Comparing Probabilistic Measures of Explanatory Power

Jonah N. Schupbach*

March 26, 2010

Abstract

Recently, in attempting to account for explanatory reasoning in probabilistic terms, Bayesians have proposed several measures of the degree to which a hypothesis explains a given set of facts. These candidate measures of "explanatory power" are shown to have interesting *normative* interpretations and consequences. What has not yet been investigated, however, is whether any of these measures are also *descriptive* of people's actual explanatory judgments. Here, I present my own experimental work investigating this question. I argue that one measure in particular is an accurate descriptor of explanatory judgments. Then, I discuss some interesting implications of this result for both the epistemology and the psychology of explanatory reasoning.

*University of Pittsburgh, Department of History & Philosophy of Science; 1017 Cathedral of Learning; Pittsburgh, Pennsylvania, USA; 15260.

1 Explanatory Reasoning and Bayesianism

Humans are, it seems, constantly making use of explanatory considerations when they reason. Mundane examples abound: I notice that the books on my shelf are disarranged, and I reason that my two year old has been playing in the office; I observe a large crowd waiting at the bus stop and I hypothesize that the 10:00 bus has not yet been by. In both of these cases, there is seen to be some reason in a hypothesis's favor precisely because of its ability to explain some observed fact. Explanatory considerations not only pervade everyday human reasoning, but they are also ubiquitous in intellectual practices such as science and medicine: scientists reason to the existence of a hitherto unobserved planet given their observations of the motion of Uranus; a doctor diagnoses a patient with the measles after considering that patient's symptoms.

In recent years, Bayesians have turned their attention to the epistemology of such "explanatory reasoning." Okasha (2000, pp. 702-706) proposes necessary probabilistic conditions for one hypothesis's being a *better explanation* than another. Lipton (2004, ch. 7; see also his 2001), on the other hand, suggests various possible probabilistic renderings of his notions of explanatory *loveliness* and *likeliness*. The culmination of this work, however, has been the attempt to define a measure of the degree to which a particular hypothesis h is able to explain some given set of information d – i.e., h's degree of "explanatory power" relative to d. Interestingly, this project was undertaken decades ago by Popper (1959) and by Good (1960, 1968). Very recently, however, McGrew (2003), Glass (2007), and Schupbach and Sprenger (2010) all propose specific such measures in their work. All of these corresponding measures, along with some other plausible candidate measures, are shown in **Table 1**.

$E_D(d,h) = Pr(d\|h) - Pr(d)$	
$E_C(d,h) = Pr(d\|h) - Pr(d\|\neg h)$	
$E_P(d,h) = \dfrac{Pr(d\|h) - Pr(d)}{Pr(d\|h) + Pr(d)}$	(Popper, 1959)
$E_M(d,h) = ln\left[\dfrac{Pr(d\|h)}{Pr(d)}\right]$	(Good, 1960; McGrew, 2003)
$E_G(d,h) = \dfrac{Pr(d \wedge h)}{Pr(d \vee h)} = \left[\dfrac{1}{Pr(h\|d)} + \dfrac{1}{Pr(d\|h)} - 1\right]^{-1}$	(Glass, 2007)
$\mathscr{E}(d,h) = \dfrac{Pr(h\|d) - Pr(h\|\neg d)}{Pr(h\|d) + Pr(h\|\neg d)}$	(Schupbach and Sprenger, 2010)

Table 1. Candidate Measures of Explanatory Power.

Measures E_M and $\mathscr{E}$ are both related to Bayesian measures of confirmation; in fact, these measures just *are* the confirmation measures of Keynes (1921) and Ke-

meny and Oppenheim (1952) respectively but with each reference to the evidence e in these original measures replaced with a reference to h and a reference to the explanandum d substituted for each reference to h in the confirmation measures. Measures E_D and E_C have been built in the same way from two other confirmation measures – due to (Eells, 1982) and (Christensen, 1999) respectively – and added to the list of measures to consider here.

Popper's measure of explanatory power E_P is closely linked in two different ways to two other measures on this list. It is, first of all, a renormalization of E_D as seen by the fact that the numerator of E_P just is E_D. But more importantly, E_P is *ordinally equivalent* to measure E_M. In more detail, this means that, for any h, h', d, and d', $E_P(d,h) > (=,<) E_P(d',h')$ if and only if $E_M(d,h) > (=,<) E_M(d',h')$.[1] Thus, E_P and E_M always impose the same *ordinal* relations on judgments of explanatory power.

Measure E_G is unique insofar as it is the only proposed *coherence*-theoretic measure of explanatory power. Glass (2007) argues that the explanatory power of h relative to explanandum d just is measured by the degree to which h coheres with d. Glass thus analyzes explanatory power in terms of his favorite Bayesian measure of coherence, which was first proposed by himself (Glass, 2002) and independently by Olsson (2002).

Each of the measures shown in **Table 1** does its part to clarify the precise formal relations that attain between hypotheses and data when the former are explanatory of the latter. These measures thus enable us to ask and answer substantive questions about the epistemic value of explanatory power. As a matter of fact, all four of the measures of explanatory power that have been put forward in the literature have also been used to defend explanatory reasoning as having normative merit. Schupbach and Sprenger (2010, p. 20) show, for instance, that "when all else is equal, the probability of an explanatory hypothesis in the light of some evidence is directly proportional to that hypothesis's ability to explain that evidence." McGrew (2003, p. 558) provides a defense of a similar ceteris paribus theorem in terms of his measure. Glass (2007, p. 294) argues that, according to his account, "good explanations will be probable explanations and so someone who reasons [explanatorily] will tend to make probable inferences." And Popper (1959, p. 401) shows that the amount of explanatory power that a hypothesis has relative to some evidence is positively related to the degree of "corroboration" that the former receives from the latter.

