Determinants of judgments of explanatory power: Credibility, Generality, and Statistical Relevance

Matteo Colombo, Leandra Bucher, & Jan Sprenger

Abstract Explanation is a central concept in human psychology. Drawing upon philosophical theories of explanation, psychologists have recently begun to examine the relationship between explanation, probability and causality. Our study advances this growing literature in the intersection of psychology and philosophy of science by systematically investigating how judgments of explanatory power are affected by (i) the prior credibility of a potential explanation, (ii) the causal framing used to describe the explanation, (iii) the generalizability of the explanation, and (iv) its statistical relevance for the evidence. Collectively, the results of our five experiments support the hypothesis that the prior credibility of a causal explanation plays a central role in explanatory reasoning: first, because of the presence of strong main effects on judgments of explanatory power, and second, because of the gate-keeping role it has for other factors. Highly credible explanations were not susceptible to causal framing effects. Instead, highly credible hypotheses were sensitive to the effects of factors which are usually considered relevant from a normative point of view: the generalizability of an explanation, and its statistical relevance for the evidence. These results advance current literature in the philosophy and psychology of explanation in three ways. First, they yield a more nuanced understanding of the determinants of judgments of explanatory power, and the interaction between these factors. Second, they illuminate the close relationship between prior beliefs and explanatory power. Third, they clarify the relationship between abductive and probabilistic reasoning.

Keywords

Explanation Prior credibility Causal framing Generality Statistical relevance

Explanation is a central concept in human psychology. It supports a wide array of cognitive functions, including reasoning, categorization, learning, inference, and decision-making (Lombrozo, 2006; Keil & Wilson, 2000; Keil, 2006). When presented with an explanation of why a certain event occurred, how a certain mechanism works, or why people behave the way they do, both scientists and laypeople have strong intuitions about what counts as a good explanation. Yet, more than sixty years after philosophers of science began to elucidate the nature of explanation (Craik, 1943; Hempel & Oppenheim, 1948; Hempel, 1965; Carnap, 1966; Salmon, 1971), the determinants of judgments of explanatory power remain unclear.

In this paper, we present five experiments on factors that may affect judgments of explanatory power. Motivated by a large body of theoretical results in epistemology and philosophy of science, as well as by a growing amount of empirical work in cognitive psychology (for respective surveys see Woodward, 2014; Lombrozo, 2012), we examined how judgments of explanatory power are affected by (i) the prior credibility of a potential explanation, (ii) the causal framing used to describe the explanation, (iii) the generalizability of the explanation, and (iv) its statistical relevance for the evidence.

Specifically, we set out to test four hypotheses. First, we hypothesized that the prior credibility of a causal explanation predicts judgments of explanatory power. Throughout all five experiments, we manipulated the prior credibility of different explanations, and examined the effects of this manipulation on explanatory judgments. We also wanted to understand how low and high prior credibility interacted with other possible psychological determinants of explanatory power.

Our focus on the prior credibility of causal explanation was motivated by the fact that most philosophical and psychological analyses of explanatory power agree that powerful explanations provide information about credible causal relationships (Salmon, 1984; Lewis, 1986; Dowe, 2000). Credible causal information facilitates the manipulation and control of nature (Pearl, 2000; Woodward, 2003; Strevens, 2008) and plays distinctive roles in human psychology (Lombrozo,

2011; Sloman & Lagnado, 2015). For example, credible causal information guides categorization (Carey, 1985, 201ff; Murphy & Medin, 1985; Lombrozo, 2009), supports inductive inference and learning (Holyoak & Cheng, 2011; Legare & Lombrozo, 2014; Walker et al. 2014), and calibrates metacognitive strategies involved in problem-solving (Chi et al, 1994; Aleven & Koedinger, 2002).

Our second, related hypothesis was that presenting an explanatory hypothesis in causal terms predicts judgments of its explanatory power. Thus, we wanted to find out whether people's explanatory judgments are sensitive to causal framing effects.

The importance of this issue should be clear in the light of the fact that magazines and newspapers very often, even when it's not warranted, describe scientific explanations in terms of causal language (e.g., 'Processed meat causes cancer' or 'Economic recession leads to xenophobic violence') with the aim of capturing readers' attention and boosting their sense of understanding (Entmann 1993; Scheufele & Scheufele 2010). By combining prior credibility and causal framing as predictors of judgments of explanatory power, Experiment 1 and 2 examined the impact of causality on the explanatory power of scientific hypotheses.

With Experiment 3, we tested the hypothesis that the generalizability (or scope) of a hypothesis determines its explanatory power. Specifically, we isolated the effects of generalizability on judgments´ explanatory power and its interaction with the prior credibility of an explanation, while controlling for causal framing and statistical relevance.

