrant corresponds to a specific entropy measure, each line corresponds to a distinct one-parameter generalized entropy function. Several special cases are highlighted. (Relevant references and formulas are listed in Table 4.)

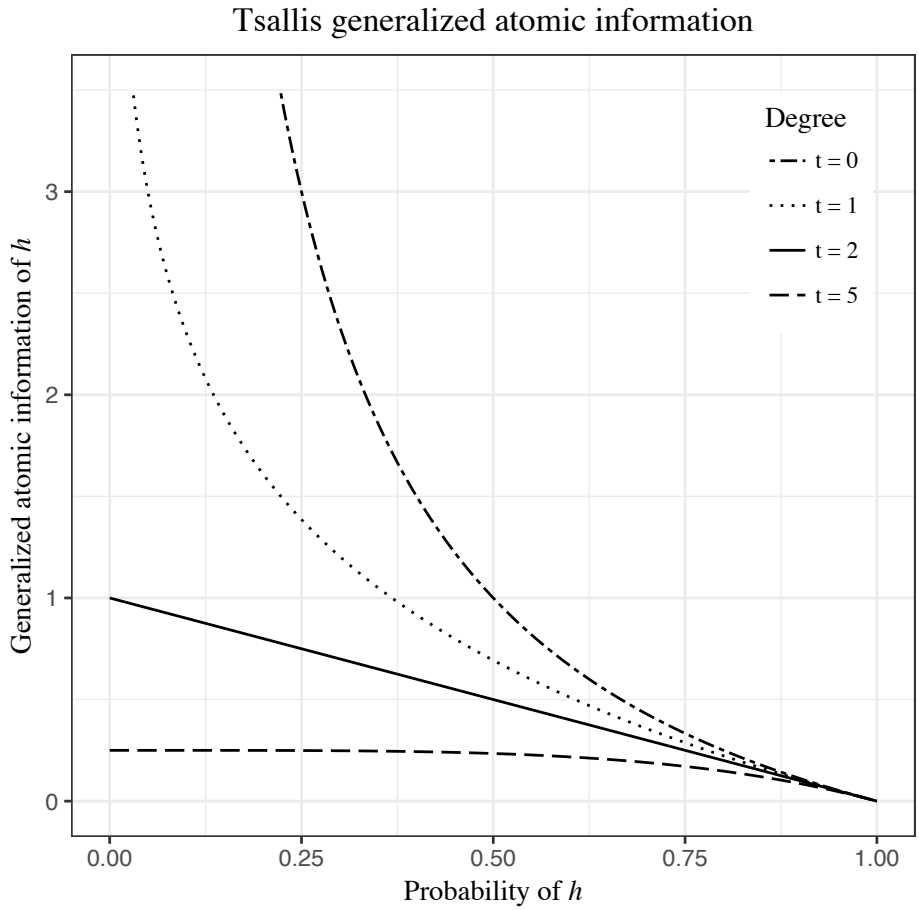

Figure 3. A graphical illustration of the generalized atomic information function $ln_t(1/P(h))$ for four different values of the parameter t (0, 1, 2, and 5, respectively, for the curves from top to bottom). Appropriately, the amount of information arising from finding out that h is the case is a decreasing function of $P(h)$. For high values of t, however, such decrease is flattened: with $t = 5$ (the lowest curve in the figure) finding out that h is true provides almost the same amount of information for a large set of initial probability assignments.

A good deal more can be said about the scope of this approach: see Figures 2 and 3, Table 4, and Suppl Mat (section 3) for additional material. Here, we will only mention briefly *three* important further points about *R*-measures in the Sharma-Mittal framework and their meaning for modelling information search behavior. They are as follows.

Additivity of expected entropy reduction: For any *H*,*E*,*F* and $P(H, E, F)$, $R_P^{SM_{(r,t)}}(H, E \times F) = R_P^{SM_{(r,t)}}(H, E) + R_P^{SM_{(r,t)}}(H, F|E)$.

This statement means that, for any Sharma-Mittal *R*-measure, the informational utility of a combined test $E \times F$ for H amounts to the sum of the plain utility of E and the utility of F that

is expected considering all possible outcomes of E [Suppl Mat, 4]. (Formally, $R_P^{SM_{(r,t)}}(H, F|E) = \sum_{e_j \in E} R_P^{SM_{(r,t)}}(H, F|e_j)P(e_j)$, while $R_P^{SM_{(r,t)}}(H, F|e_j)$ denotes the expected entropy reduction of H provided by F as computed when all relevant probabilities are conditionalized on e_j.) According to Nelson's (2008) discussion, this elegant additivity property of expected entropy reduction is important and highly desirable as concerns the analysis of the rational assessment of tests or queries. Moreover, one can see that the additivity of expected entropy reduction can be extended to any finite chain of queries and thus be applied to sequential search tasks such as those experimentally investigated by Nelson et al. (2014).

Irrelevance: For any H,E and $P(H, E)$, if either $E = \{e\}$ or $H \perp_P E$, then $R_P^{SM_{(r,t)}}(H, E) = 0$.

This statement says that two special kinds of queries can be known in advance to be of no use, that is, informationally inconsequential relative to the hypothesis set of interest. One is the case of an empty test $E = \{e\}$ with a single possibility that is already known to obtain with certainty, so that $P(e) = 1$. As suggested vividly by Floridi (2009, p. 26), this would be like consulting the raven in Edgar Allan Poe's famous poem, which is known to give one and the same answer no matter what (it always spells out "Nevermore"). The other case is when variables H and E are unrelated, that is, statistically independent according to P ($H \perp_P E$ in our notation). In both of these circumstances, $R_P^{SM_{(r,t)}}(H, E) = 0$ simply because the prior and posterior distribution on H are identical for each possible value of E, so that no entropy reduction can ever obtain.

By the irrelevance condition, empty and unrelated queries have zero expected utility — but can a query E have a *negative* expected utility? If so, a rational agent would be willing to pay a cost just for not being told what the true state of affairs is as concerns E, much as an abandoned lover who wants to be spared being told whether her/his beloved is or is not happy because s/he expects more harm than good. Note, however, that for the lover non-informational costs are clearly involved, while we are assuming queries or tests to be assessed in purely informational terms, bracketing all further factors (see, e.g., Raiffa & Schlaifer, 1961, Meder & Nelson, 2012, and Markant & Gureckis, 2012, for work involving situation-specific payoffs). In this perspective, it is reasonable and common to see irrelevance

as the worst-case scenario and exclude the possibility of informationally harmful tests: an irrelevant test (whether empty or statistically unrelated) simply can not tell us anything of interest, but that is as bad as it can get (see Good, 1967, and Goosens, 1976, for seminal analyses; also see Dawid, 1998).[6]

Interestingly, not all Sharma-Mittal measures of expected entropy reduction are non-negative. Some of them do allow for the controversial idea that there could exist detrimental tests in purely informational terms, such that an agent should rank them worse than an irrelevant search and take active measures to avoid them (despite them having, by assumption, no intrinsic cost). Mathematically, a non-negative measure $R_P(H, E)$ is generated if and only if the underlying entropy measure is a *concave* function [Suppl Mat, 4], and the conditions for concavity are as follows:

Concavity: $ent_P^{SM_{(r,t)}}(H)$ is a concave function of $\{P(h_1), \ldots, P(h_n)\}$ just in case $t \geq 2 - 1/r$.[7]

In terms of Figure 2, this means that any entropy (represented by a point) below the Arimoto curve is not generally concave (see Figure 4 for a graphical illustration of a strongly non-concave entropy measure). Thus, if the concavity of *ent* is required (to preserve the non-

[6] In theories of so-called imprecise probabilities, the notion arises of a detrimental experiment *E* in the sense that interval probability estimates for each element in a hypothesis set of interest *H* can be properly included in the corresponding interval probability estimates conditional on *each* element in *E*. This phenomenon is known as *dilation*: one's initial state of credence about *H* becomes less precise (thus *more uncertain*, under a plausible interpretation) no matter how an experiment turns out. The strongly unattractive character of this implication has been sometimes disregarded (see Tweeney et al., 2010, for an example in the psychology of reasoning), but the prevailing view is that appropriate moves are required to avoid it or dispel it (for recent discussions, see Bradley & Steele, 2014; Pedersen & Wheeler, 2014).

[7] This important result is proven in Hoffmann (2008), and already mentioned in Taneja et al. (1989, p. 61), who in turn refer to van der Pyl (1978) for a proof. We did not posit concavity as a defining property of entropies, and that's how it should be, in our opinion. Concavity may definitely be convenient or even required in some applications, but barring non-concave functions would be overly restrictive as concerns the formal notion of entropy. In physics, for instance, concavity is taken as directly relevant for generalized thermodynamics (Beck, 2009, p. 499; Tsallis, 2004, p. 10). In biological applications, on the other hand, concavity was suggested by Lewontin (1972; also see Rao 2010, p. 71), but seen as having "no intuitive motivation" by Patil and Taille (1982, p. 552).

negativity of R), then many prominent special cases are retained (including Quadratic, Hartley, Shannon, and Error entropy), but a significant bit of the whole Sharma-Mittal parameter space is ruled out. This concerns, for instance, entropies of degree 1 and order higher than 1 (see Ben-Bassat & Raviv, 1978).

Table 4. *A summary of the Sharma-Mittal framework and several of its special cases, including a specification of their structure in the general theory of means and a key reference for each.*

	(r,t)-setting	**Algebraic form of $ent_P(H)$**	**Generalized mean construction**		
			Characteristic function and its inverse		Atomic information
Sharma-Mittal Sharma & Mittal (1975)	$r \geq 0$ $t \geq 0$	$\frac{1}{t-1}\left[1-\left(\sum_{h_i \in H} P(h_i)^r\right)^{\frac{t-1}{r-1}}\right]$	$g(x) = ln_r(e_t^x)$	$g^{-1}(x) = ln_t(e_r^x)$	$inf(x) = ln_t\left(\frac{1}{x}\right)$
Effective Numbers Hil (1973)	$r \geq 0$ $t = 0$	$\left(\sum_{h_i \in H} P(h_i)^r\right)^{\frac{1}{1-r}} - 1$	$g(x) = ln_r(1+x)$	$g^{-1}(x) = e_r^x - 1$	$inf(x) = \frac{1-x}{x}$
Rényi Rényi (1961)	$r \geq 0$ $t = 1$	$\frac{1}{1-r} ln\left(\sum_{h_i \in H} P(h_i)^r\right)$	$g(x) = ln_r(e^x)$	$g^{-1}(x) = ln(e_r^x)$	$inf(x) = ln\left(\frac{1}{x}\right)$
Power entropies Laakso & Taagepera (1979)	$r \geq 0$ $t = 2$	$1-\left(\sum_{h_i \in H} P(h_i)^r\right)^{\frac{1}{r-1}}$	$g(x) = ln_r\left(\frac{1}{1-x}\right)$	$g^{-1}(x) = 1-(e_r^x)^{-1}$	$inf(x) = 1-x$
Gaussian Frank (2004)	$r = 1$ $t \geq 0$	$\frac{1}{t-1}\left[1-e^{(1-t)\left[\sum_{h_i \in H} P(h_i) ln\left(\frac{1}{P(h_i)}\right)\right]}\right]$	$g(x) = ln(e_t^x)$	$g^{-1}(x) = ln_t(e^x)$	$inf(x) = ln_t\left(\frac{1}{x}\right)$
Arimoto Arimoto (1971)	$r \geq ½$ $t = 2-\frac{1}{r}$	$\frac{r}{r-1}\left[1-\left(\sum_{h_i \in H} P(h_i)^r\right)^{\frac{1}{r}}\right]$	$g(x) = ln_r\left[1+\left(\frac{1-r}{r}\right)x\right]^{\frac{r}{1-r}}$	$g^{-1}(x) = \frac{r}{r-1}\left[1-(e_r^x)^{\frac{1-r}{r}}\right]$	$inf(x) = \frac{r}{r-1}\left[1-x^{\frac{r-1}{r}}\right]$
Tsallis Tsallis (1988)	$r = t \geq 0$	$\frac{1}{t-1}\left(1-\sum_{h_i \in H} P(h_i)^t\right)$	$g(x) = x$	$g^{-1}(x) = x$	$inf(x) = ln_t\left(\frac{1}{x}\right)$
Quadratic Gini (1912)	$r = t = 2$	$1-\sum_{h_i \in H} P(h_i)^2$	$g(x) = x$	$g^{-1}(x) = x$	$inf(x) = 1-x$
Shannon Shannon (1948)	$r = t = 1$	$\sum_{h_i \in H} P(h_i)\, ln\left(\frac{1}{P(h_i)}\right)$	$g(x) = x$	$g^{-1}(x) = x$	$inf(x) = ln\left(\frac{1}{x}\right)$
Hartley Hartley (1928)	$r = 0$ $t = 1$	$ln\left(\sum_{h_i \in H} P(h_i)^0\right)$	$g(x) = e^x - 1$	$g^{-1}(x) = ln(1+x)$	$inf(x) = ln\left(\frac{1}{x}\right)$

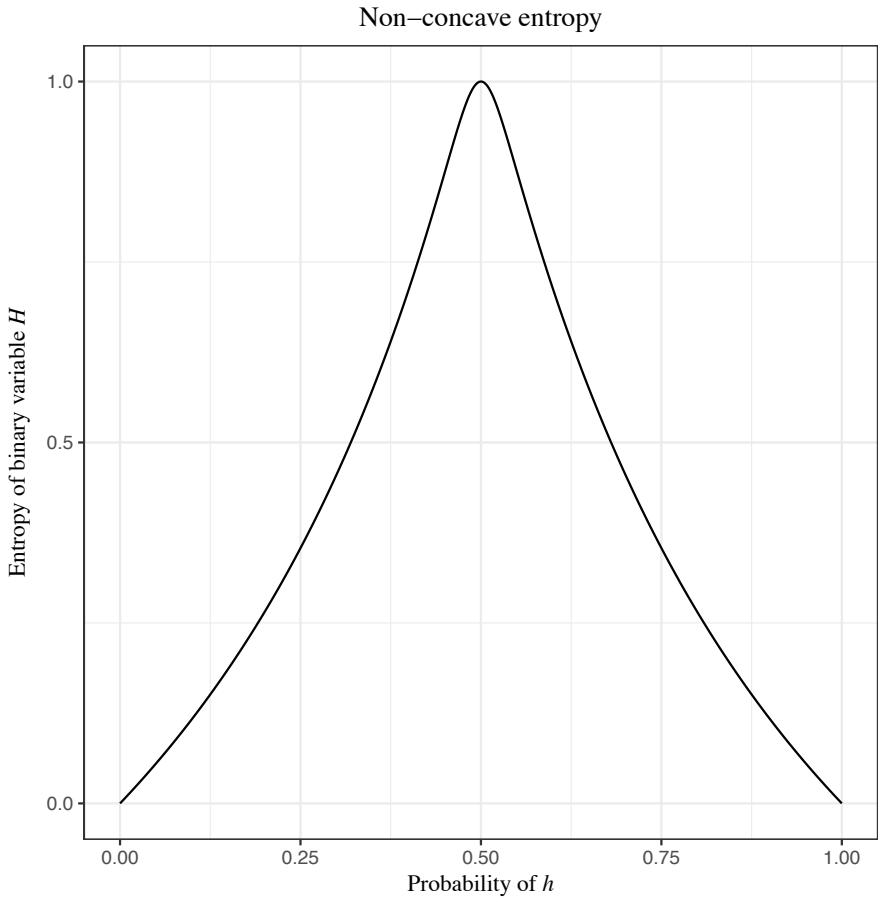

Figure 4. Graphical illustration of the non-concave entropy $ent^{SM_{(20,0)}}$ for a binary hypothesis set $H = \{h, \overline{h}\}$ as a function of the probability of *h*.

4.2. Psychological interpretation of the order and degree parameter

The order parameter r*: Imbalance and continuity*. What is the meaning of the order parameter in the Sharma-Mittal formalism when entropies and expected entropy reduction measures represent uncertainty and the value of queries, respectively? To clarify, let us consider what happens with extreme values of *r*, i.e., if *r* = 0 or goes to infinity, respectively [Suppl Mat, 3]:

$$ent_P^{SM_{(0,t)}}(H) = ln_t\left[\sum_{h_i \in H} P(h_i)^0\right]$$

$$ent_P^{SM_{(\infty,t)}}(H) = ln_t\left[\frac{1}{\max_{h_i \in H}[P(h_i)]}\right]$$

Given the convention $0^0 = 0$, $\sum_{h_i \in H} P(h_i)^0$ simply computes the number of all elements in *H* with a non-null probability. Accordingly, when *r* = 0, entropy becomes a (increasing) function of the mere number of the "live" (non-zero probability) options in *H*. When *r* goes to infinity,

on the other hand, entropy becomes a (decreasing) function of the probability of a single element in *H*, i.e., the most likely hypothesis. This shows that the order parameter *r* is an index of the *imbalance* of the entropy function, which indicates how much the entropy measure discounts minor (low probability) hypotheses. For order-0 measures, the actual probability distribution is neglected: non-zero probability hypothesis are just counted, as if they were all equally important (see Gauvrit & Morsanyi, 2014). For order-∞ measures, on the other hand, only the most probable hypothesis matters, and all other hypotheses are disregarded altogether. For intermediate values of *r*, more likely hypotheses count more, but less likely hypotheses do retain some weight. The higher [lower] *r* is, the more [less] the likely hypotheses are regarded and the unlikely hypotheses are discounted. Importantly, for extreme values of the order parameter, an otherwise natural idea of *continuity* fails in the measurement of entropy: when *r* goes to either zero or infinity, it is not the case that small (large) changes in the probability distribution *P*(*H*) produce comparably small (large) changes in entropy values.

To see better how order-0 entropy measures behave, consider the simplest of them:

$$ent_P^{SM_{(0,0)}}(H) = n^+ - 1$$

where $n^+ = \sum_{h_i \in H} P(h_i)^0$, so n^+ denotes the number of hypotheses in *H* with a non-null (strictly positive) probability. Given the –1 correction, $ent^{SM_{(0,0)}}$ can be interpreted as the "number of contenders" for each entity in set *H*, because it takes value 0 when only one element is left. For future reference, we will label $ent^{SM_{(0,0)}}$ *Origin entropy* because it marks the origin of the graph in Figure 3. Importantly, the expected reduction of Origin entropy is just *the expected number of hypotheses in H conclusively falsified by a test E*.