These measures thus attempt to provide *normative* accounts of explanatory power and explanatory reasoning. First and foremost, they tell us how we ought to think of the concept of explanatory power. Additionally, they each assert that, un-

[1] *Proof:* Dividing the numerator and denominator of E_P through by $Pr(d)$ gives the following:

$$E_P(d,h) = \frac{Pr(d|h)/Pr(d) - 1}{Pr(d|h)/Pr(d) + 1}$$

And, for values of $r \in [0,\infty)$, which is the range of the ratio $Pr(d|h)/Pr(d)$, $f(r) = (r-1)/(r+1)$ is a monotonically increasing function of r. Thus, E_P is an increasing function of the ratio $Pr(d|h)/Pr(d)$. But, of course, $E_M = ln[Pr(d|h)/Pr(d)]$ is also a monotonically increasing function of the ratio $Pr(d|h)/Pr(d)$. □

der certain conditions, explanatory considerations do guide us to hypotheses which are more probable. Thus, they tell us that we ought to reason explanatorily under such conditions. These measures unquestionably thus have interesting normative interpretations and consequences.

What has not yet been investigated regarding these measures is the separate question of whether any of them are also *descriptive* of people's actual explanatory judgments. Of course, the normative bearings of these measures does not imply their descriptive accuracy. It may well be that a measure accurately represents the way people generally *ought* to think about explanatory power and that, if they think about it in this way, then they *ought* to reason in favor of good explanations; and it may simultaneously be true that people do not do as they epistemically ought. Alternatively, if some candidate normative measure also doubles as a good descriptor of people's explanatory judgments, then we have the makings of an interesting defense of human explanatory reasoning. The issue then is whether people actually think about explanatory power in the way that these epistemologists have said that they should.

But the descriptive question also has important bearing for the normative analyses themselves. Here, the question is whether any of the formal accounts fit with the concept of explanatory power as it is generally used. If all of the measures diverge widely from people's actual explanatory intuitions, then it may be that people do not understand explanatory power in the way that they should; however, it might more plausibly be the case that the analyses are just wrong. On the other hand, if any particular candidate measure fits well with such intuitions, then this not only reflects nicely on everyday human intuitions, but it also provides some support for the general accuracy of that particular measure.

This paper empirically investigates the descriptive question. As such, and in light of the above, it holds interest both to philosophers interested in the epistemology of explanatory reasoning and to psychologists interested in human reasoning.

2 Experiment: Comparing the Descriptive Merits of the Measures

In this section, I summarize my own recent experimental research investigating the descriptive question. The overarching goal of this project was to test and compare the relative descriptive merits of the aforementioned candidate measures of explanatory power. In order to do this, I used an experimental design based closely upon a chance-setup previously applied by Phillips and Edwards (1966) and more recently by Tentori et al. (2007) in their comparison of various Bayesian measures of *confirmation*.

2.1 Materials and Procedure

In this experiment, participants were asked to judge how well various hypotheses explain certain sets of data. These judgments were elicited during an individual

Urn	Number of Black Balls	Number of White Balls
A	30	10
B	15	25

Table 2. Respective Contents of Urns A and B.

interview involving a probabilistic scenario of black and white balls being drawn without replacement from one of two possible urns. During the interview, participants were first presented with two opaque urns, and then informed of their respective contents. The urns were composed of black and white balls as specified in **Table 2**. At this point, participants were also given a visual representation of the urns' contents, which they were free to refer to throughout the experiment.

The decision of which urn to use throughout the remainder of the interview was next decided via an actual flip of a fair coin. Participants saw that the coin flip determined our choice of urn; however, whether the chosen urn was A or B was left hidden. The experiment then proceeded with a series of ten random drawings without replacement from the chosen urn. These drawings and the corresponding results were performed in full view of the participants. Additionally, balls that were the results of prior drawings were lined in front of the participants in the order in which they had been withdrawn; thus, at any time in the interview, participants could refer to all of the results up to that point. Throughout each interview, the coin flip and drawings were truly chance events so that which urn was used and which balls were withdrawn differed between participants. Participants were faced with six tasks after each individual drawing.

Task 1. Participants were first asked to make a mark on an "impact scale" representing the degree to which "the hypothesis that urn A was chosen $[(H_A)]$ explains the results from all of the drawings so far." Each impact scale was printed on a strip of paper and consisted of a dotted line with arrows pointing out of either end. The following five descriptive labels were spaced evenly from left to right over the line (with the line extending in both directions beyond the labels):

- This hypothesis is an **extremely poor** explanation of the results collected so far
- This hypothesis is a **poor** explanation of the results collected so far
- This hypothesis is **neither a good nor a poor** explanation of the results collected so far
- This hypothesis is a **good** explanation of the results collected so far
- This hypothesis is an **extremely good** explanation of the results collected so far

Fresh copies of the scale were used for each of the ten drawings, and all of a participant's previously marked judgments were organized in his or her view to

refer to if desired. On a given impact scale, the marked distance from the neutral point was used to quantify judged degrees of explanatory power. Upon receiving the impact scale, participants were told that the scale was intended to be continuous and that distances would matter to how their responses were recorded.

Task 2. Next, participants were asked to repeat the first task but this time with regard to the hypothesis that urn B was chosen (H_B). Ultimately then, participants were asked to make 20 judgments of explanatory power throughout the experiment (10 pertaining to H_A, and 10 pertaining to H_B).

Tasks 3 and 4. In tasks 3 through 6, participants estimated various relevant probabilities. For the first two of these tasks, participants were faced with the following two questions (in the questions listed below, n was set to the number of balls that had been drawn at that point in the interview):

- Considering the color of the first n balls, what now is the probability that the urn selected is A?
- Considering the color of the first n balls, what now is the probability that the urn selected is B?

Participants were instructed that their answers could be written in whatever format they preferred (decimals, fractions, or percentages); however, they had to sum either to 1 (if they chose to write decimals or fractions) or 100%.

Tasks 5 and 6. For the final two tasks performed with each drawing, participants were asked the following two questions:

- Assuming that the selected urn is A, what at this point was the probability of drawing a ball of this color?
- Assuming that the selected urn is B, what at this point was the probability of drawing a ball of this color?

Again, participants were instructed that their answers could be written in whatever format they preferred; for these two questions, it was pointed out that there was no need for the two answers to sum to 1 (or 100%).

Tasks 3 and 4 were used to assess participants' judgments about the probabilities of the respective hypotheses conditional upon all of the "evidence" received from the drawings. That is, in the n'th round of the interview, each participant's response to task 3 was interpreted as that person's subjective probability for H_A conditional upon the n results of all of the drawings up to that point: $Pr_{Subj}(H_A|d_1 \wedge d_2 \wedge ... \wedge d_n)$. Similarly, participants' responses to task 4 were taken to provide values for $Pr_{Subj}(H_B|d_1 \wedge d_2 \wedge ... \wedge d_n)$.

On the other hand, tasks 5 and 6 assessed participant judgments about the probabilities of the latest result conditional upon the respective hypotheses and all preceding results. That is, in the n'th round of the interview, each participant's response to task 5 was interpreted as that person's subjective probability for the result of the n'th drawing conditional upon H_A and upon the $n-1$ preceding results: $Pr_{Subj}(d_n|H_A \wedge d_1 \wedge d_2 \wedge ... \wedge d_{n-1})$. Similarly, responses to task 6 were taken to provide values for $Pr_{Subj}(d_n|H_B \wedge d_1 \wedge d_2 \wedge ... \wedge d_{n-1})$.

Given the chance nature and the quantitative details of this experimental design, the following, corresponding objective probabilities were calculated for each drawing in each interview: $Pr_{Obj}(H_A|d_1 \wedge d_2 \wedge ... \wedge d_n)$, $Pr_{Obj}(H_B|d_1 \wedge d_2 \wedge ... \wedge d_n)$, $Pr_{Obj}(d_n|H_A \wedge d_1 \wedge d_2 \wedge ... \wedge d_{n-1})$, and $Pr_{Obj}(d_n|H_B \wedge d_1 \wedge d_2 \wedge ... \wedge d_{n-1})$.

These probabilities (collected in both their subjective and objective varieties) were sufficient to derive corresponding degrees of explanatory power for H_A and H_B (relative to the various sets of data) from all of the candidate measures in **Table 1**. In this way, this experiment elicited a host of participant judgments about explanatory power along with the same number of corresponding results derived from each measure (first using subjective probabilities, and then also derived using the objective probabilities).

2.2 Participants

26 undergraduate students from the University of Pittsburgh participated in this study in exchange for $10 each. The average age of the participants was 20 years. Among the participants, there were 14 men and 12 women.

3 Results

3.1 Preparing the Measures for Comparison

In order to compare the descriptive accuracies of the measures, we rely first upon the measure of the Euclidean distance between participant judgments and the "theoretical results" derived from each particular candidate measure of explanatory power. This distance (in n-dimensional space) between a set of n judged degrees of explanatory power and a corresponding set of n theoretical degrees is given by the following equation – where $J(d_i, h_i)$ represents participant judgments of the degree to which h_i explains d_i, and E stands in for any particular candidate measure of explanatory power:

$$\mathrm{d}(J, E) = \sqrt{\sum_{i=1}^{n} (J(d_i, h_i) - E(d_i, h_i))^2}$$

That is, the Euclidean distance d between participant judgments J and the theoretical results derived from E is given by summing the squares of the "residuals" (the differences between each judged value and theoretical value) and then calculating that sum's square root. The lower the value of d, the closer E is to participant judgments J.

This choice of measure requires defense especially in light of Tentori et al.'s (2007) similar study comparing the desciptive merits of various confirmation measures. Tentori et al. rely primarily on a Pearson correlation test to decide which confirmation measure "corresponds most closely to judged evidential impact" (p. 115). The experimental design applied here is based upon that used by Tentori et al.; furthermore, the nature of our experimental results and our aims in analyzing

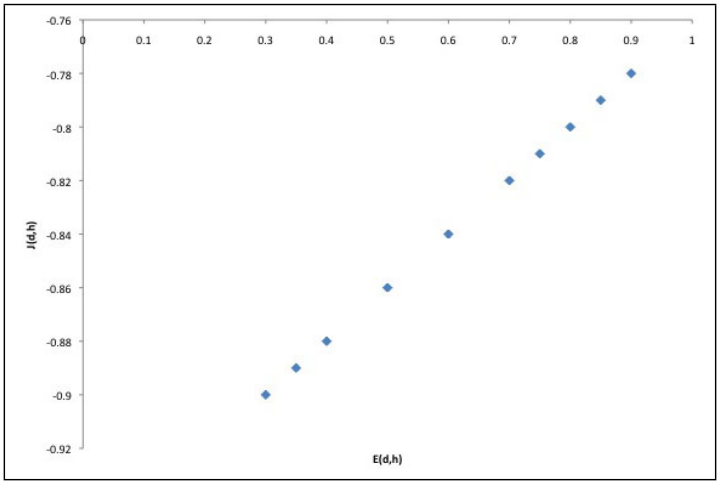

Figure 1. $E(d,h)$ perfectly correlated with $J(d,h)$ but giving vastly different values.