To the extent that all details of the prior and posterior probability distribution over *H* are neglected, computational demands are significantly decreased with order-0 entropies. As a consequence, measures of the expected reduction of an order-0 entropy (and especially Origin entropy) also amount to comparably frugal, heuristic or quasi-heuristic models of information search (see Baron et al.'s model, 1988, p. 106). Lack of continuity, too, is associated with heuristic models, which often rely on discrete elements instead of continuous representations (see Gigerenzer, Hertwig, & Pachur, 2011; Katsikopoulos, Schooler, & Hertwig, 2010). More

generally, when the order parameter approaches 0, entropy measures become more and more balanced, meaning that they treat all live hypotheses more and more equally. What happens to the associated expected entropy reduction measures is that they become more and more "Popperian" in spirit. In fact, for order-0 relevance measures, a test *E* will deliver some non-null expected informational utility about hypothesis set *H* if and only if some of the possible outcomes of *E* can conclusively rule out some element in *H*. Otherwise, the expected entropy reduction will be zero, no matter how large the changes in probability that might arise from *E*. Cognitively, relevance measures of low order would then describe the information search preferences of an agent who is distinctively eager to prune down the list of candidate hypotheses, an attitude which might prevail in earlier stages of an inquiry, when such a list can be sizable.

Among entropy measures of order infinity, we already know $ent^{SM_{(\infty,2)}} = 1 - \max_{h_i \in H}[P(h_i)]$ as Error entropy. What this illustrates is that, when *r* goes to infinity, entropy measures become more and more decision-theoretic in a short-sighted kind of way: in the limit, they are only affected by the probability of a correct guess given the currently available information. A notable consequence for the associated measures of expected entropy reduction is that a test *E* can deliver some non-null expected informational utility only if some of the possible outcomes of *E* can alter the probability of the modal hypothesis in *H*. If that is not the case, then the expected utility will be zero, no matter how significant the changes in the probability distribution arising from *E*. Cognitively, then, *R*-measures of very high order would describe the information search preferences of an agent who is predominantly concerned with an estimate of the probability of error in an impending choice from set *H*.

The degree parameter t*: Perfect tests and certainty*. Let us now consider briefly the meaning of the degree parameter *t* in the Sharma-Mittal formalism when entropies and relevance measures represent uncertainty and the value of queries, respectively. A remarkable fact about the degree parameter *t* is that (unlike the order parameter *r*) it does *not* affect the ranking of entropy values. Indeed, one can show that any Sharma-Mittal entropy measure is a strictly increasing function of any other measure *of the same order* r, *regardless of the degree* (for any hypothesis set *H* and any probability distribution *P*) [Suppl Mat, 4]. Thus, concerning

the ordinal comparison of entropy values, only if the *order* differs can divergences between pairs of SM entropy measures arise. On the other hand, the implications of the degree parameter for measures of expected entropy reduction are significant and have not received much attention.

As a useful basis for discussion, suppose that variables H and E are independent, in the standard sense that for any $h_i \in H$ and any $e_j \in E$, $P(h_i \cap e_j) = P(h_i)P(e_j)$, denoted as $H \perp_P E$. Then we have [Suppl Mat, 4]:

$$R_P^{SM_{(r,t)}}(E,E) - R_P^{SM_{(r,t)}}(H \times E,E) = (t-1)ent_P^{SM_{(r,t)}}(H)ent_P^{SM_{(r,t)}}(E)$$

If expected entropy reduction is interpreted as a measure of the informational utility of queries or tests, this equality governs the relationship between the computed utilities of E in case it is a "perfect" (conclusive) test and in case it is not. More precisely, the first term on the left, $R_P^{SM_{(r,t)}}(E,E)$, measures the expected informational utility of a perfect test because the test itself and the target of investigation are the same, hence finding out the true value of E removes all relevant uncertainty. On the other hand, E is not anymore a perfect test in the second term of the equation above, $R_P^{SM_{(r,t)}}(H \times E,E)$, for here a more fine-grained hypothesis set $H \times E$ is at issue, thus a more demanding epistemic target; hence finding out the true value of E would not remove all relevant uncertainty. (Recall that, by assumption, H is statistically independent from E, so the uncertainty about H would remain untouched, as it were, after knowing about E.). With entropies of degree 1 (including Shannon), the associated measures of expected entropy reduction imply that E has *exactly identical* utility in both cases, because $t = 1$ nullifies the right-hand side of the equation, regardless of the order parameter r. With $t > 1$ the right-hand side is positive, so E is a strictly more useful test when it is conclusive than when it is not. With $t < 1$, on the contrary, the right-hand side is negative, so E is strictly *less* useful a test when it is conclusive than when it is not. Note that these are ordinal relationships (rankings). In comparing the expected informational utility of queries, the degree parameter t can thus play a crucial role. Crupi and Tentori (2014, p. 88) provided some simple illustrations which can be adapted as favoring an entropy with $t > 1$ as the basis for the R-measure of the expected utility of queries (here, we present an illustration in Figure 5).

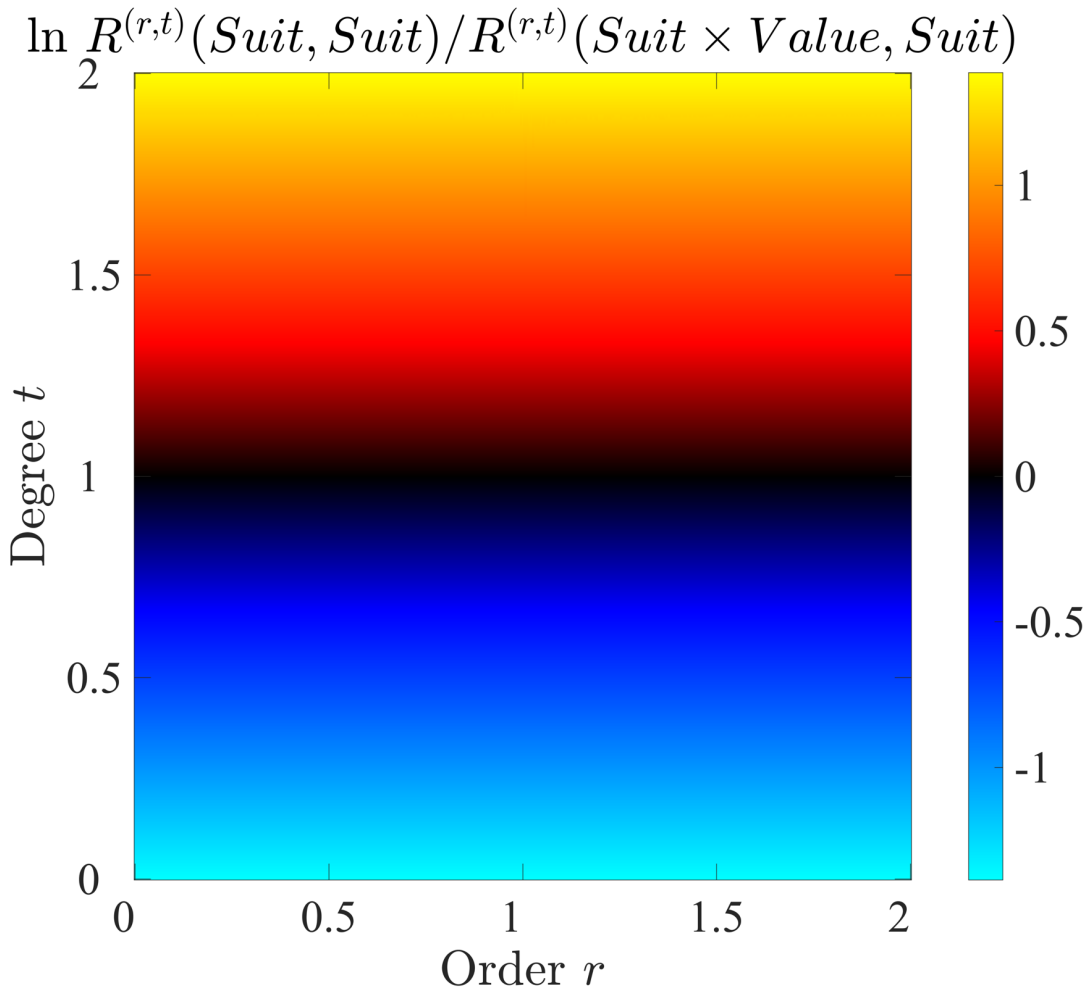

Figure 5. Consider a standard 52-card playing deck, with *Suit* corresponding to the 4 equally probable suits, *Value* corresponding to the 13 equally probable numbers (or faces) that a card can take (2 through 10, Jack, Queen, King, Ace), and *Suit*×*Value* corresponding to the 52 equally probable individual cards in the deck. Suppose that you will be told the suit of a randomly chosen card. Is this more valuable to you if (i) (perfect test case) your goal is to learn the suit, i.e., R_P(*Suit*, *Suit*), or (ii) (inconclusive test case) your goal is to learn the specific card, i.e., R_P(*Suit*×*Value*, *Suit*)? What is the ratio of the value of the expected entropy reduction in (i) vs. (ii)? For degree 1, the information to be obtained has equal value in each case. For degrees greater than 1, the perfect test is more useful. For degrees less than one, the inconclusive test is more useful. Interestingly, as the figure shows, the degree parameter uniquely determines the relative value of R_P(*Suit*, *Suit*) and R_P(*Suit*×*Value*, *Suit*), regardless of the order parameter. In the Figure, values of the order parameter *r* and of the degree parameter *t* lie on the *x*– and *y*–axis, respectively. Color represents the log of the ratio between the conclusive test and the inconclusive test case in the card example above: black means that the information values of the tests are equal (log of the ratio is 0); warm / cool shades indicate that the conclusive test has a higher / lower value, respectively (log of the ratio is positive / negative).

The meaning of a high degree parameter is of particular interest in so-called Tsallis family of entropy measures, obtained from $ent^{SM_{(r,t)}}$ when $r = t$ (see Table 4). Consider Tsallis entropy of degree 30, that is $ent^{SM_{(30,30)}}$. With this measure, entropy remains very close to a upper bound value of $1/(t-1) \approx 0.0345$ unless the probability distribution reflects near-certainty about the true element in the hypothesis set *H*. For instance, for as uneven a distribution as {0.90, 0.05, 0.05}, $ent^{Tsallis_{(30)}}$ yields entropy 0.03330, still close to 0.0345, while it quickly approaches 0 when the probability of one hypothesis exceeds 0.99. *Non-Certainty entropy* seems a useful label for future reference, as measure $ent^{Tsallis_{(30)}}$ essentially implies that entropy is almost invariant as long as an appreciable lack of certainty (a "reasonable doubt", as it were) endures. Accordingly, the entropy reduction from a piece of evidence *e* is largely negligible unless one is led to acquire a very high degree of certainty about *H*, and it approaches the upper bound of $1/(t-1)$ as the posterior probability comes close to matching a truth-value assignment (with $P(h_i) = 1$ for some *i* and 0 for all other *h*s). Up to the inconsequential normalizing constant $t - 1$, the expected reduction of this entropy, $R^{Tsallis_{(30)}}$, amounts to a smooth variant of Nelson's et al. (2010) "probability-of-certainty heuristic", where a datum $e_i \in E$ has informational utility 1 if it reveals the true element in *H* with certainty and utility 0 otherwise, so that the expected utility of *E* itself is just the overall probability that certainty about *H* is eventually achieved by that test. These remarks further illustrate that a larger degree *t* implies an increasing tendency of the corresponding *R*-measure to value highly the attainment of certainty or quasi-certainty about the target hypothesis set when assessing a test.

5. A systematic exploration of how key information search models diverge

Depending on different entropy functions, two measures *R* and R^* of the expected reduction of entropy as the informational utility of tests may disagree in their rankings. Formally, there exist variables *H*, *E*, and *F* and probability distribution *P*(*H*,*E*,*F*) such that $R_P(H,E) > R_P(H,F)$ while $R_P^*(H,E) < R_P^*(H,F)$; thus, *R*-measures are *not* generally *ordinally equivalent*. In the following, we will focus on an illustrative sample of measures in the Sharma-Mittal framework and show that such divergences can be widespread, strong, and telling about the specific

tenets of those measures. This means that different entropy measures can provide markedly divergent implications in the assessment of possible queries' expected usefulness. Depending on the interpretation of the models, this in turn implies conflicting empirical predictions and/or incompatible normative recommendations.

Our list will include three classical models that are standard at least in some domains, namely Shannon, Quadratic, and Error entropy. It also includes three measures which we previously labelled heuristic or quasi-heuristic in that they largely or completely disregard quantitative information conveyed by the relevant probability distribution *P*: these are Origin entropy (or the "number of contenders"), Hartley entropy, and Non-Certainty entropy, as defined above. For a wider coverage and comparison, we also include an entropy function lying well below the Arimoto curve in Figure 2, that is, $ent^{SM_{(20,0)}}$, and thus labelled *Non-Concave* (see Figure 4).

We ran simulations to identify cases of strong disagreement between our seven measures of expected entropy reduction, on a pairwise basis, about which of two tests is taken to be more useful. In each simulation, we considered a scenario with a threefold hypothesis space $H = \{h_1, h_2, h_3\}$, and two binary tests, $E = \{e,\overline{e}\}$ and $F = \{f, \overline{f}\}$.[8] The goal of each simulation was to find a case — that is, a specific joint probability distribution $P(H,E,F)$ — where two *R*-measures strongly disagree about which of two tests is most useful. The ideal scenario here is a case where expected reduction of one kind of entropy (say, Origin) implies that *E* is as useful as can possibly be found, while *F* is as bad as it can be, and the expected reduction of another kind of entropy (say, Shannon) implies the opposite, with equal strength of conviction.

The quantification of the disagreement between two *R*-measures in a given case — for a given $P(H,E,F)$ — arises from three steps (also see Nelson et al., 2010). (i) *Normalization*: for each measure, we divide nominal values of expected entropy reduction (for each of *E* and *F*) by the expected entropy reduction of a *conclusive* test for three equally probable hypotheses,

[8] We used three-hypothesis scenarios to illustrate the differences among our selected sample of *R* measures, because scenarios of this kind appeared to offer a reasonable balance of being simple yet powerful enough to deliver divergences that are strong and intuitively clear. Note however that two-hypothesis scenarios can also clearly differentiate many of the *R* measures (see the review of behavioral research on binary classification tasks in the subsequent sections).

that is, by $R_U(H, H)$. (ii) *Preference Strength*: for each measure, we compute the simple difference between the (normalized) expected entropy reduction for test *E* and for test *F*, that is, $\frac{R_P(H,E)}{R_U(H,H)} - \frac{R_P(H,F)}{R_U(H,H)}$. (iii) *Disagreement Strength* (*DS*): if the two measures agree on whether *E* or *F* is most useful, *DS* is defined as zero; if they disagree, *DS* is defined as the geometric mean of those measures' respective absolute preference strengths in step (ii).

In the simulations, a variety of techniques were involved in order to maximize disagreement strength, including random generation of prior probabilities over *H* and of likelihoods for *E* and *F*, optimization of likelihoods alone, and joint optimization of likelihoods and priors. Each example reported here was found in the attempt to maximize *DS* for a particular pair of measures. We relied on the simulations largely as a heuristic tool, thus selecting and slightly adapting the numerical examples to make them more intuitive and improve clarity.[9]

For each pair of *R*-measures in our sample of seven, at least one case of moderate or strong disagreement was found (Table 5). Thus, for each pairwise comparison one can identify probabilities for which the models make diverging claims about which test is more useful. In what follows, we append a short discussion to the cases in which Shannon entropy strongly disagrees with each competing model. Such discussion is illustrative and qualitative, to intuitively highlight the underlying properties of different models. Similar explications could be provided for all other pairwise comparisons, but are omitted for the sake of brevity.

Shannon vs. *Non-Certainty Entropy* (case 3 in Table 5; *DS* = 0.30). In its purest form, Non-Certainty entropy equals 0 if one hypothesis in *H* is known to be true with certainty, and 1 otherwise. As a consequence, the entropy reduction expected from a test *E* just amounts to the probability that full certainty will be achieved after *E* is performed. Within the Sharma-Mittal framework, this behavior can be often approximated by an entropy measure such as Tsallis of degree 30, as explained above.[10] One example where the expected reduction of Shannon and

[9] It is important to note that the procedures we used do not guarantee finding globally maximal solutions; thus, a failure to find a case of strong disagreement does not necessarily entail that no such case exists.

[10] One should note, however, that Tsallis 30, unlike pure non-certainty entropy, is a continuous function. As a consequence, the approximation described eventually fails when one gets very close to limiting cases. More

Non-Certainty entropy disagree significantly involves a prior $P(H) = \{0.67, 0.10, 0.23\}$. The Non-Certainty measure rates very poorly a test E such that $P(H|e) = \{0.899, 0.100, 0.001\}$, $P(H|\overline{e}) = \{0.001, 0.100, 0.899\}$, and $P(e) = 0.74$, and strongly prefers a test F such that $P(H|f) = \{1, 0, 0\}$, $P(H|\overline{f}) = \{0.40, 0.18, 0.42\}$, and $P(f) = 0.45$, because the probability to attain full certainty from F is sizable (45%). The expected reduction of Shannon entropy implies the opposite ranking, because test E, while unable to provide full certainty, will invariably yield a highly skewed posterior as compared to the prior.

Shannon vs. *Origin and Hartley Entropy* (case 5 in Table 5; $DS = 0.56$ and $DS = 0.48$, respectively). The reduction of both Origin and Hartley entropy share similar ideas of counting how many hypotheses are conclusively ruled out by the evidence. For example, with prior $P(H) = \{0.500, 0.499, 0.001\}$, the expected reduction of either Origin or Hartley entropy assigns value zero to test E such that $P(H|e) = \{0.998, 0.001, 0.001\}$, $P(H|\overline{e}) = \{0.001, 0.998, 0.001\}$, and $P(e) = 0.501$, because no hypothesis is ever ruled out conclusively, and rather prefers test F such that $P(H|f) = \{0.501, 0.499, 0\}$, $P(H|\overline{f}) = \{0, 0.499, 0.501\}$, and $P(f) = 0.998$. The expected reduction of Shannon entropy implies the opposite ranking, because F will almost always yield only a tiny change in overall uncertainty.