them are closely related. So why this change in how we proceed with the analysis? The answer is that a correlation test will inevitably fall short of the sort that we want to utilize in our comparison.[2]

Pearson's correlation test measures the degree of linear dependence that holds between two variables. As such, it provides a powerful tool for showing the degree to which the values of one variable can be predicted as a linear function of another variable (whose values are known). More specific to our context, if J and a particular set of theoretical results derived from a measure E are shown to be highly correlated, then this would constitute evidence that E could be used as a predictor of people's explanatory judgments. This would surely be an interesting finding. However, a measure of the degree of explanatory power which hopes to be descriptively valid claims to be more than merely capable of being made into a good predictor of such judgments; indeed, the most descriptively accurate measure will be the one whose results actually correspond most closely to judged degrees of explanatory power themselves. This notion of proximity is just what is measured by a distance measure such as d. On the other hand, the concept of correlation can diverge significantly from this notion. Indeed, two variables can be *perfectly* correlated even while having vastly different corresponding values (as in **Figure 1**). Thus, in order to test the full descriptive merits or our measures, we opt for a distance measure.

Our choice to use a distance measure does, however, lead to a new complication. In order for us to compare the distances between each of our measures and actual human judgments, we must first and foremost make sure that all derived and judged degrees of explanatory power are on the same scale. Participants' marked

[2]This is not intended to be a criticism of Tentori et al.'s use of this test. Pearson's correlation test *does* seem to be well-suited for their purposes but not so for our own given the differences between our respective concepts of interest.

judgments are easily placed onto a $[-1,1]$ scale with the extreme left point of the dotted line on the impact scale representing -1, the center point 0, and the extreme right point 1. Moreover, E_D, E_C, E_P, and $\mathscr{E}$ are all on the same $[-1,1]$ scale with interpretations corresponding to the labels provided with the impact scale.[3] Measure E_G has a finite range of $[0,1]$; thus, it can quickly be placed on the same scale as the other measures if we consider the rescaled verion, $E_{G'}(d,h) = 2 \times E_G(d,h) - 1$. On the other hand, rescaling measure E_M proves to be a much more complicated affair.

Measure E_M agrees with our other candidate measures of explanatory power on its neutral point. That is, (substituting the rescaled $E_{G'}$ for E_G) all of the measures agree that the value 0 is to be interpreted as the neutral point at which h is "explanatorily irrelevant" to d. However, while all other candidate measures are finite, E_M has the range $(-\infty,\infty)$. In order to measure the distance between the results provided by such a measure and a set of judged degrees on a finite scale then, E_M must be "rescaled" down to a finite scale.

This can be done by feeding the results of E_M into any function that has all of the real numbers as its domain and the real numbers from -1 to 1 as its range. More specifically, such a function minimally ought to satisfy the following conditions of adequacy in order to rescale E_M appropriately:

Finite Boundedness. The function F must have all of the real numbers as its domain and the set of real numbers from -1 to 1 as its range: $F : \mathbb{R} \to [-1,1]$.

Monotonicity. F must be monotonically increasing: $\forall(x)(F'(x) \geq 0)$.

Neutrality. $F(x) = 0$ if and only if $x = 0$.

Asymptotic Behavior. The rate at which $F(x)$ increases or decreases approaches 0 for the limiting points: $\lim_{x\to\infty} F'(x) = 0$ and $\lim_{x\to-\infty} F'(x) = 0$.

These conditions of adequacy are all easily motivated as requirements for our function F. **Finite Boundedness** has already been discussed above. **Monotonicity** is required given that we want E_M's ordinal judgments to be preserved under the transformation affected by F; i.e., for any two pairs $\langle d_i, h_i\rangle$ and $\langle d_j, h_j\rangle$, $E_M(d_i,h_i) < (=,>) E_M(d_j,h_j)$ if and only if $F(E_M(d_i,h_i)) < (=,>) F(E_M(d_j,h_j))$. We also want F to preserve the fact that E_M is normalized around 0 with this value representing explanatory irrelevance; thus, we require **Neutrality**. Finally, as values of E_M increase (or decrease) without bound, corresponding degrees of explanatory power become less distinguishable and their differences less meaningful. Accordingly, we enforce the **Asymptotic Behavior** requirement for F.

Hartmann and Sprenger (2010) introduce (for purposes entirely different than our own) a family of functions that, with a minor modification,[4] elegantly satisfies our conditions of adequacy. This family is defined by the following equation:

[3] For example, $\mathscr{E}(d,h) = 1$ is interpreted as the point at which h provides a full explanation of d, $\mathscr{E}(d,h) = 0$ the point at which h is judged to be explanatorily irrelevant to d, and $\mathscr{E}(d,h) = -1$ the point at which h provides a full explanation of $\neg d$ (Schupbach and Sprenger, 2010, p. 5).

[4] For their purposes, Hartmann and Sprenger introduce the measure $L_\alpha(x) = 1 - e^{-\frac{1}{2\alpha^2}x^2}$ defined over domain $\mathbb{R}^{\geq 0}$. Here, we need a function defined generally over $\mathbb{R}$ that is monotonically decreasing as $x \to -\infty$. The modied measure L_α achieves these purposes.

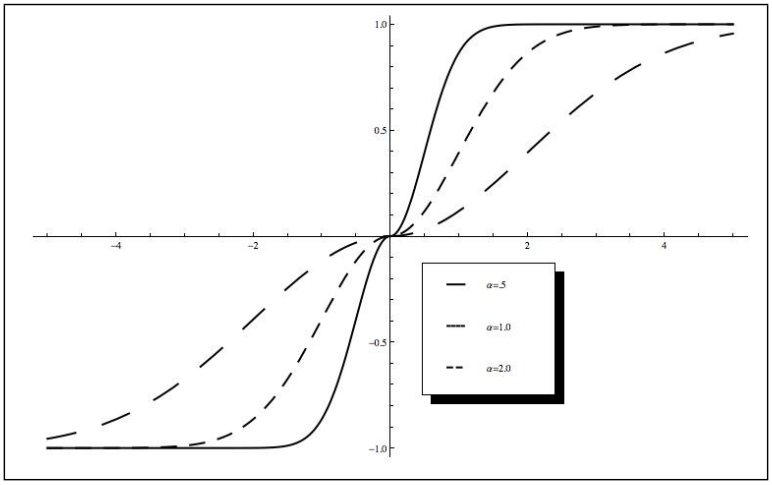

Figure 2. Three members of the L_α family.

$$L_\alpha(x) = \begin{cases} 1 - e^{-\frac{1}{2\alpha^2}x^2} & \text{if } x \geq 0 \\ -1 + e^{-\frac{1}{2\alpha^2}x^2} & \text{if } x < 0 \end{cases}$$

L_α provides us with any number of functional rescalings of E_M depending upon the parameter α (three members of the L_α family are pictured in **Figure 2**). This fact constitutes a significant advantage for E_M when it comes to testing and comparing our measures' proximities to participant judgments. To measure E_M's distance from participant judgments, we can essentially evaluate a wide range of the members of L_α and then choose that member of L_α that is closest. In this sense, E_M is much more flexible and thereby has an a priori advantage over the other measures.

3.2 Comparing the Measures

We are now prepared to compare the descriptive merits of our various candidate measures of explanatory power. We first apply the Euclidean distance measure d to the results derived from each of our candidate measures of explanatory power via participants' subjective probabilities. Results (over 260 judgments for each hypothesis) are displayed in **Table 3**. These results change somewhat if we now apply the measure d to the results derived from the candidate measures using *objective* probabilities. Results are displayed in **Table 4**.

These tables reveal several interesting findings. First, the last row in each table provides the distance between participant judgments and the corresponding posterior probabilities (rescaled to $[-1, 1]$) that the urn chosen is A (column 2) or is B (column 3) in light of d. These probabilities come remarkably close to participant judgments of explanatory power. In particular, the *subjective* posterior probabilities come closest to participant judgments about H_A while these probabilities are second only to $\mathscr{E}$ in proximity to judgments about H_B. These results might suggest

Measure	Distance from $J(d, H_A)$	Distance from $J(d, H_B)$
E_D	8.563	7.726
E_C	8.455	7.755
E_P	5.437	6.144
$E_{G'}$	15.048	14.940
$\mathscr{E}$	5.597	5.211
$L_{.5}$	6.928	8.197
L_1	5.935	6.233
L_2	6.376	6.024
$2 \times Pr_{Subj}(H_A[H_B]\|d) - 1$	5.132	5.404

Table 3. Distances between participant judgments and measures (subjective probabilities).

Measure	Distance from $J(d, H_A)$	Distance from $J(d, H_B)$
E_D	8.497	7.596
E_C	8.356	7.555
E_P	5.392	5.952
$E_{G'}$	14.520	14.887
$\mathscr{E}$	5.617	6.218
$L_{.5}$	6.217	7.190
L_1	6.118	6.312
L_2	6.502	6.218
$2 \times Pr_{Obj}(H_A[H_B]\|d) - 1$	6.587	8.318

Table 4. Distances between participant judgments and measures (objective probabilities).

either of the following two hypotheses. First, it could be that participants confuse the concepts of explanatory power and probability; in this case, when asked to judge how well a hypothesis explains some set of data, participants tend to read the question as asking for their judgment of how probable the hypothesis is in light of that data. Alternatively, participants may have distinct concepts of explanatory power and posterior probability that are nevertheless closely related (as the normative implications of our candidate measures would imply). In either case, we would expect participant judgments of one of these concepts to track judgments of the other. We will have more to say below about the relative merits of these two hypotheses.

These tables also reveal $E_{G'}$ to be a uniquely *bad* descriptor of participants' explanatory judgments. As mentioned previously, $E_{G'}$ also happens to be unique insofar as it is the only formal attempt to analyze explanatory power in terms of coherence. Consequently, the descriptive prospects for a coherence-theoretic analysis of explanatory power look bleak. At least with regards to the notion of coherence that Glass (2007) has in mind when he introduces E_G, this study suggests that participants are *not* thinking about how well hypotheses cohere with d when making judgments about how well they explain d.

Third, the tables show that, whether we use subjective or objective probabilities in our derivations, measures E_P, $\mathscr{E}$, and various rescalings of E_M consistently come the closest of all of the considered candidate measures of explanatory power to participant judgments. This observation immediately leads to a further question insofar as we want a *full* comparison of the descriptive merits of our measures. Recall that measure E_M has the advantage of corresponding to any number of rescaled measures L_α. While E_P and $\mathscr{E}$ look as though they generally come closer to participant judgments than $\mathrm{L}_{.5}$, L_1, or L_2, it may be that *some other* rescaling of E_M nonetheless outperforms E_P and $\mathscr{E}$. To investigate this possibility, we must run a more careful analysis of the L_α family to get a closer estimate of which of its members comes the closest to participant judgments. Then, we can compare that member to E_P and $\mathscr{E}$. **Figures 3** and **4** summarize the results of such an analysis. Looking at these figures, we can see that the overall Euclidean distance (over all 520 participant judgments – 260 pertaining to H_A and 260 pertaining to H_B) corresponding to members of L_α never dips below that for E_P or for $\mathscr{E}$. We can also now estimate which member of the L_α family is the closest competitor to E_P and $\mathscr{E}$. When using subjective probabilities, we estimate the best performing member of L_α to be $\mathrm{L}_{1.25}$; when using objective probabilities, we choose $\mathrm{L}_{.9}$.