Shannon vs. *Non-Concave Entropy* (case 6 in Table 5; $DS = 0.26$). For non-concave entropies, the expected entropy reduction may turn out to be negative, thus indicating an allegedly detrimental query, that is, a test where expected utility is lower than that of a completely irrelevant test. This feature yields cases of significant disagreement between the expected reduction of our illustrative Non-Concave entropy, $ent^{SM_{(20,0)}}$, and of classical concave measures such as Shannon. With a prior $P(H) = \{0.66, 0.17, 0.17\}$, the Non-Concave measure rates a test E such that $P(H|e) = \{1, 0, 0\}$, $P(H|\overline{e}) = \{1/3, 1/3, 1/3\}$, and $P(e) = 0.49$ much lower than an irrelevant test F such that $P(H|f) = P(H|\overline{f}) = P(H)$. Indeed, the non-concave R-measure assigns a significant *negative* value to test E. This critically depends on one interesting fact: for

precisely, Tsallis 30 entropy rapidly decreases for *almost* certain distributions such as, say, $P(H) = \{0.998, 0.001, 0.001\}$. In fact, Tsallis 30 entropy is sizable and almost constant if $P(H)$ conveys a *less-than-almost-certain* state of belief, and becomes largely negligible otherwise.

Non-Concave entropy, going from $P(H)$ to a completely flat posterior, $P(H|\overline{e})$, is an extremely aversive outcome (i.e. it implies a very large increase in uncertainty), while the 49% chance of achieving certainty by datum *e* is not highly valued (a feature of low degree measures, as we know). The expected reduction of Shannon entropy implies the opposite ranking instead, as it conveys the principle that no test can be informationally less useful than an irrelevant test (such as *F*).

Shannon vs. *Quadratic Entropy* (case 8 in Table 5; *DS* = 0.09). Shannon and Quadratic entropies are similar in many ways, yet at least cases of moderate disagreement can be found. One is with prior $P(H) = \{0.50, 0.14, 0.36\}$. Test *E* is such that $P(H|e) = \{0.72, 0.14, 0.14\}$, $P(H|\overline{e}) = \{0.14, 0.14, 0.72\}$, and $P(e) = 0.62$, while with test *F* one has $P(H|f) = \{0.5, 0.5, 0\}$, $P(H|\overline{f}) = \{0.5, 0, 0.5\}$, and $P(f) = 0.28$. Expected Quadratic entropy reduction ranks *E* over *F*, as it puts a particularly high value on posterior distributions where one single hypothesis comes to prevail. In comparison, this is less important for the reduction of Shannon entropy, as long as some hypotheses are completely (or largely) ruled out, as occurs with *F*. Accordingly, the Shannon measure prefers *F* over *E*.

Shannon vs. *Error Entropy* (case 9 in Table 5; *DS* = 0.20). A stronger disagreement arises between Shannon and Error entropy. Consider prior $P(H) = \{0.50, 0.18, 0.32\}$, a test *E* such that $P(H|e) = \{0.65, 0.18, 0.17\}$, $P(H|\overline{e}) = \{0.17, 0.18, 0.65\}$, and $P(e) = 0.69$, and a test *F* such that $P(H|f) = \{0.5, 0.5, 0\}$, $P(H|\overline{f}) = \{0.5, 0, 0.5\}$, and $P(f) = 0.36$. The expected reduction of Error entropy is significant with *E* but zero with *F*, because the latter will leave the modal probability untouched. (Note that it does not matter that the hypotheses with the maximum probability changed.) However, test *F*, unlike *E*, will invariably rule out an hypothesis that was a priori significantly probable, and for this reason is preferred by the Shannon *R*-measure.

Table 5. *Cases of strong disagreement between seven measures of expected entropy reduction. Two binary tests E and F are considered for a ternary hypothesis set H. Preference strength is the difference between (normalized) values of expected entropy reduction for E and F, respectively: it is positive if test E is strictly preferred, negative if F is strictly preferred, and null if they're rated equally. The most relevant preference values to be compared are highlighted in bold: they illustrate that, for each pair of R-measures in our sample of seven, the table includes at least one case of moderate or strong disagreement.*

n.	$P(H)$	Test E		Test F		Preference strength in the expected reduction of entropy						
		$P(H\|e)$ *vs.* $P(H\|\bar{e})$	$P(e)$ *vs.* $P(\bar{e})$	$P(H\|f)$ *vs.* $P(H\|\bar{f})$	$P(f)$ *vs.* $P(\bar{f})$	Non-Certainty	Origin	Hartley	Non-Concave	Shannon	Quadratic	Error
1	{0.50, 0.25, 0.25}	{0.5, 0.5, 0} {0.5, 0, 0.5}	0.5 0.5	{1, 0, 0} {1/3, 1/3, 1/3}	0.25 0.75	**−0.250**	**0.250**	**0.119**	0.250	0.119	0	0
2	{0.67, 0.17, 0.17}	{0.82, 0.17, 0.01} {0.01, 0.17, 0.82}	0.8 0.2	{1, 0, 0} {1/3, 1/3, 1/3}	0.49 0.51	**−0.487**	−0.490	−0.490	**0.394**	0.046	0.062	0.240
3	{0.67, 0.10, 0.23}	{0.899, 0.1, 0.001} {0.001, 0.1, 0.899}	0.74 0.26	{1, 0, 0} {0.40, 0.18, 0.42}	0.45 0.55	**−0.409**	−0.450	−0.450	0.342	**0.218**	**0.249**	**0.329**
4	{0.6, 0.1, 0.3}	{1, 0, 0} {1/3, 1/6, 1/2}	0.4 0.6	{0.7, 0.3, 0} {0.55, 0. 0.45}	1/3 2/3	0.400	**−0.100**	**0.031**	0.045	0.051	0.155	0.150
5	{0.5, 0.499, 0.001}	{0.998, 0.001, 0.001} {0.001, 0.998, 0.001}	0.501 0.499	{0.501, 0.499, 0} {0, 0.499, 0.501}	0.998 0.002	0.942	**−0.500**	**−0.369**	**0.499**	**0.617**	**0.744**	**0.746**
6	{0.66, 0.17, 0.17}	{1, 0, 0} {1/3, 1/3, 1/3}	0.49 0.51	{0.66, 0.17, 0.17} {0.66, 0.17, 0.17}	0.5 0.5	0.490	0.490	0.490	**−0.236**	**0.288**	**0.250**	0
7	{0.53, 0.25, 0.22}	{1, 0, 0} {0.295, 0.375, 0.330}	1/3 2/3	{0.53, 0.25, 0.22} {0.53, 0.25, 0.22}	0.5 0.5	0.333	0.333	0.333	**−0.123**	0.261	0.249	**0.080**
8	{0.50, 0.14, 0.36}	{0.72, 0.14, 0.14} {0.14, 0.14, 0.72}	0.62 0.38	{0.5, 0.5, 0} {0.5, 0, 0.5}	0.28 0.72	0	−0.500	−0.369	0.293	**−0.085**	**0.086**	0.330
9	{0.50, 0.18, 0.32}	{0.65, 0.18, 0.17} {0.17, 0.18, 0.65}	0.69 0.31	{0.5, 0.5, 0} {0.5, 0, 0.5}	0.36 0.64	0	−0.180	−0.133	0.213	**−0.179**	−0.024	**0.225**
10	{0.42, 0.42, 0.16}	{0.5, 0.5, 0} {0, 0, 1}	0.84 0.16	{0.66, 0.24, 0.10} {0.10, 0.66, 0.24}	0.57 0.43	0.160	0.580	0.470	−0.146	0.241	**0.115**	**−0.120**

6. Model comparison: Prediction and behavior

Now that we have seen examples illustrating the theoretical properties of a variety of Sharma-Mittal relevance measures, we turn to addressing whether the Sharma-Mittal measures can help with psychological or normative theory of the value of information.

6.1. Comprehensive analysis of Wason's abstract selection task

The single most widely studied experimental information search paradigm is Wason's (1966) selection task. In the classical, abstract version, participants are presented with a conditional hypothesis (or "rule"), h = "if A [antecedent], then C [consequent]". The hypothesis concerns some cards, each of which has a letter on one side and a number on the other, for instance A = "the card has a vowel on one side" and C = "the card has an even number on the other side". One side is displayed for each of four cards: one instantiating A (e.g., showing letter E), one instantiating not-A (e.g., showing letter K), one instantiating C (e.g., showing number 4), and one instantiating not-C (e.g., showing number 7). Participants have therefore four information search options in order to assess the truth or falsity of hypothesis h: turning over the A, the not-A, the C, or the not-C card. They are asked to choose which ones they would pick up as useful to establish whether the hypothesis holds or not. All, none, or any subset of the four cards can be selected.

According to Wason's (1966) original, "Popperian" reading of the task, the A and not-C search options are useful because they could falsify h (by possibly revealing a even number and a vowel, respectively), so a rational agent should select them. The not-A and C options, on the contrary, could not provide conclusively refuting evidence, so they're worthless in this interpretation. However, observed choice frequencies depart markedly from these prescriptions. In Oaksford and Chater's (1994, p. 613) metaanalysis, they were 89%, 16%, 62%, and 25% for A, not-A, C, and not-C, respectively. Oaksford and Chater (1994, 2003) devised Bayesian models of the task in which agents treat the four cards as sampled from a larger deck and are assumed to maximize the expected reduction of uncertainty, with Shannon entropy as the standard measure. Oaksford and Chater postulated a foil hypothesis $\overline{h}$ in which A and C are statistically independent and a target hypothesis h under

which *C* always (or almost always) follows *A*. In Oaksford and Chater's (1994) "deterministic" analysis, *C* always followed *A* under the dependence hypothesis *h*. A key innovation in Oaksford and Chater (2003, p. 291) was the introduction of an "exception" parameter, such that $P(C|A) = 1 - P(\text{exception})$ under *h*. The model also requires parameters α and γ for the probabilities $P(A)$ and $P(C)$ of the antecedent and consequent of *h*. We implement Oaksford and Chater's (2003) model, positing $\alpha = 0.22$ and $\gamma = 0.27$ (according to the "rarity" assumption), and an uniform prior on $H = \{h, \overline{h}\}$, as suggested in Oaksford and Chater (2003, p. 296). We explored the implications of calculating the expected usefulness of turning over each card, not only according to Shannon entropy reduction, but for the whole set of entropy measures from the Sharma-Mittal framework. [11]

Empirical Data. We first address how well different expected entropy reduction measures correspond to empirical aggregate card selection frequencies in the task, with respect to Oaksford and Chater's (2003) model. For the selection frequencies, we use the abstract selection task data as reported by Oaksford and Chater (1994, p. 613) and mentioned above (89%, 16%, 62%, and 25% for *A*, not-*A*, *C*, and not-*C*, respectively).

Figure 6 (top row) shows the rank correlation between relevance values and empirical selection frequencies for each order and degree value from 0 to 20, in steps of 0.25. First consider results for the model with $P(\text{exception}) = 0$ (Figure 6, top left subplot). A wide range of measures, including expected reduction of Shannon and Quadratic entropy, of some non-concave entropies (e.g., $R^{SM_{(10,1.5)}}$) and of measures with fairly high degree (e.g., $R^{SM_{(10,8)}}$) correlate perfectly with the rank of selection frequencies. However, if a high degree measure with moderate or high order is used, the rank correlation is not perfect.

[11] To fit the relevant patterns of responses, we pursued a variety of methods, including optimizing Hattori's "selection tendency function" (which maps expected entropy reduction onto the predicted probability that a card will be selected, see Hattori, 1999, 2002; also see Stringer, Borsboom, & Wagenmakers, 2011), or taking previously reported parameters for Hattori's selection tendency function; Spearman rank correlation coefficients; and Pearson correlations. Similar results were obtained across these methods. Because the rank correlations are simple to discuss, we focus on those here. Full simulation results for these and other measures, model variants with other values of *P*(exception), and Matlab code, are available from J.D.N.

Consider for instance the Tsallis measure of degree 20 (i.e. $R^{SM_{(20,20)}}$). This leads to relevance values for the *A*, not-*A*, *C*, and not-*C* cards of 0.0281, 0.0002, 0.0008, and 0.0084, respectively. Because the relative ordering of the *C* and the not-*C* card is incorrect (from the perspective of observed choices), the rank correlation is only 0.8. The same rank correlation of 0.8 is obtained, but for a different reason, from strongly non-concave relevance measures. $R^{SM_{(20,0)}}$, for instance, gives values of 1.181, 0.380, 1.054, and 0.372 (again for the *A*, not-*A*, *C*, and not-*C* cards, respectively), so that the not-*A* card is deemed more informative than the not-*C* card by this relevance measure.

Let us now consider expected reduction of Origin entropy, $R^{SM_{(0,0)}}$, as an example of the 0-order measures. It gives relevance values of 0.527, 0, 0, and 0.159 for the *A*, not-*A*, *C*, and not-*C* cards, respectively. This is similar to Wason's analysis of the task: only the *A* and the not-*C* cards can falsify a hypothesis (namely, the dependence hypothesis *h*), thus only those two cards have value. The other cards could change the relative plausibility of *h* vs. $\overline{h}$; however, according to 0-order measures, no informational value is achieved because no hypothesis is definitely ruled out. In this sense, 0-order measures can be thought of as bringing elements of the original logical interpretation of the selection task into the same unified information-theoretic framework including Shannon and generalized entropies (see below for more on this). Interestingly, this does not imply that the *A* and the not-*C* cards are equally valuable: in the model, the *A* card offers a higher chance of falsifying *h* than the not-*C* card, so it is more valuable, according to this analysis. Thus, while incorporating the basic idea of the importance of possible falsification, the 0-order Sharma-Mittal formalization of informational value offers something that the standard logical reading does not: a rationale for assessing the relative value among those queries (the *A* and the not-*C* card) providing the possibility of falsifying a hypothesis. The Origin entropy values and the empirical data agree that the *A* card is most useful and (up to a tie) that the not-*A* card is least useful, but disagree on virtually everything else; $R^{SM_{(0,0)}}$'s rank correlation to empirical card selection frequencies is 0.6325.

What if Oaksford and Chater's (2003) model is combined with exception parameter *P*(exception) = 0.1, rather than 0? In this case, the empirical selection frequencies perfectly correlate with the theoretical values for an even wider range of measures than for the "deterministic" model (Figure 6, top right plot). For instance, Tsallis of degree 11, i.e. $R^{SM_{(11,11)}}$,

which had rank correlation of 0.8 with P(exception) = 0, has a perfect rank correlation with 0.1. This is due to the relative ordering of the not-A and C cards. For the P(exception) = 0 model, the A, not-A, C, and not-C cards had $R^{SM_{(11,11)}}$ relevance of 0.059, 0.002, 0.012, and 0.016, respectively; with P(exception) = 0.1, the cards' respective relevance values are 0.019, 0.001, 0.007, and 0.005. In addition, a dramatic difference between P(exception) = 0 and P(exception) = 0.1 arises for the 0-order measures. If P(exception) > 0, even if very small, no amount of obtained data can ever lead to ruling out a hypothesis in the model. Therefore, with P(exception) = 0.1 all cards have zero value for 0-order measures, and the correlation with behavioral data is undefined (plotted black in Figure 6).

A probabilistic understanding of Wason's normative indications. Finally, we discuss how well the expected informational value of the cards, as calculated using Oaksford and Chater's (2003) model and various Sharma-Mittal measures, corresponds to Wason's original interpretation of the task. We thus conducted the same analyses as above, but instead of using the human selection frequencies we assumed that the A card was selected with 100%, the not-A card with 0%, the C card with 0%, and the not-C card with 100% probability. The 0-order relevance measures, again within Oaksford and Chater's (2003) model with P(exception) = 0, provide a probabilistic understanding of Wason's normative indications. Like Wason, the 0-order measures deem only the A and the not-C cards to be useful when P(exception) = 0. The rank correlation with theoretical selection frequencies from Wason's analysis is 0.94 (see Figure 6, bottom left plot). Why is the correlation not perfect? The probabilistic understanding proposed, as discussed above, goes beyond the logical analysis: because the A card offers a higher probability of falsification than the not-C card does in the probability model, the 0-order relevance measures value the former more than the latter. Recall that our hypothetical participants always select both cards that entail the possibility of falsifying the dependence hypothesis; thus, the correlation is less than one. The worst correlation with Wason's ranking is from the strongly non-concave measures, such as $R^{SM_{(20,0)}}$; this correlation is exactly zero.