In light of the preceding discussion, at least two important questions still remain. First, do participants simply conflate the notions of explanatory power and posterior probability, or do they take these to be distinct, albeit closely related to one another? Second, $\mathscr{E}$ and E_P are generally shown by d to be closer to participant judgments than the other measures. Yet, one might still wonder what degree of confidence we can have in this conclusion given our data and whether we can run a distinct comparison between these two measures which will single one out as providing the best fit with participants' judgments.

As it turns out, we can shed light on both of these questions by performing a more sophisticated comparison of our measures. Specifically, we calculate and

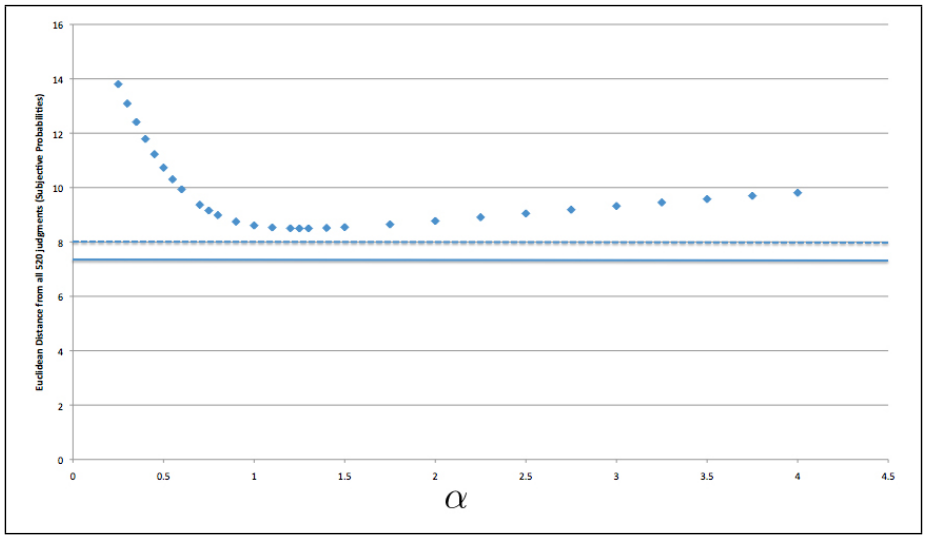

Figure 3. Distances of members of L_α versus that of E_P (dotted line) and $\mathscr{E}$ (solid line) – calculated using subjective probabilities.

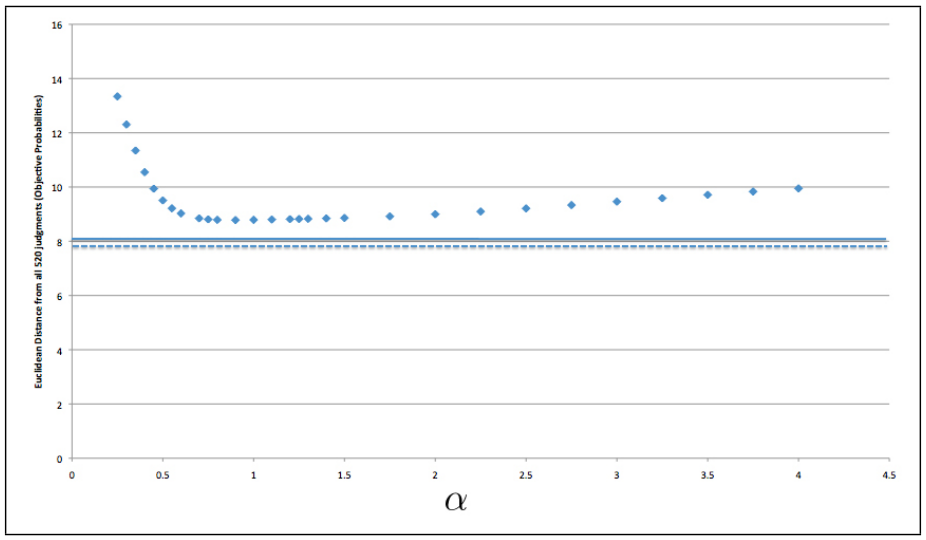

Figure 4. Distances of members of L_α versus that of E_P (dotted line) and $\mathscr{E}$ (solid line) – calculated using objective probabilities.

Measure	Mean Residual	σ	Measure	Mean Residual	σ
E_D	-.098	.497	E_D	-.095	.491
E_C	-.095	.495	E_C	-.095	.485
E_P	.077	.352	E_P	.081	.343
$E_{G'}$	.749	.551	$E_{G'}$	.728	.550
$\mathrm{L}_{1.25}$	.112	.356	$\mathrm{L}_{.9}$	.134	.362
$2 \times Pr_{Subj} - 1$	-.095	.313	$2 \times Pr_{Obj} - 1$	-.095	.456
$\mathscr{E}$	-.015	.335	$\mathscr{E}$	.071	.361

Table 5. Sample statistics (using subjective probabilities on left, objective probabilities on right).

	E_D	E_C	E_P	$E_{G'}$	$\mathrm{L}_{1.25}$	$2 \times Pr_{Subj} - 1$
$\mathscr{E}$	$t = 5.915$	$t = 6.000$	$t = -7.543$	$t = -49.702$	$t = -11.783$	$t = 7.833$
	$p < .001$	$p < .001$	$p < .001$	$p < .001$	$p < .001$	$p < .001$
	E_D	E_C	E_P	$E_{G'}$	$\mathrm{L}_{.9}$	$2 \times Pr_{Obj} - 1$
$\mathscr{E}$	$t = 8.092$	$t = 8.628$	$t = -2.963$	$t = -32.441$	$t = -11.896$	$t = 13.074$
	$p < .001$	$p < .001$	$p < .005$	$p < .001$	$p < .001$	$p < .001$

Table 6. Comparison of $\mathscr{E}$ with other measures (theoretical results calculated using subjective probabilities for top half and objective probabilities for bottom half). Each cell reports the results of a paired t-test between residuals obtained with $\mathscr{E}$ and those obtained with the measure in the associated column. For each test, $N = 520$, corresponding to the total number of participant judgments.

compare the means of the residuals (i.e., $J(d_i, h_i) - E(d_i, h_i)$) between the theoretical results provided by each candidate measure and participant judgments. These mean residuals (and corresponding standard deviations) are displayed in **Table 5**. As this table shows, $\mathscr{E}$'s results have the mean residual that comes closest to the ideal value of 0, and this is true whether we are using subjective or objective probabilities to derive our theoretical values. Furthermore, **Table 6** reveals results from a series of paired t-tests collectively showing that the differences between $\mathscr{E}$'s mean residual and those corresponding to the other measures are all quite significant. Note, in particular, that $\mathscr{E}$'s mean residual is significantly closer to 0 than that of E_P and $\mathrm{L}_{1.25}$ (when using subjective probabilities) and E_P and $\mathrm{L}_{.9}$ (when using objective probabilities). Accordingly, from our experimental data, we can now conclude that $\mathscr{E}$ comes significantly closer to participant judgments than any other candidate measure (including any functional rescaling of E_M).

Importantly, $\mathscr{E}$ not only does comparatively well in this regard, but it also does remarkably well on its own. In particular, the mean residual between $\mathscr{E}$'s results (calculated using subjective probabilities) and participant judgments (**Table 5**) does not differ significantly from 0 ($N = 520$, $t = -1.012$, $p = .312$). This result

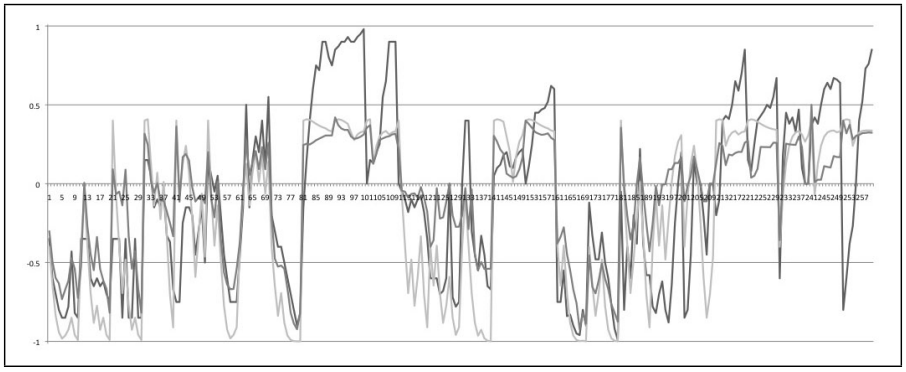

Figure 5. Participant judgments about H_A (darkest line) plotted with values derived from $\mathscr{E}$ using subjective probabilities and objective probabilities (lightest line).

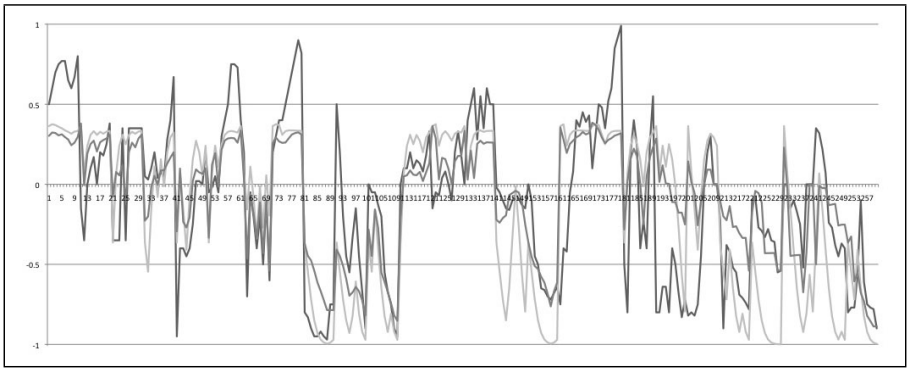

Figure 6. Participant judgments about H_B (darkest line) plotted with values derived from $\mathscr{E}$ using subjective probabilities and objective probabilities (lightest line).

does not hold true for any other measure; in all other cases (using either subjective or objective probabilities) a measure's mean residual differs significantly from the ideal value 0 (for all of these comparisons, $p < .0001$). **Figures 5** and **6** give visual representations of the fit between $\mathscr{E}$ and participant judgments.