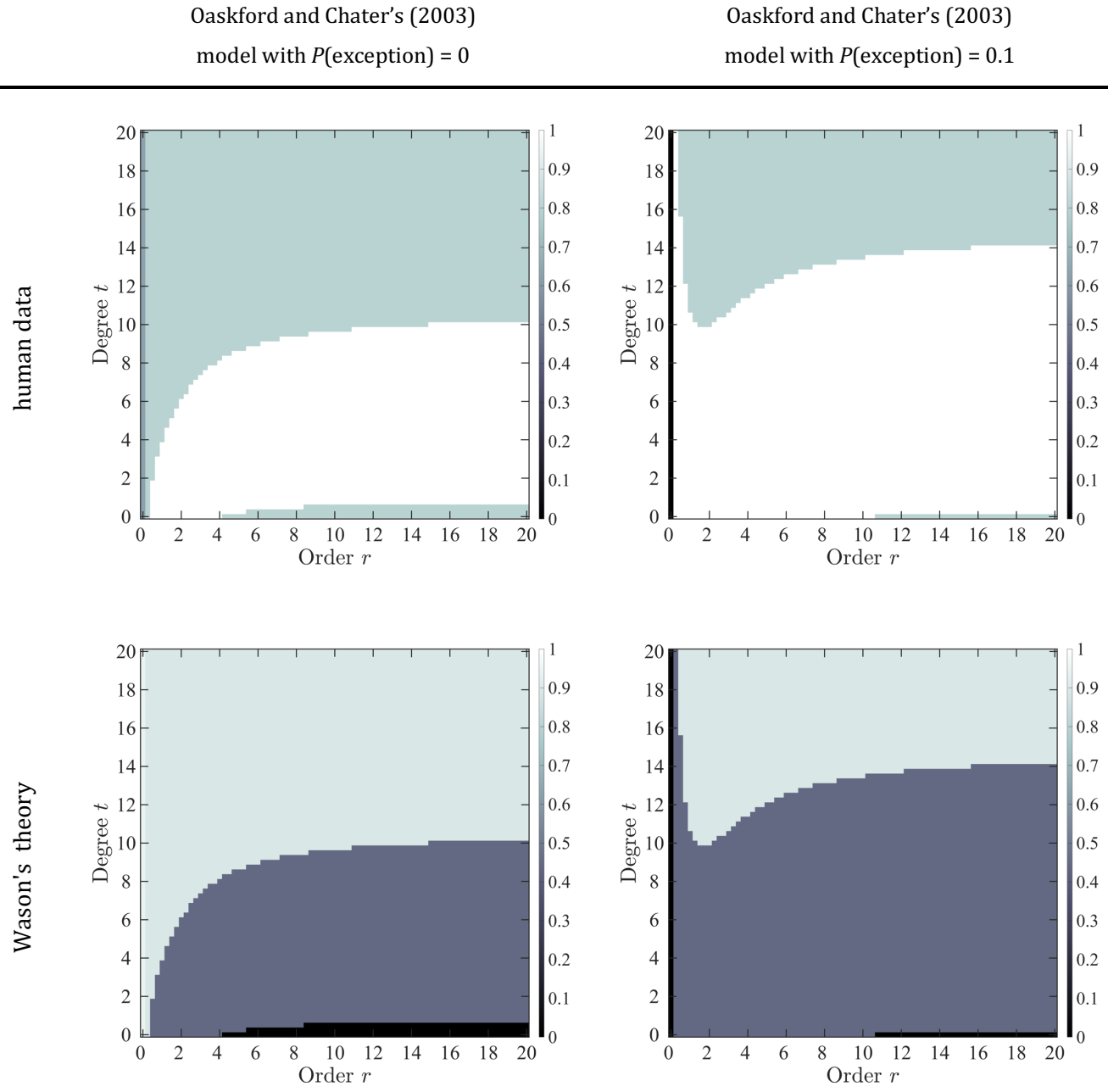

Figure 6. Plots of rank correlation values for the expected reduction of various Sharma-Mittal entropies in Oaksford and Chater's (2003) model of the Wason selection task. In the top row, models of expected entropy reduction are compared with empirical aggregate card selection frequencies. In the bottom row, instead, the comparison is with theoretical choices implied by Wason's original analysis of the task. In the left *vs*. right columns the conditional probability representation of "if vowel, then even number" rules out expections or allows for them (with probability 0.1), respectively.

The Wason selection task illustrates the theoretical potential of the Sharma-Mittal framework. Whereas other authors noted the robustness of probabilistic analyses of the task across different measures of informational utility (see Fitelson & Hawthorne, 2010; Nelson, 2005, pp. 985-986; Oaksford & Chater, 2007), the variety of measures involved in those analyses arose in an *ad hoc* way. We extend those results, and show that even the traditional, allegedly anti-Bayesian reading of the task can be recovered smoothly in one overarching framework. In particular, the implications of Wason's Popperian interpretation can be represented well by the maximization of the expected reduction of an entirely balanced (order-0) Sharma-Mittal measure (such as Origin or Hartley entropy) in a deterministic reading of the task (i.e., with $P(\text{exception}) = 0$). Conversely, this means that adopting a probabilistic approach to Wason's task is not by itself sufficient to account for observed behavior. Even then, in fact, people's choices would still diverge from at least some theoretically viable models of information search.

6.2. Information search in experience-based studies

Is the same expected uncertainty reduction measure able to account for human behavior across a variety of tasks? To explore this issue, we reviewed experimental scenarios employed in experience-based investigations of information search behavior. In this experimental paradigm, participants learn the underlying statistical structure of an environment where items (plankton specimens) are visually displayed and subject to a binary classification (kind *A vs*. *B*) for which two binary features (yellow *vs*. black eye; dark *vs*. light claw) are potentially relevant. Immediate feedback is provided after each trial in a learning phase, until a performance criterion is reached, indicating adequate mastery of the environmental statistics. In a subsequent information-acquisition test phase of this procedure, both of the two features (eye and claw) are obscured, and participants have to select the most informative / useful feature relative to the target categories (kinds of plankton). (See Nelson et al., 2010, for a detailed description.) In our current terms, these scenarios concern a binary hypothesis space H = {specimen of kind *A*, specimen of kind *B*} and two binary tests E = {yellow eye, black eye} and F = {dark claw, light claw}. In each

case, the experience-based learning phase conveyed the structure of the joint probability distribution $P(H,E,F)$ to participants. The test phase, in which either feature E or F can be viewed, represents a way to see whether the participants deemed $R_P(H,E)$ or $R_P(H,F)$ to be greater.

Table 6. *Choices between two binary tests / experiments* (E *vs.* F) *for a binary classification problem* (H) *in experience-based experimental procedures. Cases 1-3 are taken from Nelson et al. (2010, Exp. 1); cases 4-5 from Exp. 3 in the same article; case 6 is an unpublished study using the same experimental procedure; cases 7-8 are from Meder and Nelson (2012, Exp. 1).*

n.	$P(H)$	Test E		Test F		% observed choices of E
		$P(H\|e)$ *vs.* $P(H\|\bar{e})$	$P(e)$ *vs.* $P(\bar{e})$	$P(H\|f)$ *vs.* $P(H\|\bar{f})$	$P(f)$ *vs.* $P(\bar{f})$	
1	{0.7, 0.3}	{0, 1} {0.754, 0.246}	0.072 0.928	{1, 0} {0.501, 0.499}	0.399 0.601	82% (23/28)
2	{0.7, 0.3}	{0, 1} {0.767, 0.233}	0.087 0.913	{1, 0} {0.501, 0.499}	0.399 0.601	82% (23/28)
3	{0.7, 0.3}	{0.109, 0.891} {0.978, 0.022}	0.320 0.680	{1, 0} {0.501, 0.499}	0.399 0.601	97% (28/29)
4	{0.7, 0.3}	{0, 1} {0.733, 0,.267}	0.045 0.955	{1, 0} {0.501, 0.499}	0.399 0.601	89% (8/9)
5	{0.7, 0.3}	{0.201, 0.799} {0.780, 0.220}	0.139 0.861	{1, 0} {0.501, 0.499}	0.399 0.601	70% (14/20)
6	{0.7, 0.3}	{0.135, 0.865} {0.848, 0.152}	0.208 0.792	{1, 0} {0.501, 0.499}	0.399 0.601	70% (14/20)
7	{0.44, 0.56}	{0.595, 0.405} {0.331, 0.669}	0.414 0.586	{0, 1} {0.502, 0.498}	0.123 0.877	60% (12/20)
8	{0.36, 0.64}	{0.090, 0.910} {0.707, 0.293}	0.562 0.438	{0, 1} {0.501, 0.499}	0.282 0.118	79% (15/19)

Overall, we found eight relevant experimental scenarios from the experimental paradigm described above (they are listed in Table 6) in which there was at least some interesting disagreement among the Sharma-Mittal measures about which feature is more useful. For each, we derived values of expected uncertainty reduction from Sharma-Mittal

measures of order and degree from 0 to 20, in increments of 0.25, and we computed the simple proportion of cases in which each measure's ranking of $R_P(H,E)$ and $R_P(H,F)$ matched the most prevalent observed choice.

Nelson et al. (2010) devised their scenarios to dissociate predictions from a sample of competing and historically influential models of rational information search. Their conclusion was that the expected reduction of Error entropy (expected *probability gain*, in their terminology) accounted for participants' behavior and outperformed the expected reduction of Shannon entropy (expected *information gain*, in their terminology). A more comprehensive analysis within our current approach implies a richer picture. The data set employed can be accurately represented in the Sharma-Mittal framework for a significant range of degree values provided that the order parameter is high enough (the results are displayed in Figure 7, left side). Observed choices are especially consistent with expected reduction of a quite unbalanced (e.g., $r \geq 4$), concave or quasi-concave (t close to 2) Sharma-Mittal entropy measure. Importantly, there is overlap between results from modeling the Wason selection task and these experience-based learning data, giving hope to the idea that a unified theoretical explanation of human behavior may extend across several tasks.

6.3. Information search in words-and-numbers studies

The experience-based learning tasks discussed above were inspired by analogous tasks in which the prior probabilities of categories and feature likelihoods were presented to participants using words and numbers (e.g., Skov and Sherman, 1986). We refer to such tasks as Planet Vuma experiments, reflecting the typically whimsical content, such as classifying species of aliens on Planet Vuma, designed to not conflict with people's experience with real object categories.

Whereas expected reduction of Error entropy, and other models as discussed above, gives a plausible explanation of the experience-based learning task data, individual data in words-and-numbers studies are very noisy, and no attempt has been made to see whether a unified theory could account for the modal responses across these tasks. We therefore re-analyzed empirical data from several Planet Vuma experiments, in a manner analogous to

our analyses of the experience-based learning data above (Figure 7). What do the results show? To our surprise, the results suggest that there may be a systematic explanation of people's behavior on words-and-numbers-based tasks.

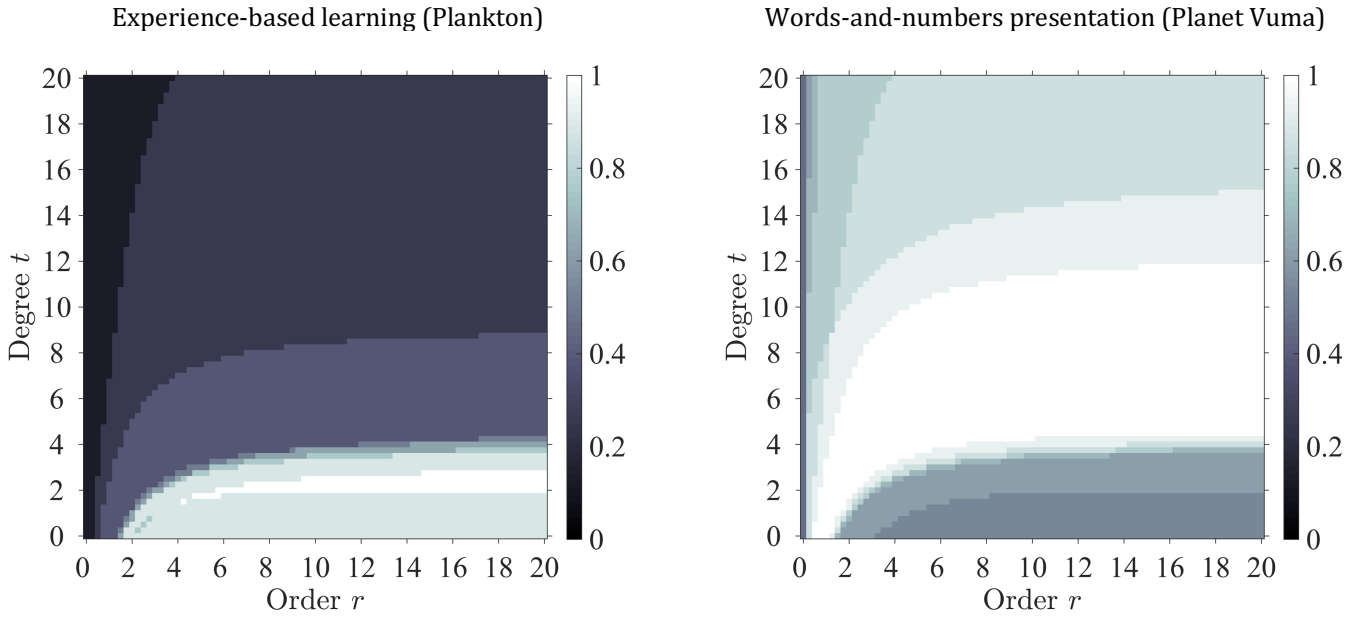

Figure 7. On the left, a graphical illustration of the empirical accuracy of Sharma-Mittal measures relative to binary information search choices in 8 experience-based experimental scenarios (described in Table 6). The shade at each point illustrates the proportion of choices (out of 8) correctly predicted by the expected reduction of the corresponding underlying entropy, with white and black indicating maximum (8/8) and minimum (0/8) accuracy, respectively. Results suggest that an Arimoto metric of moderate or high order is highly consistent with human choices. On the right, illustration of the empirical accuracy of Sharma-Mittal measures in theoretically similar tasks, but where probabilistic information is presented in a standard explicit format (with numeric prior probabilities and test likelihoods). In these tasks, individual participants' test choices are highly noisy. Can a systematic theory still account for the modal results across tasks? We analyzed 13 cases (described in Table 7) of binary information search preferences. The shade at each point illustrates the proportion of comparisons (out of 13) correctly predicted by the expected reduction of the corresponding underlying entropy, with white and black again indicating maximum (13/13) and minimum (0/13) accuracy, respectively. Results show that a wide range of measures is consistent with available experimental findings, including Shannon entropy as well as a variety of high-degree measures (degree much higher than the Arimoto curve).

Table 7. *Choices between two binary tests / experiments (E vs. F) for a binary classification problem (H) in words-and-numbers (Vuma Planet) experiments. Cases 1-6 are from Nelson (2005); case 7 is from Skov & Sherman (1986), cases 8-10 are from Nelson et al. (2010, Exp. 1); cases 11-13 from Wu et al. (in press, Exp. 1-3). In each case, test E was deemed more useful than test F by the participants. We only report scenarios for which at least two Sharma-Mittal measures strictly disagree about which of the tests has higher expected usefulness. (Thus, not all feature queries involved in the original articles are listed here.) Nelson (2005) asked participants to give a rank ordering among four possible features' information values. Here we list the six corresponding pairwise comparisons, in each case labeling the feature that was ranked higher as the favorite one (E). Wu et al. (in press) studied 14 different probability, natural frequency, and graphical information formats for the presentation of relevant probabilities. For comparison with other studies, we take results only from the standard probability format here.*

n.	$P(H)$	Test E		Test F	
		$P(e\|h)$	$P(e\|\overline{h})$	$P(f\|h)$	$P(f\|\overline{h})$
1	{0.5, 0.5}	0.70	0.30	0.99	1.00
2	{0.5, 0.5}	0.30	0.0001	0.99	1.00
3	{0.5, 0.5}	0.01	0.99	0.99	1.00
4	{0.5, 0.5}	0.30	0.0001	0.70	0.30
5	{0.5, 0.5}	0.01	0.99	0.30	0.0001
6	{0.5, 0.5}	0.01	0.99	0.70	0.30
7	{0.5, 0.5}	0.90	0.55	0.65	0.30
8	{0.7, 0.3}	0.57	0	0	0.24
9	{0.7, 0.3}	0.57	0	0	0.29
10	{0.7, 0.3}	0.05	0.95	0.57	0
11	{0.7, 0.3}	0.41	0.93	0.03	0.30
12	{0.7, 0.3}	0.43	1.00	0.04	0.37
13	{0.72, 0.28}	0.03	0.83	0.39	1.00

The degree of the most plausible measures is considerably above the Arimoto curve, although not as high as, for instance, Non-Certainty entropy (order 30). From a descriptive psychological standpoint, a plausible interpretation is that when confronted with words-and-numbers-type tasks, people have a strong focus on the chances of obtaining a certain or near-to-certain result, and are less concerned with (or, perhaps, attuned to) the details of the individual items in the probability distribution. The Sharma-Mittal framework provides potential explanation for heretofore perplexing experimental results, while also highlighting key questions (e.g., how much preference for near-certainty, exactly, do subjects have) for future empirical research on words-and-numbers tasks.

6.4. Unifying theory and intuition in the Person Game (Having your cake and eating it too)

In this section, we introduce another theoretical conundrum from the literature, and show how the Sharma-Mittal framework may help solve it. As pointed out above, the expected reduction of Error entropy had appeared initially to provide the best explanation of people's intuitions and behavior on experience-based-learning-based information search tasks (Nelson et al., 2010). But this model leads to potentially counterintuitive behavior on another interesting kind of information search task, namely the Person Game (a variant of the Twenty Questions game). In this game, n cards (say, 20) with different faces are presented. One of those faces has been chosen at random (with equal probability) to be the correct face in a particular round of the game. The player's task is to find the true face in the smallest number of yes/no questions about physical features of the faces. For instance, asking whether the person has a beard would be a possible question, $E = \{e,\overline{e}\}$, with e = beard and $\overline{e}$ = no beard. If $k < n$ is the number of characters with a beard, then $P(e) = k/n$ and $P(\overline{e}) = (n - k)/n$. Moreover, a "yes" answer will leave k equiprobable guesses still in play, and a "no" answer $n - k$ such guesses.

Several papers have reported (see Nelson et al., 2014, for references) that people preferentially ask about features that are possessed by close to 50% of the remaining possible items, thus with $P(e)$ close to 0.5. This strategy can be labelled the *split-half*

heuristic. It is optimal to minimize the expected number of questions needed under some task variants (Navarro & Perfors, 2011), although not in the general case (Nelson, Meder, & Jones, 2016), and can be accounted for using expected Shannon entropy reduction. But expected Shannon entropy reduction cannot account for people's behavior on experience-based learning information search tasks, as our above analyses show. Can expected Error entropy reduction account for these results and intuitions? Put more broadly, can the same entropy model provide a satisfying account for both the Person Game and the experience-based learning tasks? As it happens, Error entropy cannot account for the preference to split the remaining items close to 50%. In fact, every possible question (unless its answer is known already, because none or all of the remaining faces have the feature) has exactly the same expected Error entropy reduction, namely $1/k$, where there are k items remaining (Nelson, Meder, & Jones, 2016). This might lead us to wonder whether we must have different entropy/information models to account for people's intuitions and behavior across these different tasks. Indeed, it would call into question the potential for a unified and general purpose theory of the psychological value of information.

It turns out that the findings on why expected Shannon entropy reduction favors questions close to a 50:50 split, and why Error entropy has no such preference, apply much more generally than to Shannon and Error entropy. In fact, for all Sharma-Mittal measures, the ordinal evaluation of questions on the Person Game is solely a function of the degree of the entropy measure, and has nothing to do with the order of the measure [Suppl Mat, 5]. Among other things, this implies that all entropy-based measures with degree $t = 1$ have the exact same preferences as expected Shannon entropy reduction, and all of them quantify the usefulness of querying a feature as a function of the proportion of remaining items that possess that feature. Similarly, all degree-2 measures, and not only Error entropy, deem all questions to be equally useful in the Person Game. The core of this insight stems from the fact that, if a probability distribution is uniform, then the entropy of that distribution depends only on the degree of a Sharma-Mittal entropy measure. More formally, for any set of hypotheses $H = \{h_1, h_2, \ldots h_n\}$ with a uniform probability distribution $U(H)$:

$$ent_U^{SM_{(r,t)}}(H) = ln_t(n)$$

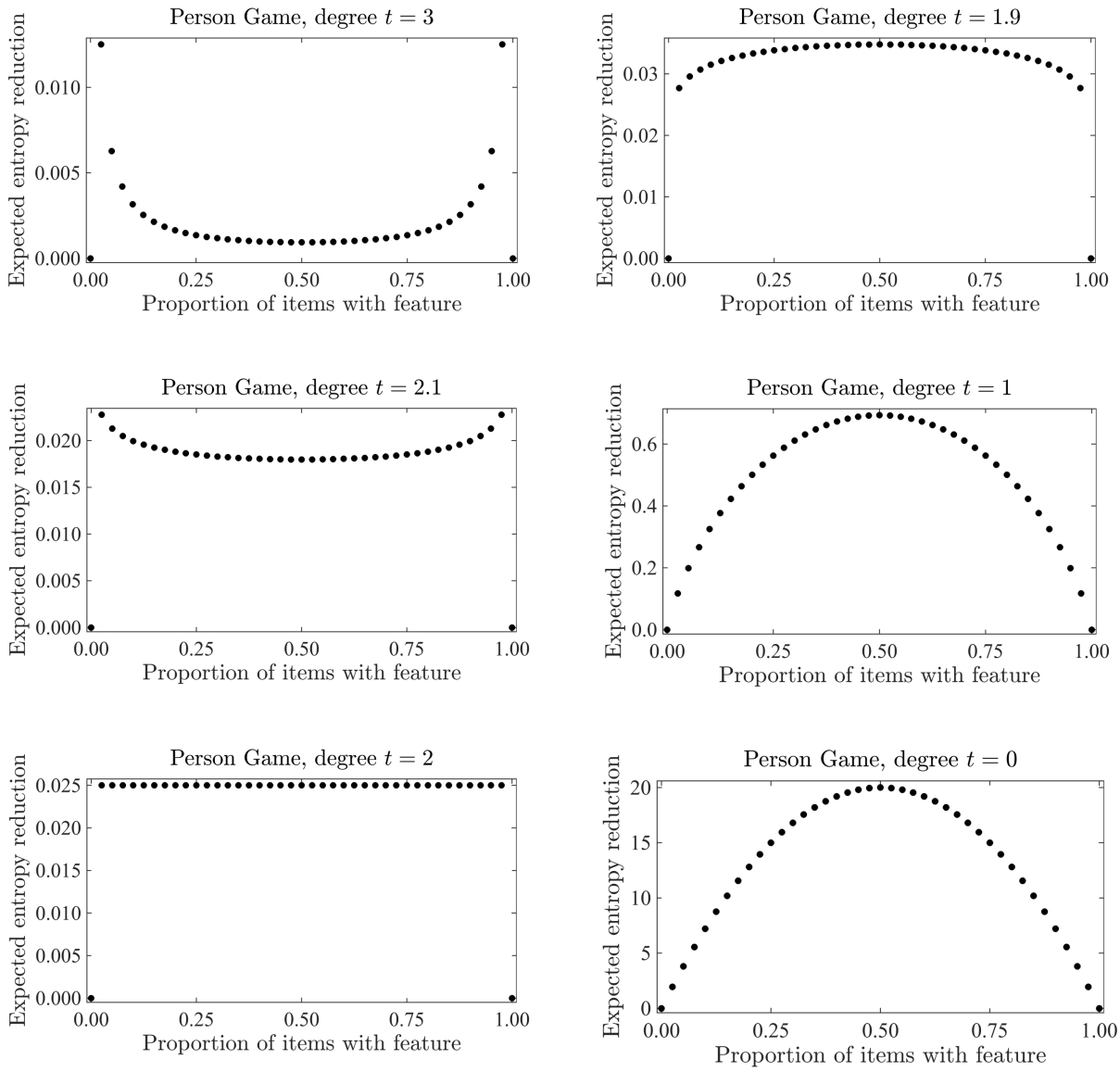

Figure 8. The expected entropy reduction of a binary question $E = \{e,\overline{e}\}$ in the Person Game with a hypothesis set H of size 40 (the possible guesses, that is, characters initially in play) as a function of the proportion of possible guesses remaining after getting datum e (e.g., a "yes" answer to "has the chosen person a beard?"). Questions are deemed most valuable with the zero-degree entropy measures (bottom right plot). Although the shape of the curve is similar for the degree $t = 0$ and degree $t = 1$ measures, the actual information value (see the y axis) decreases as the degree increases. For degree $t = 2$ (for example for Error entropy), every question is equally useful (provided that there is some uncertainty about the answer; bottom left plot). If the degree is greater than 2, then the least-equally-split questions (e.g., 1:39 questions, in the case of 40 items) are deemed most useful (left column, top and middle row). The order parameter is irrelevant for purposes of evaluating questions' expected usefulness in the Person Game, because all prior and possible posterior probability distributions are uniform (see text).

Figure 8 shows how possible questions are valued, in the Person Game, as a function of the proportion of remaining items that possess a particular feature. We see that if $t = 1$, as for Shannon and all Rényi entropies, questions with close to a 50:50 split are preferred. If the degree t is greater than 1 but less than 2, questions with close to a 50:50 split are still preferred, but less so. If $t = 2$, then 1:99 and 50:50 questions are deemed equally useful. Remarkably, if the degree is greater than 2, then a 1:99 question is preferred to a 50:50 question.

While the choice of particular Sharma-Mittal measures is only partly constrained by observed preferences in the Person Game alone (and specifically the value of the order parameter r is not), nothing in principle would guarantee that a joint and coherent account of such behavior and other findings exists. It is then important to point out that one can, in fact, pick up an entropy measure whereby the experience-based data above follow along with a greater informative value for 50:50 questions than for 1:99 questions in the Person Game. For instance, medium-order Arimoto entropies (such as $ent^{SM_{(10,1.9)}}$) will work.

7. General discussion

In this paper, we have presented a general framework for the formal analysis of uncertainty, the Sharma-Mittal entropy formalism. This framework generates a comprehensive approach to the informational value of queries (questions, tests, experiments) as the expected reduction of uncertainty. The amount of theoretical insight and unification achieved is remarkable, in our view. Moreover, such a framework can help us understand existing empirical results, and point out important research questions for future investigation of human intuition and reasoning processes as concerns uncertainty and information search.

Mathematically, the parsimony of the Sharma-Mittal formalism is appealing and yields decisive advantages in analytic manipulations, derivations, and calculations, too. Within the domain of cognitive science, no earlier attempt has been made to unify so many existing models concerning information search / acquisition behavior. Notably, this involves both popular candidate rational measures of informational utility (such as the expected

reduction of Shannon or Error entropy) and avowed heuristic models, such as Baron et al.'s (1988, 106) quasi-Popperian heuristic (maximization of the expected number of hypotheses ruled out, i.e., the expected reduction of Origin entropy) and Nelson et al.'s (2010, 962) "probability-of-certainty" heuristic (closely approximated by the expected reduction of a high degree Tsallis entropy, or a similar measure). In addition, once applied to uncertainty and information search, the Sharma-Mittal parameters are not dumb mathematical construals, but rather capture cognitively and behaviorally meaningful ideas. Roughly, the order parameter, *r*, captures how much one disregards minor hypotheses (via the kind of means applied to the probability values in $P(H)$). The degree parameter *t*, on the other hand, captures how much one cares about getting (very close) to certainty (via the behavior of the surprise / atomic information function; see Figure 3). Thus, high order indicates a strong focus on the prevalent (most likely) element in the hypothesis set and lack of consideration for minor possibilities. A very low order, on the other hand, implies a Popperian or quasi-Popperian attitude in the assessment of tests, with a marked appreciation of potentially falsifying or almost falsifying evidence. The degree parameter, in turn, has important implications for how much potentially conclusive experiments are valued, as compared to experiments that are informative but not conclusive. Moreover, for each particular order, if the degree is higher than the corresponding Arimoto entropy (and in any case if the order is less than 0.5 or the degree is at least 2), then the concavity of the entropy measure guarantees that no experiment will be rated as having negative expected usefulness.

Even according to fairly cautious views such as Aczel's (1984), the above remarks seem to provide a fairly strong motivation to consider pursuing a generalized approach. Here is another possible concern, however. Uncertainty and the informational value of tests may be involved in many arguments concerning human cognition. Now we see that those notions can be formalized in many different ways, such that different properties (say, additivity, or non-negativity) are or are not implied. Thus, the arguments at issue might be valid for some choices of the corresponding measures and not for others. This point has been labelled the issue of *measure-sensitivity* in related areas (Fitelson, 1999) — is it something to be worried about? Does it raise problems for our proposal?

It is not uncommon for measure-sensitivity to foster skeptical or dismissive reactions on the prospects of the formal analysis of the concept at issue (e.g. Hurlbert, 1971, Kyburg & Teng, 2001, pp. 98 ff.). However, measure-sensitivity is a widespread and mundane phenomenon. In areas related to the formal analysis of reasoning, the issue arises, for instance, for Bayesian theories of inductive confirmation (e.g., Brössel 2013; Crupi & Tentori, 2016; Festa & Cevolani, 2016; Glass, 2013; Hájek & Joyce, 2008; Roche & Shogenji, 2014), scoring rules and measures of accuracy (e.g., D'Agostino & Sinigaglia, 2010; Leitgeb & Pettigrew, 2010a,b; Levinstein, 2012; Predd *et al*., 2009), and measures of causal strength (e.g., Griffiths & Tenenbaum, 2005, 2009; Fitelson & Hitchcock, 2011; Meder, Mayrhofer, & Waldmann, 2014; Sprenger, 2016). Our treatment contributes to make the same point explicit for measures of uncertainty and the informational value of experiments. This we see as a constructive contribution. The prominence of one specific measure in one research domain may well have been partly affected by historical contingencies. As a consequence, when a theoretical or experimental inference relies on the choice of one measure, it makes sense to check how robust it is across different choices or, alternatively, to acknowledge which measure-specific properties support the conclusion and how compelling they are. Having a plurality of related measures available is indeed an important opportunity. It prompts thorough investigation of the features of alternative options and their relationships (e.g., Crupi, Chater, & Tentori, 2013; Huber & Schmidt-Petri, 2009; Nelson, 2005, 2008), it can provide a rich source of tools for both theorizing and the design of new experimental investigations (e.g., Rusconi et al., 2014; Schupbach, 2011; Tentori et al., 2007), and it makes it possible to tailor specific models to varying tasks and contexts within an otherwise coherent approach (e.g., Crupi & Tentori, 2014; Dawid & Musio, 2014; Oaksford & Hahn, 2007).

Which Sharma-Mittal measures are more consistent with observed behavior overall? According to our analyses, a subset of Sharma-Mittal information search models receives a significant amount of convergent support. We found that measures of high but finite order accounting for the experience-based (plankton task) data (Figure 7, left side) are also empirically adequate for abstract selection task data (Figure 6, top row) and results from a Twenty Questions kind of task such as the Person Game (Figure 8). On the other hand, the

best fit with words-and-numbers (Planet Vuma) information search tasks indicates a different kind of model within the Sharma-Mittal framework (Figure 7, right side). For these cases, our analysis thus suggests that people's behavior may comply with different measures in different situations, so a key question arises about the features of a task which affect such variation in a consistent way, such as a comparably stronger appreciation of certainty or quasi-certainty as prompted by an experimental procedure conveying environmental statistics by explicit verbal and numerical stimuli.

Beyond this broad outlook, our discussion also allows for the resolution of a number of puzzles. Let us mention a last one. Nelson et al. (2010) had concluded from their experimental investigations that human information search in an experience-based setting was appropriately accounted for by maximization of the expected reduction of Error entropy. This specific model, however, exhibits some questionable properties related to its lack of mathematical continuity: in particular, if the most likely hypothesis in H is not changed by any possible evidence in E, then the latter has no informational utility whatsoever according to R^{Error}, no matter if it can rule out other non-negligible hypotheses in the set (see, e.g., cases 1 and 6 in Table 6). Findings from Baron et al. (1988) suggest that this might not describe human judgment adequately. In that study, participants were given a fictitious medical diagnosis scenario with $P(H) = \{0.64, 0.24, 0.12\}$, and a series of possible binary tests including E such that $P(H|e) = \{0.47, 0.35, 0.18\}$, $P(H|\overline{e}) = \{1,0,0\}$ and $P(e) = 0.68$ and another completely irrelevant test F (with an even chance of a positive / negative result on each one of the elements in H, so that $P(H|f) = P(H|\overline{f}) = P(H)$). According to R^{Error}, tests E and F are both equally worthless — $R_P^{Error}(H,E) = R_P^{Error}(H,F) = 0$ — because hypothesis $h_1 \in H$ remains the most likely no matter what. Participants' mean ratings of the usefulness of E and F were markedly different, however: 0.48 *vs*. 0.09 (on a 0-1 scale). Indeed, rating E higher than F seems at least reasonable, contrary to what R^{Error} implies. In the Sharma-Mittal framework, reconciliation is possible: expected reduction of a relatively high order (say, 10) entropy measure from the Arimoto family would account for Nelson et al.'s (2010) and similar findings (see Figure 7), and still would not put test E above on a par with the entirely pointless test F. Indeed, given our theoretical background and the limited empirical

indications available, such a measure would count as a plausible choice in our view, had one to pick up a specific entropy underlying a widely applicable model of the informational utility of experiments. Moreover, this kind of operation has wider scope. Origin entropy, for instance, may imply largely appropriate ratings in some contexts (say, biological) and yet not be well-behaved because of its discontinuities: a Sharma-Mittal measure such as $ent^{SM_{(0.1,0.1)}}$ would then closely approximate the former while avoiding the latter.

Many further empirical issues can be addressed. For one instance, our analysis of human data in Tables 6-7 and Figure 7 provides relatively weak and indirect evidence against non-concave entropy measures as a basis for the assessment of the informational utility of queries by human agents. However, strongly diverging predictions can be generated from concave *vs*. non-concave measures (as illustrated in cases 6 and 7, Table 5), and hence put to empirical test. Moreover, our explanatory reanalysis of prior work was based on the aggregate data reported in earlier articles — but how does this extend to individual behavior? We are aware of no studies that address questions of whether there are meaningful individual differences in the psychology of information. Thus, while inferences about individuals should be the goal (Lee, 2011), this requires future research, perhaps with adaptive Bayesian experimental design techniques (Kim et al., 2014). Better models of individual-level psychology could also serve the goal of identifying the information that would be most informative for individual human learners (Gureckis & Markant, 2012), potentially enhancing automated tutor systems. Another idea concerns the direct assessment of uncertainty, e.g., whether more uncertainty is perceived in, say, $P(H)$ = {0.49, 0.49, 0.02} *vs*. $P^*(H)$ = {0.70, 0.15, 0.15}. Judgments of this kind are likely to play a role in human reasoning and decision-making and may be plausibly modulated by a number of interesting factors. Moreover, an array of relevant predictions can be generated from the Sharma-Mittal framework to dissociate subsets of entropy measures. Yet as far as we know, and rather surprisingly, no established experimental procedure exists for a direct behavioral measurement of the judged overall uncertainty concerning a hypothesis set; this is another important area for future investigation.

REFERENCES

Aczél J. (1984). Measuring information beyond communication theory: Why some generalized information measures may be useful, others not. *Aequationes Mathematicae*, *27*, 1-19.

Aczél J. (1987). Characterizing information measures: Approaching the end of an era. *Lecture Notes in Computer Science*, *286*, 357-384.

Aczél J., Forte B., & Ng C.T. (1974). Why the Shannon and Hartley entropies are "natural". *Advances in Applied Probability*, *6*, 131-146.

Arimoto S. (1971). Information-theoretical considerations on estimation problems. *Information and Control*, *19*, 181-194.

Austerweil J.L. & Griffiths T.L. (2011). Seeking confirmation is rational for deterministic hypotheses. *Cognitive Science*, *35*, 499-526.

Bar-Hillel Y. & Carnap R. (1953). Semantic information. *British Journal for the Philosophy of Science*, *4*, 147-157.

Baron J. (1985). *Rationality and intelligence*. New York: 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.

Barwise J. (1997). Information and possibilities. *Notre Dame Journal of Formal Logic*, *38*, 488-515.

Beck C. (2009). Generalised information and entropy measures in physics. *Contemporary Physics*, *50*, 495-510.

Ben-Bassat M. & Raviv J. (1978). Rényi's entropy and the probability of error. *IEEE Transactions on Information Theory*, *24*, 324-331.

Benish W.A. (1999). Relative entropy as a measure of diagnostic information. *Medical Decision Making*, *19*, 202-206.

Boztas S. (2014). On Rényi entropies and their applications to guessing attacks in cryptography. *IEICE Transactions on Fundamentals of Electronics, Communication, and Computer Sciences*, *97*, 2542-2548.

Bradley S. & Steele K. (2014). Uncertainty, learning and the 'problem' of dilation. *Erkenntnis*, *79*, 1287–1303.

Bramley N.R., Lagnado D., & Speekenbrink M. (2015). Conservative forgetful scholars: How people learn causal structure through sequences of interventions. *Journal of Experimental Psychology*: *Learning, Memory, and Cognition*, *41*, 708-731.