We may now return to the question of whether participants are simply conflating the notions of explanatory power and posterior probability. If this were true, then we would expect the mean residual corresponding to the posterior probability to be very close to 0. This should particularly prove true in cases where the residuals represent the differences in a participant's judged degree of explanatory power and that same participant's own stated subjective posterior probability. In the subjective and objective cases, however, the mean residual is $-.095$. This means that, on average (over 520 data points), participants judge explanatory power to be significantly lower than the corresponding posterior probability. Thus, our experimental data provides us with evidence that, even while intuitions about explana-

tory power are linked closely to judgments of posterior probability (as evidenced by their close Euclidean distance), these notions remain conceptually distinct.

4 Discussion

This experiment has important implications both for the epistemology and psychology of explanatory reasoning. Regarding the former, Schupbach and Sprenger (2010) argue that measure $\mathscr{E}$ corresponds most closely to our notion of explanatory power because this measure alone satisfies several intuitive conditions of adequacy for such an analysis. This paper augments that case for $\mathscr{E}$ with empirical evidence suggesting that this measure also does the best at predicting people's explanatory judgments in general. The case for $\mathscr{E}$ as our most accurate formal analysis of explanatory power thus looks to be strong indeed.

Regarding this experiment's implications for psychology, the results here support the claim that $\mathscr{E}$ is a useful predictor of human explanatory judgments. At worst then, $\mathscr{E}$ provides psychologists with a useful, but merely instrumental theory of explanatory reasoning. On the other hand, at best, $\mathscr{E}$ may lend insight into some of the mental heuristics, and ultimately the cognitive mechanisms, that people use in making judgments pertaining to explanation and probability. To take one example, from these experiments, we see clear signs that participants' judgments of explanatory power are closely aligned with, though distinct from, their judgments of probability. This finding accords well with the normative implications of $\mathscr{E}$. It also suggests that people may well use their intuitions about how well a hypothesis explains data as a heuristic when trying to gauge that hypothesis's probability in light of that data. As Peter Lipton (2004, p. 121) repeatedly quips: "explanatory loveliness is a guide to judgments of likeliness."

Last, and of interest to both philosophers and psychologists, these experiments form the basis of a normative defense of everyday human explanatory reasoning. If, as suggested here, people's explanatory judgments fit well with the formal analysis $\mathscr{E}$, then their judgments will tend to benefit from this measure's positive, normative implications. Consequently, given that, according to $\mathscr{E}$, the best explanation of some set of facts d must also be the most probable hypothesis in the light of d (under certain formal conditions), and given that people's explanatory judgments tend to agree with the results of $\mathscr{E}$, then (given certain corresponding conditions) people will tend to choose more probable hypotheses when they reason explanatorily. The specific conditions that must attain in order for this to hold true can be spelled out quite precisely (thanks to the clarity inherent in the formal analysis $\mathscr{E}$). Such further work, moreover, can give us a better sense of where we should expect explanatory reasoning to break down and lead humans astray given $\mathscr{E}$. Such a project is beyond the scope of the present paper, however, and thus constitutes one possible route for further research in the light of this study.

References

Christensen, D. (1999, September). Measuring Confirmation. *Journal of Philosophy 96*(9), 437–461.

Eells, E. (1982). *Rational Decision and Causality*. Cambridge: Cambridge University Press.

Glass, D. H. (2002). Coherence, Explanation, and Bayesian Networks. In M. O'Neill, R. F. E. Sutcliffe, C. Ryan, M. Eaton, and N. J. L. Griffith (Eds.), *Artificial Intelligence and Cognitive Science*, pp. 177–182. New York: Springer-Verlag.

Glass, D. H. (2007). Coherence Measures and Inference to the Best Explanation. *Synthese 157*, 275–296.

Good, I. J. (1960). Weight of Evidence, Corroboration, Explanatory Power, Information and the Utility of Experiments. *Journal of the Royal Statistical Society. Series B (Methodological) 22*(2), 319–331.

Good, I. J. (1968, August). Corroboration, Explanation, Evolving Probability, Simplicity and a Sharpened Razor. *British Journal for the Philosophy of Science 19*(2), 123–143.

Hartmann, S. and J. Sprenger (Forthcoming, 2010). The Weight of Competence under a Realistic Loss Function. *The Logic Journal of the IGPL*.

Kemeny, J. G. and P. Oppenheim (1952). Degree of Factual Support. *Philosophy of Science 19*, 307–324.

Keynes, J. M. (1921). *A Treatise on Probability*. London: Macmillan.

Lipton, P. (2001). Is Explanation a Guide to Inference? A Reply to Wesley C. Salmon. In G. Hon and S. S. Rakover (Eds.), *Explanation: Theoretical Approaches and Applications*, pp. 93–120. Dordrecht: Kluwer Academic.

Lipton, P. (2004). *Inference to the Best Explanation* (2nd ed.). New York, NY: Routledge. 1st ed. published in 1991.

McGrew, T. (2003). Confirmation, Heuristics, and Explanatory Reasoning. *British Journal for the Philosophy of Science 54*, 553–567.

Okasha, S. (2000). Van Fraassen's Critique of Inference to the Best Explanation. *Studies in the History and Philosophy of Science 31*(4), 691–710.

Olsson, E. J. (2002). What is the Problem of Coherence and Truth? *Journal of Philosophy 94*, 246–272.

Phillips, L. D. and W. Edwards (1966). Conservatism in a Simple Probability Inference Task. *Journal of Experimental Psychology 72*(3), 346–354.

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

Schupbach, J. N. and J. Sprenger (Under Review, 2010). The Logic of Explanatory Power.

Tentori, K., V. Crupi, N. Bonini, and D. Osherson (2007). Comparison of Confirmation Measures. *Cognition 103*, 107–119.