Brier G.W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Report*, *78*, 1-3.

Brössel P. (2013). The problem of measure sensitivity redux. *Philosophy of Science*, *80*, 378–397.

Carnap R. (1952). *The Continuum of Inductive Methods*. Chicago: University of Chicago Press.

Cho A. (2002). A fresh take on disorder, or disorderly science. *Science*, 297, 1268–1269.

Crupi V. & Girotto V. (2014). From *is* to *ought*, and back: How normative concerns forster progress in reasoning research. *Frontiers in Psychology*, *5*, 219.

Crupi V. & Tentori K. (2014). Measuring information and confirmation. *Studies in the History and Philosophy of Science*, *47*, 81-90.

Crupi V. & Tentori K. (2016). Confirmation theory. In A. Hájek & C. Hitchcock (eds.), *Oxford Handbook of Philosophy and Probability* (pp. 650-665). Oxford: Oxford University Press.

Crupi V., Chater N., & Tentori K. (2013). New axioms for probability and likelihood ratio measures. *British Journal for the Philosophy of Science*, *64*, 189–204.

Crupi V., Tentori K., & Lombardi L. (2009). Pseudodiagnosticity revisited. *Psychological Review*, *116*, 971-985.

Csizár I. (2008). Axiomatic characterizations of information measures. *Entropy*, *10*, 261-273.

D'Agostino M. & Sinigaglia C. (2010). Epistemic accuracy and subjective probability. In M. Suárez, M. Dorato, & M. Rèdei (eds.), *Epistemology and Methodology of Science* (pp. 95-105). Berlin: Springer.

Daróczy Z. (1970). Generalized information functions. *Information and Control*, *16*, 36-51.

Dawid A.P. (1998). Coherent measures of discrepancy, uncertainty and dependence, with applications to Bayesian predictive experimental design. Technical Report 139, Department of Statistical Science, University College London (http://www.ucl.ac.uk/Stats/research/pdfs/139b.zip).

Dawid A.P. & Musio M. (2014). Theory and applications of proper scoring rules. *Metron*, *72*, 169-183.

Denzler J. & Brown C.M (2002). Information theoretic sensor data selection for active object recognition and state estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *24*, 145-157.

Evans J.St.B.T. & Over D.E. (1996). Rationality in the selection task: Epistemic utility versus uncertainty reduction. *Psychological Review*, *103*, 356-363.

Fano R. (1961). *Transmission of Information: A Statistical Theory of Communications.* Cambridge: MIT Press.

Festa R. (1993). *Optimum Inductive Methods*. Rijksuniversiteit Groningen.

Festa R. & Cevolani G. (2016). Unfolding the grammar of Bayesian confirmation: Likelihood and anti-likelihood principles. *Philosophy of Science*, forthcoming.

Fitelson B. (1999). The plurality of Bayesian measures of confirmation and the problem of measure sensitivity. *Philosophy of Science*, *66,* S362–S378.

Fitelson B. & Hawthorne J. (2010). The Wason task(s) and the paradox of confirmation. *Philosophical Perspectives*, *24* (*Epistemology*), 207-241.

Fitelson B. & Hitchcock C. (2011). Probabilistic measures of causal strength. In P. McKay Illari, F. Russo, & J. Williamson (eds.), *Causality in the Sciences* (pp. 600–27). Oxford: Oxford University Press.

Floridi L. (2009). Philosophical conceptions of information. In G. Sommaruga (ed.), *Formal Theories of Information* (pp. 13-53). Berlin: Springer.

Floridi L. (2013). Semantic conceptions of information. In E. Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2013 Edition). url = http://plato.stanford.edu/archives/spr2013/entries/information-semantic.

Frank T. (2004). Complete description of a generalized Ornstein-Uhlenbeck process related to the non-extensive Gaussian entropy. *Physica A*, *340*, 251-256.

Gauvrit N. & Morsanyi K. (2014). The equiprobability from a mathematical and psychological perspective. *Advances in Cognitive Psychology*, *10*, 119-130.

Gibbs J.P. & Martin W.T. (1962). Urbanization, technology, and the division of labor. *American Sociological Review*, *27*, 667-677.

Gigerenzer G., Hertwig R., & Pachur T. (eds.) (2011). *Heuristics: The Foundations of Adaptive Behavior*. New York: Oxford University Press.

Gini C. (1912). Variabilità e mutabilità. In *Memorie di metodologia statistica*, I: *Variabilità e concentrazione* (pp. 189-358). Milano: Giuffrè, 1939.

Glass D.H. (2013). Confirmation measures of association rule interestingness. *Knowledge-Based Systems*, *44*, 65-77.

Gneiting T. & Raftery A.E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, *102*, 359-378.

Good I.J. (1950). *Probability and the Weight of Evidence*. New York: Griffin.

Good I.J. (1952). Rational Decisions. *Journal of the Royal Statistical Society* B, *14*, 107-114.

Good I.J. (1967). On the principle of total evidence. *British Journal for the Philosophy of Science*, *17*, 319-321.

Goosens W.K. (1976). A critique of epistemic utilities. In R. Bogdan (ed.), *Local induction* (pp. 93-114). Dordrecht: Reidel.

Griffiths T.L. & Tenenbaum J.B. (2005). Structure and strength in causal induction. *Cognitive Psychology*, *51*, 334-384.

Griffiths T.L. & Tenenbaum J.B. (2009). Theory-based causal induction. *Psychological Review*, *116*, 661-716.

Gureckis T.M. & Markant D.B. (2012). Self-directed learning: A cognitive and computational perspective. *Perspectives on Psychological Science*, *7*, 464-481.

Hájek A. & Joyce J. (2008). Confirmation. In S. Psillos & M. Curd (eds.), *Routledge Companion to the Philosophy of Science* (pp. 115-129). New York: Routledge.

Hartley R. (1928). Transmission of information. *Bells Systems Technical Journal*, *7*, 535-563.

Hasson U. (2016). The neurobiology of uncertainty: Implications for statistical learning. *Philosophical Transactions B*, *371*, 20160048.

Hattori M. (1999). The effects of probabilistic information in Wason's selection task: An analysis of strategy based on the ODS model. In *Proceedings of the 16th Annual Meeting of the Japanese Cognitive Science Society* (pp. 623-626).

Hattori M. (2002). A quantitative model of optimal data selection in Wason's selection task. *Quarterly Journal of Experimental Psychology*, *55*, 1241-1272.

Havrda J. & Charvát F. (1967). Quantification method of classification processes. Concept of structural *a*-entropy. *Kybernetica*, *3*, 30-35.

Hill M. (1973). Diversity and evenness: A unifying notation and its consequences. *Ecology*, *54*, 427-431.

Hoffmann S. (2008). Generalized distribution-based diversity measurement: Survey and unification. Faculty of Economics and Management Magdeburg, Working Paper 23 (http://www.ww.uni-magdeburg.de/fwwdeka/femm/a2008_Dateien/2008_23.pdf).

Horwich P. (1982). *Probability and Evidence*. Cambridge, UK: Cambridge University Press.

Huber F. & Schmidt-Petri C. (eds.) (2009). *Degrees of Belief*. Dordrecht: Springer.

Hurlbert S.H. (1971). The non-concept of species diversity: A critique and alternative parameters. *Ecology*, *52*, 577-586.

Jost L. (2006). Entropy and diversity. *Oikos*, *113*, 363-375.

Kaniadakis G., Lissia M., & Scarfone A.M. (2004). Deformed logarithms and entropies. *Physica A*, *340*, 41-49.

Katsikopoulos K.V., Schooler L.J., & Hertwig R. (2010). The robust beauty of ordinary information, *Psychological Review*, *117*, 1259-1266.

Keylock J.C. (2005). Simpson diversity and the Shannon-Wiener index as special cases of a generalized entropy. *Oikos*, *109*: 203-207.

Kim W., Pitt M.A., Lu Z.L., Steyvers M., & Myung, J.I. (2014). A hierarchical adaptive approach to optimal experimental design. *Neural Computation*, *26*, 2465-2492.

Klayman J. & Ha Y.-W. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review*, *94*, 211-228.

Kyburg H.E. & Teng C.M. (2001). *Uncertain Inference*. New York: Cambridge University Press.

Laakso M. & Taagepera R. (1979). "Effective" number of parties – A measure with application to West Europe. *Comparative Political Studies*, *12*, 3-27.

Lande R. (1996). Statistics and partitioning of species diversity, and similarity among multiple communities. *Oikos*, *76*, 5-13.

Lee M.D. (2011). How cognitive modeling can benefit from hierarchical Bayesian models. *Journal of Mathematical Psychology*, *55*, 1-7.

Legge G.E., Klitz T.S., & Tjan B.S. (1997). Mr. Chips: An ideal observer model of reading. *Psychological Review*, *104*, 524-553.

Leitgeb H. & Pettigrew R. (2010a). An objective justification of Bayesianism I: Measuring inaccuracy. *Philosophy of Science*, *77*, 201-235.

Leitgeb H. & Pettigrew R. (2010b). An objective justification of Bayesianism II: The consequences of minimizing inaccuracy. *Philosophy of Science*, *77*, 236-272.

Levinstein B. (2012). Leitgeb and Pettigrew on accuracy and updating. *Philosophy of Science*, *79*, 413-424.

Lewontin R.C. (1972). The apportionment of human diversity. *Evolutionary Biology*, *6*, 381-398.

Lindley D.V. (1956). On a measure of the information provided by an experiment. *Annals of Mathematical Statistics*, *27*, 986-1005.

Markant D. & Gureckis T.M. (2012). Does the utility of information influence sampling behavior? In N. Miyake, D. Peebles, & R.P. Cooper (eds.), *Proceedings of the 34th Annual Conference of the Cognitive Science Society* (pp. 719-724). Austin, TX: Cognitive Science Society.

Masi M. (2005). A step beyond Tsallis and Rényi entropies. *Physics Letters A*, 338, 217-224.

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

Meder B., Mayrhofer R., & Waldmann M.R. (2014). Structure induction in diagnostic causal reasoning. *Psychological Review*, *121*, 277-301.

Muliere P. & Parmigiani G. (1993). Utility and means in the 1930s. *Statistical Science*, *8*, 421-432.

Najemnik J. & Geisler W.S. (2005). Optimal eye movement strategies in visual search. *Nature*, *434*, 387-391.

Najemnik J. & Geisler W.S. (2009). Simple summation rule for optimal fixation selection in visual search. *Vision Research*, *49*, 1286-1294.

Naudts J. (2002). Deformed exponentials and logarithms in generalized thermostatistics. *Physica A*, *316*, 323-334.

Navarro D.J. & Perfors A.F. (2011). Hypothesis generation, sparse categories, and the positive test strategy. *Psychological Review*, *118*, 120-134.

Nelson J.D. (2005). Finding useful questions: On Bayesian diagnosticity, probability, impact, and information gain. *Psychological Review*, *112*, 979-999.

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

Nelson J.D. & Cottrell G.W. (2007). A probabilistic model of eye movements in concept formation. *Neurocomputing*, *70,* 2256-2272.

Nelson J.D., Divjak B., Gudmundsdottir G., Martignon L., & Meder B. (2014). Children's sequential information search is sensitive to environmental probabilities. *Cognition*, *130*, 74-80.

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.

Nelson J.D., Meder B., & Jones M. (2016). On the fine line between "heuristic" and "optimal" sequential question strategies. Submitted.

Nelson J.D., Tenenbaum J.B., & Movellan J.R. (2001). Active inference in concept learning. In J.D. Moore & K. Stenning (eds.), *Proceedings of the 23rd Conference of the Cognitive Science Society* (pp. 692-697). Mahwah (NJ): Erlbaum.

Niiniluoto I. & Tuomela R. (1973). *Theoretical Concepts and Hypothetico-Inductive Inference*. Dordrecht: Reidel.

Oaksford M. & Chater N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review*, *101*, 608-631.

Oaksford M. & Chater N. (2003). Optimal data selection: Revision, review, and re-evaluation. *Psychonomic Bulletin & Review*, *10*, 289-318.

Oaksford M. & Hahn U. (2007). Induction, deduction, and argument strength in human reasoning and argumentation. In A. Feeney & E. Heit (eds.), *Inductive Reasoning: Experimental, Developmental, and Computational Approaches* (pp. 269-301). Cambridge (UK): Cambridge University Press.

Oaksford M. & Chater N. (2007). *Bayesian rationality: The probabilistic approach to human reasoning*. Oxford (UK): Oxford University Press.

Patil G. & Taille C. (1982). Diversity as a concept and its measurement. *Journal of the American Statistical Association*, *77*, 548-561.

Pedersen P. & Wheeler G. (2014). Demystifying dilation. *Erkenntnis*, *79*: 1305-1342.

Pettigrew R. (2013). Epistemic utility and norms for credences. *Philosophy Compass*, *8*, 897-908.

Popper K.R. (1959). *The Logic of Scientific Discovery*. London: Routledge.

Predd J.B., Seiringer R., Lieb E.J., Osherson D., Poor H.V., & Kulkarni S.R. (2009). Probabilistic coherence and proper scoring rules. *IEEE Transactions on Information Theory*, *55*, 4786-4792.

Raiffa H. & Schlaifer R. (1961). *Applied Statistical Decision Theory*. Boston: Clinton Press.

Raileanu L.E. & Stoffel K. (2004). Theoretical comparison between the Gini Index and Information Gain criteria. *Annals of Mathematics and Artificial Intelligence*, *41*, 77-93.

Ramírez-Reyes A., Hernández-Montoya A.R., Herrera-Corral G., & Domínguez-Jiménez I. (2016). Determining the entropic index *q* of Tsallis entropy in images through redundancy. *Entropy*, *18*, 299.

Rao C.R. (2010). Quadratic entropy and analysis of diversity. *Sankhya: The Indian Journal of Statistics*, *72A*, 70-80.

Renninger L.W., Coughlan J., Verghese P., & Malik J. (2005). An information maximization model of eye movements. *Advances in Neural Information Processing Systems*, *17*, 1121-1128.

Rényi A. (1961). On measures of entropy and information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability* I (pp. 547-556). Berkeley (CA): University of California Press.

Ricotta C. (2003). On parametric evenness measures. *Journal of Theoretical Biology*, *222*, 189-197.

Roche W. & Shogenji T. (2014). Dwindling confirmation. *Philosophy of Science*, *81*, 114-137.

Roche W. & Shogenji T. (2016). Information and inaccuracy. *British Journal for the Philosophy of Science*, forthcoming.

Ruggeri A. & Lombrozo T. (2015). Children adapt their questions to achieve efficient search. *Cognition*, *143*, 203-216.

Rusconi P., Marelli M., D'Addario M., Russo S., & Cherubini P. (2014). Evidence evaluation: Measure *Z* corresponds to human utility judgments better than measure *L* and optimal-experimental-design models. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*, 703-723.

Sahoo P.K. & Arora G. (2004). A thresholding method based on two-dimensional Rényi's entropy. *Pattern Recognition*, *37*, 1149-1161.

Savage L.J. (1972). *The foundations of statistics*. New York: Wiley.

Selten R. (1998). Axiomatic characterization of the quadratic scoring rule. *Experimental Economics*, *1*, 43–61.

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

Sharma B. & Mittal D. (1975). New non–additive measures of entropy for discrete probability distributions. *Journal of Mathematical Sciences* (Delhi), *10*, 28-40.

Simpson E.H. (1949). Measurement of diversity. *Nature*, *163*, 688.

Skov R.B. & Sherman S.J. (1986). Information-gathering processes: Diagnosticity, hypothesis-confirmation strategies, and perceived hypothesis confirmation. *Journal of Experimental Psychology*, *22*, 93-121.

Slowiaczek L.M., Klayman J., Sherman S.L., & 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.

Sprenger J. (2016). Foundations for a probabilistic theory of causal strength. See: http://philsci-archive.pitt.edu/11927/1/GradedCausation-v2.pdf.

Stringer S., Borsboom D., & Wagenmakers E.-J. (2011). Bayesian inference for the information gain model. *Behavioral Research Methods*, *43*, 297-309.

Taneja I.J., Pardo L., Morales D., & Menéndez M.L. (1989). On generalized information and divergence measures and their applications: A brief review. *Qüestiió*, *13*, 47-73.

Tentori K., Crupi V., Bonini N., and Osherson D. (2007). Comparison of confirmation measures. *Cognition*, *103*, 107-119.

Tribus M. & McIrvine E.C. (1971). Energy and information. *Scientific American*, *225*, 179-188.

Trope Y. & Bassok M. (1982). Confirmatory and diagnosing strategies in social information gathering. *Journal of Personality and Social Psychology*, *43*, 22-34.

Trope Y. & Bassok M. (1983). Information-gathering strategies in hypothesis testing. *Journal of Experimental and Social Psychology*, *19*, 560-576.

Tsallis C. (2002). Entropic non-extensivity: A possible measure of complexity. *Chaos, Solitons, and Fractals*, *13*, 371-391.

Tsallis C. (2004). What should a statistical mechanics satisfy to reflect nature? *Physica D*, *193*, 3-34.

Tsallis C. (2011). The nonadditive entropy S_q and its applications in physics and elsewhere: Some remarks. *Entropy*, *13*, 1765-1804.

Tsallis, C. (1988). Possible generalization of Boltzmann-Gibbs statistics. *Journal of Statistical Physics*, *52*, 479-487.

Tweeney R.D., Doherty M., & Kleiter G.D. (2010). The pseudodiagnosticity trap: Should participants consider alternative hypotheses? *Thinking & Reasoning*, *16*, 332-345.

Vajda I. & Zvárová J. (2007). On generalized entropies, Bayesian decisions, and statistical diversity. *Kybernetika*, *43*, 675-696.

van der Pyl T. (1978). Propriétés de l'information d'ordre α et de type β. In *Théorie de l'information: Développements récents et applications* (pp. 161-171), Colloques Internationales du CNRS, 276. Paris: Centre National de la Recherche Scientifique.

Wang G. & Jiang M. (2005). Axiomatic characterization of non-linear homomorphic means. *Mathematical Analysis and Applications*, *303*, 350-363.

Wason P. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology*, *12*, 129-140.

Wason P. (1966). Reasoning. In B. Foss (ed.), *New horizons in psychology* (pp. 135-151). Harmonsworth, (UK): Penguin.

Wason P. (1968). Reasoning about a rule. *Quarterly Journal of Experimental Psychology*, *20*, 273-281.

Wu C., Meder B., Filimon F., & Nelson J.D. (in press). Asking better questions: How presentation formats guide information search. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*, 1274-1297.

This is the **Supplementary Materials** file of:

Crupi V., Nelson J.D., Meder B., Cevolani G., and Tentori K., Generalized information theory meets human cognition: Introducing a unified framework to model uncertainty and information search. *Cognitive Science*, 2018.

1. Generalized logarithm and exponential

Consider the Tsallis logarithm, $ln_t(x) = \frac{1}{1-t}\left[x^{(1-t)} - 1\right]$, and note that $1 + (1-t)ln_t(x) = x^{(1-t)}$, therefore $x = [1 + (1-t)ln_t(x)]^{\frac{1}{1-t}}$. This shows that the generalized exponential $e_t^x = [1 + (1-t)x]^{\frac{1}{1-t}}$ just is the inverse function of $ln_t(x)$.

In order to show that the ordinary natural logarithm is recovered from $ln_t(x)$ ($x > 0$) in the limit for $t \to 1$, we posit $x = 1 - y$ and first consider $x \leq 1$, so that $|-y| < 1$. Then we have:

$$\lim_{t\to 1}\{ln_t(x)\} = \lim_{t\to 1}\{ln_t(1-y)\} = \lim_{t\to 1}\left\{\frac{1}{1-t}\left[(1-y)^{(1-t)} - 1\right]\right\}$$

By the binomial expansion of $(1-y)^{(1-t)}$:

$$\begin{aligned}
&\lim_{t\to 1}\left\{\frac{1}{1-t}\left[-1 + \left(1 + (1-t)(-y) + \frac{(1-t)(1-t-1)(-y)^2}{2!} + \frac{(1-t)(1-t-1)(1-t-2)(-y)^3}{3!} + \cdots\right)\right]\right\} \\
&\quad = \lim_{t\to 1}\left\{(-y) + \frac{(-t)(-y)^2}{2!} + \frac{(-t)(-t-1)(-y)^3}{3!} + \cdots\right\} \\
&\quad = \lim_{t\to 1}\left\{(-y) - \frac{t(-y)^2}{2!} + \frac{(t)(t+1)(-y)^3}{3!} - \cdots\right\} \\
&\quad = (-y) - \frac{(-y)^2}{2!} + \frac{2!(-y)^3}{3!} - \cdots \\
&\quad = (-y) - \frac{(-y)^2}{2} + \frac{(-y)^3}{3} - \cdots
\end{aligned}$$

which is the series expansion of $ln(1-y) = \ln(x)$ (recall that $|-y| < 1$). For the case $x > 1$, one can posit $x = 1/(1-y)$, so that again $|-y| < 1$ and compute $\lim_{t\to 1}\left\{\frac{\left(\frac{1}{1-y}\right)^{(1-t)} - 1}{1-t}\right\} = \lim_{t\to 1}\left\{-\frac{1}{t-1}\left[(1-y)^{(t-1)} - 1\right]\right\}$, thus getting the same result from a similar derivation.

Just like the natural logarithm, $ln_t(x)$ is non-negative if $x \geq 1$, because if $t < 1$, then $x^{(1-t)} \geq x^0 = 1$, therefore $\frac{1}{1-t}\left[x^{(1-t)} - 1\right] \geq 0$, while if $t > 1$, then $x^{(1-t)} \leq x^0 = 1$, therefore again $\frac{1}{1-t}\left[x^{(1-t)} - 1\right] \geq 0$. If $0 < x < 1$, $ln_t(x)$ is negative instead, again like the natural logarithm.

To show that the ordinary exponential is recovered from $e_t(x)$ ($x > 0$) in the limit for $t \to 1$, we again rely on the binomial expansion, as follows.

$$\lim_{t\to 1}\{e_t(x)\} = \lim_{t\to 1}\left\{[1 + (1-t)x]^{\frac{1}{1-t}}\right\}$$

$$=\lim_{t\to 1}\left\{1+\left(\frac{1}{1-t}\right)(1-t)x+\left(\frac{1}{1-t}\right)\left(\frac{1}{1-t}-1\right)\frac{\left((1-t)x\right)^2}{2!}+\left(\frac{1}{1-t}\right)\left(\frac{1}{1-t}-1\right)\left(\frac{1}{1-t}-2\right)\frac{\left((1-t)x\right)^3}{3!}+\cdots\right\}$$

$$=\lim_{t\to 1}\left\{1+\left(\frac{1}{1-t}\right)(1-t)x+\left(\frac{1}{1-t}\right)\left(\frac{t}{1-t}\right)\frac{(1-t)^2x^2}{2!}+\left(\frac{1}{1-t}\right)\left(\frac{t}{1-t}\right)\left(\frac{2t-1}{1-t}\right)\frac{(1-t)^3x^3}{3!}+\cdots\right\}$$

$$=1+x+\frac{x^2}{2!}+\frac{x^3}{3!}+\cdots$$

$$=1+\sum_{n=1}^{\infty}\frac{x^n}{n!}=e(x)$$

Just like the ordinary exponential, $e_t(x) \geq 1$ if $x \geq 0$, because if $t < 1$, then one has $[1 + (1 - t)x] \geq 1$ to a positive power $1/(1 - t)$, while if $t > 1$, then one has $[1 + (1 - t)x] \leq 1$ to a negative power $1/(1 - t)$. If $0 < x < 1$, $e_t(x) < 1$ instead, again like the ordinary exponential.

2. Sharma-Mittal entropies

First we will derive the Sharma-Mittal formula from its generalized mean form. We have $g(x) = ln_r e_t(x)$ and $inf(x) = ln_t(x)$. Let us find $g^{-1}(x)$, by solving $y = g(x)$ for x, as follows.

$$y = ln_r e_t(x)$$

$$= ln_r\left[(1+(1-t)x)^{\frac{1}{1-t}}\right]$$

$$= \frac{1}{1-r}\left[(1+(1-t)x)^{\frac{1-r}{1-t}}-1\right]$$

Therefore:

$$1+(1-r)y = [1+(1-t)x]^{\frac{1-r}{1-t}}$$

$$[1+(1-r)y]^{\frac{1}{1-r}} = [1+(1-t)x]^{\frac{1}{1-t}}$$

$$e_r(y) = [1+(1-t)x]^{\frac{1}{1-t}}$$

$$[e_r(y)]^{1-t} = 1+(1-t)x$$

$$\frac{1}{1-t}\{[e_r(y)]^{1-t}-1\} = x$$

$$x = ln_t e_r(y)$$

So $g^{-1}(x) = ln_t e_r(x)$. Now we have all the elements to derive the Sharma-Mittal formula.

$$ent_P^{SM_{(r,t)}}(H) = g^{-1}\left\{\sum_{h_i\in H} p_i\, g\left[inf\left(\frac{1}{p_i}\right)\right]\right\}$$

$$= ln_t e_r\left\{\sum_{h_i\in H} p_i\, ln_r e_t\left[ln_t\left(\frac{1}{p_i}\right)\right]\right\}$$

$$= ln_t e_r\left\{\sum_{h_i\in H} p_i\, ln_r\left(\frac{1}{p_i}\right)\right\}$$

$$= ln_t e_r\left\{\frac{1}{1-r}\sum_{h_i\in H} p_i\left[\left(\frac{1}{p_i}\right)^{1-r}-1\right]\right\}$$

$$= ln_t\left\{1+(1-r)\left[\frac{1}{1-r}\sum_{h_i\in H} p_i\left({p_i}^{(r-1)}-1\right)\right]\right\}^{\frac{1}{1-r}}$$

$$= ln_t\{1 + \sum_{h_i \in H} p_i \, {p_i}^{(r-1)} - \sum_{h_i \in H} p_i\}^{\frac{1}{1-r}}$$

$$= ln_t\{\sum_{h_i \in H} {p_i}^r\}^{\frac{1}{1-r}} = \frac{1}{1-t}\left[\left(\sum_{h_i \in H} {p_i}^r\right)^{\frac{1-t}{1-r}} - 1\right] = \frac{1}{t-1}\left[1 - \left(\sum_{h_i \in H} {p_i}^r\right)^{\frac{t-1}{r-1}}\right]$$

Let us note that $ent^{SM_{(r,t)}}$ satisfies the basic properties of entropy measures. As pointed out above, Tsallis logarithm $ln_t(x)$ is always non-negative if $x \geq 1$, therefore so is $\sum_{h_i \in H} p_i \, ln_r\left(\frac{1}{p_i}\right)$. Moreover, $e_r(x) \geq 1$ if $x \geq 0$ (see above), so $e_r\left\{\sum_{h_i \in H} p_i \, ln_r\left(\frac{1}{p_i}\right)\right\} \geq 1$ and finally $ln_t e_r\left\{\sum_{h_i \in H} p_i \, ln_r\left(\frac{1}{p_i}\right)\right\} \geq 0$. This proves that non-negativity holds for $ent^{SM_{(r,t)}}$. Let us then consider evenness sensitivity. We already know that $ent^{SM_{(r,t)}}$ is non-negative; also, $\sum_{h_i \in H} {p_i}^r = 1$ in case $p_i = 1$ for some i, so that $ent_V^{SM_{(r,t)}}(H) = 0$. As a consequence, for any H and $P(H)$, $ent_P^{SM_{(r,t)}}(H) \geq ent_V^{SM_{(r,t)}}(H) = 0$. In order to complete the proof of evenness sensitivity, we will now study the maximization of $ent^{SM_{(r,t)}}$ by means of so-called Lagrange multipliers. We have to maximize $\sum_{h_i \in H} p_i \, ln_r\left(\frac{1}{p_i}\right) = \frac{1}{1-r}\left[\sum_{h_i \in H}(p_i)^r - 1\right]$, so we study $f(x_1, \ldots, x_n) = \frac{1}{1-r}\left[\sum_{1 \leq i \leq n}(x_i)^r - 1\right]$ under the constraint $\sum_{1 \leq i \leq n}(x_i) = 1$. By the Lagrange multipliers method, we get a system of $n + 1$ equations as follows:

$$\begin{cases} \frac{r}{1-r} {x_1}^{(r-1)} = \lambda \\ \ldots \\ \frac{r}{1-r} {x_n}^{(r-1)} = \lambda \\ x_1 + \cdots + x_n = 1 \end{cases}$$

where $x_1 = \ldots = x_n = 1/n$ is the only solution. This means that $\sum_{h_i \in H} p_i \, ln_r\left(\frac{1}{p_i}\right)$ is either maximized or minimized for the uniform distribution $U(H)$. But actually $ent_U^{SM_{(r,t)}}(H)$ must be a maximum, so that, for any H and $P(H)$, $ent_U^{SM_{(r,t)}}(H) \geq ent_P^{SM_{(r,t)}}(H)$. In fact, $ent_U^{SM_{(r,t)}}(H)$ is *strictly* positive, because $\sum_{h_i \in H} p_i \, ln_r\left(\frac{1}{p_i}\right) = ln_r(n)$ is (recall that $n > 1$). Hence, for any H, $ent_U^{SM_{(r,t)}}(H) > ent_V^{SM_{(r,t)}}(H) = 0$, and evenness sensitivity is shown to hold.

3. Some special cases of the Sharma-Mittal family

Given the above analysis of generalized logarithms and exponentials, we have Rényi (1961) entropy as a special case of the Sharma-Mittal family as follows:

$$ent_P^{SM_{(r,1)}}(H) = ln\left\{e_r\left[\sum_{h_i \in H} p_i \, ln_r\left(\frac{1}{p_i}\right)\right]\right\}$$

$$= ln\left\{\left[1 + (1-r)\frac{1}{1-r}\left(\sum_{h_i \in H} {p_i}^r - 1\right)\right]^{\frac{1}{1-r}}\right\}$$

$$= ln\left\{\left[\sum_{h_i \in H} {p_i}^r\right]^{\frac{1}{1-r}}\right\} = \frac{1}{1-r} ln\left(\sum_{h_i \in H} {p_i}^r\right) = ent_P^{Rényi_{(r)}}(H)$$

For Shannon entropy, in particular, one only needs to note that $ent_P^{SM_{(1,1)}}(H) = ln\left\{e\left[\sum_{h_i \in H} p_i\, ln\left(\frac{1}{p_i}\right)\right]\right\} = \sum_{h_i \in H} p_i\, ln\left(\frac{1}{p_i}\right)$.

For Tsallis (1988) entropy, we have:

$$ent_P^{SM_{(t,t)}}(H) = ln_t e_t\left\{\sum_{h_i \in H} p_i\, ln_t\left(\frac{1}{p_i}\right)\right\} = \sum_{h_i \in H} p_i\, ln_t\left(\frac{1}{p_i}\right) = \frac{1}{t-1}\left[1 - \sum_{h_i \in H} {p_i}^t\right] = ent_P^{Tsallis_{(t)}}(H)$$

For another generalization of Shannon entropy, i.e. Gaussian entropy (Frank, 2004), we have:

$$ent_P^{SM_{(1,t)}}(H) = ln_t e\left\{\sum_{h_i \in H} p_i\, ln\left(\frac{1}{p_i}\right)\right\} = \frac{1}{1-t}\left\{e^{(1-t)\left[\sum_{h_i \in H} p_i ln\left(\frac{1}{p_i}\right)\right]} - 1\right\} = ent_P^{Gauss_{(t)}}(H)$$

The way in which $ent^{Gauss_{(t)}}$ recovers Shannon entropy for $t = 1$ again follows by the behavior of the generalized logarithm, because $ent_P^{Gauss_{(1)}}(H) = ln\left\{e\left[\sum_{h_i \in H} p_i\, ln\left(\frac{1}{p_i}\right)\right]\right\} = \sum_{h_i \in H} p_i\, ln\left(\frac{1}{p_i}\right)$.

For Power entropies, $ent_P^{SM_{(r,2)}}(H) = ent_P^{Power_{(r)}}(H)$ follows immediately from $ent_P^{SM_{(r,t)}}(H) = \frac{1}{t-1}\left[1 - \left(\sum_{h_i \in H} {p_i}^r\right)^{\frac{t-1}{r-1}}\right]$, and the same for Quadratic entropy, i.e., $ent_P^{SM_{(2,2)}}(H) = ent_P^{Quad}(H)$.

If we posit $t = 2 - 1/r$, we have $ent_P^{SM_{(r,2-\frac{1}{r})}}(H) = \frac{r}{r-1}\left[1 - \left(\sum_{h_i \in H} P(h_i)^r\right)^{\frac{1}{r}}\right]$, which happens to be precisely Arimoto's (1971) entropy, under an inconsequential change of parametrization (Arimoto, 1971, used a parameter β to be set to $1/r$ in our notation).

For Effective Number measures (Hill, 1973), we have:

$$ent_P^{SM_{(r,0)}}(H) = \frac{1}{-1}\left[1 - \left(\sum_{h_i \in H} {p_i}^r\right)^{\frac{-1}{r-1}}\right] = \left(\sum_{h_i \in H} {p_i}^r\right)^{\frac{1}{1-r}} - 1 = ent_P^{EN_{(r)}}(H)$$

As a further point concerning Effective Numbers, consider a Sharma-Mittal measure $ent_P^{SM_{(r,t)}}(H) = ln_t e_r\left\{\sum_{h_i \in H} p_i\, ln_r\left(\frac{1}{p_i}\right)\right\}$, for any choice of r and t (both non-negative). We ask what is the number N of equiprobable elements in a partition K such that $ent_P^{SM_{(r,t)}}(H) = ent_U^{SM_{(r,t)}}(K)$. We note that $ent_U^{SM_{(r,t)}}(K) = ln_t e_r\{ln_r(N)\} = ln_t(N)$, thus we posit:

$$ent_P^{SM_{(r,t)}}(H) = ln_t e_r\left\{\sum_{h_i \in H} p_i\, ln_r\left(\frac{1}{p_i}\right)\right\} = ln_t(N)$$

$$N = e_r\left\{\sum_{h_i \in H} p_i\, ln_r\left(\frac{1}{p_i}\right)\right\}$$

$$= \left[1 + (1-r)\frac{1}{1-r}\left(\sum_{h_i \in H} {p_i}^r - 1\right)\right]^{\frac{1}{1-r}}$$

$$= \left(\sum_{h_i \in H} {p_i}^r\right)^{\frac{1}{1-r}} = ent_P^{EN_{(r)}}(H) + 1$$

This shows that, regardless of the degree parameter *t*, for any Sharma-Mittal measure of a specified order *r*, $ent_P^{EN_{(r)}}(H) + 1$ computes the theoretical number *N* of equally probable elements that would be just as entropic as *H* under that measure and given *P*(*H*).

The derivation of the form of $ent^{SM_{(0,t)}}$ is as follows:

$$ent_P^{SM_{(0,t)}}(H) = \frac{1}{t-1}\left[1 - \left(\sum_{h_i \in H} {p_i}^0\right)^{\frac{t-1}{-1}}\right] = \frac{1}{1-t}\left[(n^+)^{(1-t)} - 1\right] = ln_t(n^+)$$

where n^+ is the number of elements in *H* with a non-null probability according to *P*(*H*) (recall that we apply the convention $0^0 = 0$, common in the entropy literature). Hartley entropy, $ent_P^{Hartley}(H) = ln(n^+)$, immediately follows as a special case for $t = 1$, just as Origin entropy, $ent_P^{Origin}(H) = n^+ - 1$, for $t = 0$. $t = 2$ yields $ent_P^{SM_{(0,2)}}(H) = -[(n^+)^{-1} - 1] = \frac{n^+ - 1}{n^+}$.

For the case of infinite order, we posit $p_* = \max_{h_i \in H}(p_i)$ and note the following (*n* is again the overall size of *H*):

$$(p_*)^r \leq \sum_{h_i \in H} {p_i}^r \leq n(p_*)^r$$

Assuming $\frac{t-1}{r-1} \geq 0$ involves no loss of generality in what follows:

$$((p_*)^r)^{\frac{t-1}{r-1}} \leq \left(\sum_{h_i \in H} {p_i}^r\right)^{\frac{t-1}{r-1}} \leq (n(p_*)^r)^{\frac{t-1}{r-1}}$$

$$ln\left[((p_*)^r)^{\frac{t-1}{r-1}}\right] \leq ln\left[\left(\sum_{h_i \in H} {p_i}^r\right)^{\frac{t-1}{r-1}}\right] \leq ln\left[n^{\frac{t-1}{r-1}}((p_*)^r)^{\frac{t-1}{r-1}}\right]$$

$$ln\left[((p_*)^r)^{\frac{t-1}{r-1}}\right] \leq ln\left[\left(\sum_{h_i \in H} {p_i}^r\right)^{\frac{t-1}{r-1}}\right] \leq ln\left(n^{\frac{t-1}{r-1}}\right) + ln\left[((p_*)^r)^{\frac{t-1}{r-1}}\right]$$

$$\lim_{r\to\infty}\left\{ln\left[((p_*)^r)^{\frac{t-1}{r-1}}\right]\right\} \leq \lim_{r\to\infty}\left\{ln\left[\left(\sum_{h_i \in H} {p_i}^r\right)^{\frac{t-1}{r-1}}\right]\right\} \leq \lim_{r\to\infty}\left\{\frac{t-1}{r-1}ln(n)\right\} + \lim_{r\to\infty}\left\{ln\left[((p_*)^r)^{\frac{t-1}{r-1}}\right]\right\}$$

$$\lim_{r\to\infty}\left\{ln\left[((p_*)^r)^{\frac{t-1}{r-1}}\right]\right\} \leq \lim_{r\to\infty}\left\{ln\left[\left(\sum_{h_i \in H} {p_i}^r\right)^{\frac{t-1}{r-1}}\right]\right\} \leq 0 + \lim_{r\to\infty}\left\{ln\left[((p_*)^r)^{\frac{t-1}{r-1}}\right]\right\}$$

Therefore:

$$\lim_{r\to\infty}\left\{ln\left[\left(\sum_{h_i \in H} {p_i}^r\right)^{\frac{t-1}{r-1}}\right]\right\} = \lim_{r\to\infty}\left\{ln\left[((p_*)^r)^{\frac{t-1}{r-1}}\right]\right\} = \lim_{r\to\infty}\left\{ln\left[\left((p_*)^{\frac{r}{r-1}}\right)^{t-1}\right]\right\}$$

The limit for $r \to \infty$ of the argument of the *ln* function exists and is finite: $\lim_{r\to\infty}\left[\left((p_*)^{\frac{r}{r-1}}\right)^{t-1}\right] = {p_*}^{(t-1)}$. For this reason, we can conclude:

$$ln\left\{\lim_{r\to\infty}\left[\left(\sum_{h_i \in H} {p_i}^r\right)^{\frac{t-1}{r-1}}\right]\right\} = ln\left\{\lim_{r\to\infty}\left[\left((p_*)^{\frac{r}{r-1}}\right)^{t-1}\right]\right\}$$

$$\lim_{r\to\infty}\left[\left(\sum_{h_i \in H} {p_i}^r\right)^{\frac{t-1}{r-1}}\right] = {p_*}^{(t-1)}$$

$$\frac{1}{1-t}\left[\lim_{r\to\infty}\left(\sum_{h_i\in H} {p_i}^r\right)^{\frac{t-1}{r-1}}-1\right]=\frac{1}{1-t}\left[{p_*}^{(t-1)}-1\right]$$

$$\lim_{r\to\infty}\left[\frac{1}{1-t}\left(\left(\sum_{h_i\in H} {p_i}^r\right)^{\frac{t-1}{r-1}}-1\right)\right]=\frac{1}{1-t}\left(\left(\frac{1}{p_*}\right)^{1-t}-1\right)$$

$$\lim_{r\to\infty}\left[ent_P^{SM_{(r,t)}}(H)\right]=ln_t\left(\frac{1}{p_*}\right)$$

Error entropy is a special case for $t = 2$, because $ln_2\left(\frac{1}{p_*}\right)=-\left(\left(\frac{1}{p_*}\right)^{-1}-1\right)=1-p_*=1-\max_{h_i\in H}(p_i)$.

$t = 1$ yields $ln\left(\frac{1}{\max_{h_i\in H}(p_i)}\right)$, while, for $t = 0$, one immediately has $\frac{1}{\max_{h_i\in H}(p_i)}-1$.

4. Ordinal equivalence, additivity, and concavity

The ordinal equivalence of any pair of Sharma-Mittal measures $ent^{SM_{(r,t)}}$ and $ent^{SM_{(r,t^*)}}$ with the same order r and different degrees, t and t^*, is easily proven on the basis of the inverse relationship of $ln_t(x)$ and $e_t(x)$. In fact, for any r, t, t^*, and any H and $P(H)$, $ent^{SM_{(r,t)}}$ is a strictly increasing function of $ent^{SM_{(r,t^*)}}$:

$$\begin{aligned} ent_P^{SM_{(r,t)}}(H) &= ln_t e_r\left[\sum_{h_i\in H} p_i\, ln_r\left(\frac{1}{p_i}\right)\right] \\ &= ln_t e_{t^*}\left\{ln_{t^*} e_r\left[\sum_{h_i\in H} p_i\, ln_r\left(\frac{1}{p_i}\right)\right]\right\} \\ &= ln_t e_{t^*}\left\{ent_P^{SM_{(r,t^*)}}(H)\right\} \end{aligned}$$

For degrees t, $t^* \neq 1$, this implies that:

$$ent_P^{SM_{(r,t)}}(H)=\frac{1}{1-t}\left\{\left[1+(1-t^*)ent_P^{SM_{(r,t^*)}}\right]^{\frac{1-t}{1-t^*}}-1\right\}$$

whereas when $t = 1$ and/or $t^* = 1$, the limiting cases of the ordinary exponential and/or natural logarithm apply. This general result is novel in the literature to the best of our knowledge. However, a well-known special case is the relationship between Rényi entropies and the Effective Number measures (see Hill, 1973, p. 428, and Ricotta, 2003, p. 191):

$$\begin{aligned} ent_P^{Rényi_{(r)}}(H) &= ent_P^{SM_{(r,1)}}(H) \\ &= ln\left\{e_0\left[ent_P^{SM_{(r,0)}}(H)\right]\right\} \\ &= ln\left\{1+ent_P^{SM_{(r,0)}}(H)\right\} \\ &= ln\left\{ent_P^{EN_{(r)}}(H)+1\right\} \end{aligned}$$

Another neat illustration involves Power entropy measures and Rényi entropies:

$$ent_P^{Power_{(r)}}(H) = ent_P^{SM_{(r,2)}}(H)$$
$$= ln_2\left\{e\left[ent_P^{SM_{(r,1)}}(H)\right]\right\}$$
$$= 1 - e^{-ent_P^{Rényi_{(r)}}(H)}$$

We will now derive the general additivity rule for Sharma-Mittal entropies concerning independent variables, i.e., when $X \perp_P Y$ holds. To simplify notation, below we will use $\sum(X)$ as a shorthand for $\left(\sum_{x_i \in X} P(x_i)^r\right)^{\frac{t-1}{r-1}}$ (the same for *Y*, and so on) and we will use $ent(X)$ as a shorthand for $ent_P^{SM_{(r,t)}}(X)$ (the same for the expected reduction of entropy, *R*).

$$ent(X) + ent(Y) - (t-1)ent(X)ent(Y)$$
$$= \frac{1}{t-1}[1 - \sum(X)] + \frac{1}{t-1}[1 - \sum(Y)] - (t-1)\,\frac{1}{t-1}[1 - \sum(X)]\frac{1}{t-1}[1 - \sum(Y)]$$
$$= \frac{1}{t-1}[1 - \sum(X)] + \frac{1}{t-1}[1 - \sum(Y)] - \frac{1}{t-1}[1 - \sum(X)][1 - \sum(Y)]$$
$$= \frac{1}{t-1} - \frac{1}{t-1}\sum(X) + \frac{1}{t-1} - \frac{1}{t-1}\sum(Y) - \frac{1}{t-1} + \frac{1}{t-1}\sum(X) + \frac{1}{t-1}\sum(Y) - \frac{1}{t-1}\sum(X)\sum(Y)$$
$$= \frac{1}{t-1} - \frac{1}{t-1}\sum(X)\sum(Y)$$
$$= \frac{1}{t-1} - \frac{1}{t-1}\left(\sum_{x_i \in X} P(x_i)^r\right)^{\frac{t-1}{r-1}}\left(\sum_{y_j \in Y} P(y_j)^r\right)^{\frac{t-1}{r-1}}$$
$$= \frac{1}{t-1}\left[1 - \left(\sum_{x_i \in X} P(x_i)^r \sum_{y_j \in Y} P(y_j)^r\right)^{\frac{t-1}{r-1}}\right]$$
$$= \frac{1}{t-1}\left[1 - \left(\sum_{x_i \in X}\sum_{y_j \in Y}\left(P(x_i)P(y_j)\right)^r\right)^{\frac{t-1}{r-1}}\right]$$
$$= \frac{1}{t-1}\left[1 - \left(\sum_{x_i \in X}\sum_{y_j \in Y} P\left(x_i \cap y_j\right)^r\right)^{\frac{t-1}{r-1}}\right] = ent(X \times Y)$$

This additivity rule in turn governs the relationship between the expected entropy reduction of a test in case it is a perfect (conclusive) experiment and in case it is not. More precisely, it implies that for independent variables *E* and *H*:

$$R(E,E) - R(H \times E, E) = (t-1)ent(H)ent(E)$$

In fact:

$$R(E,E) - R(H \times E, E) = ent(E) - \sum_{e_j \in E}\left[ent\left(E|e_j\right)\right]P\left(e_j\right) - ent(H \times E) + \sum_{e_j \in E}\left[ent\left(H \times E|e_j\right)\right]P\left(e_j\right)$$
$$= ent(E) - 0 - ent(H) - ent(E) + (t-1)ent(H)ent(E) + \sum_{e_j \in E}\left[ent\left(H \times E|e_j\right)\right]P\left(e_j\right)$$
$$= -ent(H) + (t-1)ent(H)ent(E) + \sum_{e_j \in E}\left[ent\left(H|e_j\right) + ent\left(E|e_j\right) - (t-1)ent\left(H|e_j\right)ent\left(E|e_j\right)\right]P\left(e_j\right)$$
$$= -ent(H) + (t-1)ent(H)ent(E) + \sum_{e_j \in E}\left[ent\left(H|e_j\right) + 0 - (t-1)\left(ent\left(H|e_j\right) \times 0\right)\right]P\left(e_j\right)$$
$$= -ent(H) + (t-1)ent(H)ent(E) + ent(H) = (t-1)ent(H)ent(E)$$

Sharma-Mittal measures of expected entropy reduction are also generally additive for a combination of experiments, that is, for any *H,E,F* and $P(H,E,F)$, it holds that $R(H,E\times F)=R(H,E)+R(H,F|E)$. To see this, let us first consider the entropy reduction of a specific datum *e*, $\Delta ent(H,e)=ent(H)-ent(H|e)$. Δent is clearly additive in the following way:

$$\begin{aligned}\Delta ent(H,e\cap f)&=ent(H)-ent(H|e\cap f)\\&=ent(H)-ent(H|e)+ent(H|e)-ent(H|e\cap f)\\&=\Delta ent(H,e)+\Delta ent(H,f|e)\end{aligned}$$

But this pattern carries over to the expected value $R(H,E\times F)$:

$$\begin{aligned}R(H,E\times F)&=\sum_{e_j\in E}\sum_{f_k\in F}\left[\Delta ent\left(H,e_j\cap f_k\right)\right]P\left(e_j\cap f_k\right)\\&=\sum_{e_j\in E}\sum_{f_k\in F}\left[\Delta ent\left(H,e_j\right)+\Delta ent\left(H,f_k|e_j\right)\right]P\left(f_k|e_j\right)P\left(e_j\right)\\&=\sum_{e_j\in E}\sum_{f_k\in F}\left[\Delta ent\left(H,e_j\right)\right]P\left(f_k|e_j\right)P\left(e_j\right)+\sum_{e_j\in E}\sum_{f_k\in F}\left[\Delta ent\left(H,f_k|e_j\right)\right]P\left(f_k|e_j\right)P\left(e_j\right)\\&=\sum_{e_j\in E}\left[\Delta ent\left(H,e_j\right)\right]P\left(e_j\right)+\sum_{e_j\in E}\left\{\sum_{f_k\in F}\left[\Delta ent\left(H,f_k|e_j\right)\right]P\left(f_k|e_j\right)\right\}P\left(e_j\right)\\&=R(H,E)+\sum_{e_j\in E}\left\{R\left(H,F|e_j\right)\right\}P\left(e_j\right)\\&=R(H,E)+R(H,F|E)\end{aligned}$$

This result is novel in the literature to the best of our knowledge.

Finally, we will show that, for any *H,E*, and *P*(*H,E*), $R(H,E)\geq 0$ if and only if *ent* is concave. Let $\mathbb{E}_P(v)$ be the expected value of a variable *v* for some probability distribution *P* = {p_1, …, p_m}, i.e. $\mathbb{E}_P(v)=\sum_{i=1}^{m}v_i\,p_i$. According to a multivariate version of Jensen's inequality, $g(x_1,\dots,x_n)$ is a concave function if and only if g of the expected values of its arguments is greater than (or equal to) the expected value of g, that is:

$$g[\mathbb{E}_P(x_1),\dots,\mathbb{E}_P(x_n)]\geq\mathbb{E}_P[g(x_1,\dots,x_n)]$$

Now we set $g(x_1,\dots,x_n)=ent(H|e)$ and we posit that $\mathbb{E}_P(x)$ be computed on the basis of *P*(*E*), i.e. $\mathbb{E}_P(v)=\sum_{e_j\in E}v_iP\left(e_j\right)$. Assuming that *ent* is concave, we have:

$$\frac{1}{t-1}\left[1-\left(\sum_{h_i\in H}\left(\sum_{e_j\in E}P(h_i|e_j)P(e_j)\right)^r\right)^{\frac{t-1}{r-1}}\right]\geq\sum_{e_j\in E}\left[\frac{1}{t-1}\left(1-\left(\sum_{h_i\in H}P(h_i|e_j)^r\right)^{\frac{t-1}{r-1}}\right)\right]P\left(e_j\right)$$

$$\frac{1}{t-1}\left(1-\left(\sum_{h_i\in H}P(h_i)^r\right)^{\frac{t-1}{r-1}}\right)-\sum_{e_j\in E}\left[\frac{1}{t-1}\left(1-\left(\sum_{h_i\in H}P(h_i|e_j)^r\right)^{\frac{t-1}{r-1}}\right)\right]P\left(e_j\right)\geq 0$$

$$\sum_{e_j\in E}\left[\frac{1}{t-1}\left(1-\left(\sum_{h_i\in H}P(h_i)^r\right)^{\frac{t-1}{r-1}}\right)-\frac{1}{t-1}\left(1-\left(\sum_{h_i\in H}P(h_i|e_j)^r\right)^{\frac{t-1}{r-1}}\right)\right]P\left(e_j\right)\geq 0$$

$$\sum_{e_j\in E}\Delta ent(H,e)P\left(e_j\right)\geq 0$$

$$R(H,E)\geq 0$$

5. Expected entropy reduction in the Person Game

To analyze the expected entropy reduction of one binary query in the person game, we will posit $H = \{h_1, \ldots, h_n\}$ (the set of possible guesses as to who the randomly selected character is) and $E = \{e, \overline{e}\}$ (the yes/no answers to a question such as "does the selected character have blue eyes?"; recall that "$\overline{e}$" denotes the complement or the negation of e). The joint probability distribution $P(H,E)$ is defined as follows: $P(h_i \cap e) = 1/n$ in case $i \leq k$ (with $1 \leq k < n$) and $P(h_i \cap e) = 0$ otherwise; $P(h_i \cap \overline{e}) = 0$ in case $i \leq k$ and $P(h_i \cap \overline{e}) = 1/n$ otherwise. This implies that $P(h_i) = 1/n$ for each i (all guesses are initially equiprobable), $P(h_i|e) = 1/k$ for each $i \leq k$ (the posterior given e is a uniform distribution over k elements of H), and $P(h_i|\overline{e}) = 1/(n-k)$ for each $i > k$ (the posterior given $\overline{e}$ is a uniform distribution over $n-k$ elements of H). Moreover, $P(e) = k/n$. Given the general fact that $ent_U^{SM_{(r,t)}}(H) = ln_t(n)$, we have:

$$R_P^{SM_{(r,t)}}(H,E) = [ln_t(n) - ln_t(k)]P(e) + [ln_t(n) - ln_t(n-k)]P(\overline{e})$$

Algebraic manipulations yield:

$$R_P^{SM_{(r,t)}}(H,E) = \frac{n^{1-t}}{1-t}[1 - (P(e)^{2-t} + P(\overline{e})^{2-t})]$$

In the special case $t = 2$, one then has $R_P^{SM_{(r,2)}}(H,E) = \frac{1}{n}$, so that the expected usefulness of query E is constant, regardless of the value of $P(e)$. More generally, however, the first derivative of $R_P^{SM_{(r,t)}}(H,E)$ is

$$\frac{n^{1-t}}{1-t}[(2-t)P(\overline{e})^{1-t} - (2-t)P(e)^{1-t}]$$

which equals zero for $P(e) = P(\overline{e})$, so that $R_P^{SM_{(r,t)}}(H,E)$ has a maximum or a minimum for $P(e) = ½$. The second derivative, in turn, is:

$$n^{1-t}(t-2)[P(e)^{-t} + P(\overline{e})^{-t}]$$

which is strictly positive [negative] in case t is strictly higher [lower] than 2. So, in the person game, $R_P^{SM_{(r,t)}}(H,E)$ is a strictly concave function of $P(e)$ when $t < 2$, and $P(e) = ½$ is then a maximum. When $t > 2$, on the contrary, $R_P^{SM_{(r,t)}}(H,E)$ is a strictly convex function of $P(e)$, and $P(e) = ½$ is then a minimum. This general result is novel in the literature to the best of our knowledge.