While the generalizability of scientific results is an obvious epistemic virtue that figures in the evidential assessments made by scientists, its relation to explanatory power is less clear. Previous psychological findings about the role of generalizability in explanatory reasoning are mixed. Read & Marcus-Newhall (1993) found that generalizability predicts explanatory judgments. Preston & Epley (2005) showed that hypotheses that apply to a wide range of observations are judged as more valuable. However, these studies involved no uncertainty about whether or not a causal effect was actually observed (cf., Khemlani, Sussman, & Oppenheimer, 2011). So, whether or not generalizability is a robust determinant of explanatory judgment remains unclear.

With Experiments 4 and 5, we tested our fourth and final hypothesis: that the statistical relevance of a hypothesis for a body of observed evidence is another key determinant of judgments of explanatory power.

According to several philosophers, the power of an explanation is manifest in the amount of statistical information that an *explanans* H provides about an *explanandum* E. In particular, it has to be the case that Prob (E|H&S) > Prob (E|S) (Jeffrey, 1969; Greeno, 1970; Salmon, 1970). Suppose, for example, that Jones has strep infection, and his doctor gives him penicillin. After Jones has taken penicillin, he recovers within one week. When we explain why Jones recovered, we usually cite statistically relevant facts, such as the different recovery rates among treated and untreated patients.

Developing this idea, several research groups have put forward probabilistic measures of explanatory power (McGrew, 2003; Schupbach & Sprenger, 2011; Crupi & Tentori, 2012). Their approach is that a hypothesis is the more explanatorily powerful the less surprising it makes the observed evidence. Results from experimental psychology confirm this insight. Schupbach (2011) provided evidence that Schupbach & Sprenger's (2011) probabilistic measure is an accurate predictor of people's explanatory judgments in abstract reasoning problems. Colombo, Postma, & Sprenger (2016) found that explanatory judgments about everyday situations are strongly affected by changes in statistical relevance. Despite these results, it remains unclear how statistical relevance interacts with other, probabilistic and non-probabilistic factors to determine explanatory power, in particular the prior credibility of an explanation. Experiment 4 and 5 examine the influence of statistical relevance in this regard, both for numerical and for visual representation of the statistical information.

Clarifying the respective impact of prior credibility and statistical relevance on judgments of explanatory power matters to another central topic in the philosophy and psychology of explanation: *abductive reasoning* (Lipton, 2004; Douven, 2011; Schupbach, 2016). When people engage in abductive reasoning, they rely on explanatory considerations to justify the conclusion that

a certain hypothesis is true. Specifically, people often infer the truth of that hypothesis H1 from a pool of candidate hypotheses H1, H2, …, Hn, that best explains available evidence E (Peirce, 1931; Thagard, 1989; Douven, 2011). However, whether "best explains" consists in high statistical relevance, generalizability, provision of a plausible cause or some other explanatory virtue remains controversial (van Fraassen, 1989; Okasha, 2000; Lipton, 2001, 2004; Glymour, 2014; Douven & Schupbach, 2015).

In summary, bringing together different strands of research from philosophy and psychology, our study asks: How do the credibility, causal framing, statistical relevance and generalizability of a hypothesis influence judgments of explanatory power? The pattern of our experimental findings supports the hypothesis that the prior credibility of a causal explanation plays a central role in explanatory reasoning: first, because of the presence of strong main effects on judgments of explanatory power, and second, because of the gate-keeping role it has for other factors. Highly credible explanations were not susceptible to causal framing effects, which may lead astray explanatory judgment. Instead, highly credible hypotheses were sensitive to the effects of factors which are usually considered relevant from a normative point of view: the generalizability of an explanation, and its statistical relevance for the study´s results.

These results advance current literature in the philosophy and psychology of explanation in three ways. First, our results yield a more nuanced understanding of the determinants of judgments of explanatory power, and the interaction between these factors. Second, they illuminate the close relationship between prior beliefs and explanatory power. Third, they clarify the relationship between abductive and probabilistic reasoning.

Overview of the experiments and pre-tests

We conducted five experiments, where we systematically examined the influence of the possible determinants of explanatory judgment: prior credibility, causal framing, generalizability, and statistical relevance. To warrant the validity of the experimental material, we conducted a series of

pre-studies, where participants evaluated different levels of causal framing, credibility, and generalizability. Materials which corresponded to high, low, and neutral levels of these three factors were implemented in the vignettes of our five experiments, either as independent variables or as control variables. Material evaluation and main experiments were both conducted online on Amazon Mechanical Turk, utilizing the Qualtrics Survey Software. We only allowed workers with an approval rate > 95% and with a number of HITs approved > 5000 to submit responses. Instructions and material were presented in English. None of the participants took part in more than one experiment.

Causal Framing

In a pre-study, a sample of N = 44 participants (mean age 30.5 years, *SD* = 7.3, 28 male) from America (n = 27) and other countries rated eight brief statements, expressing relations between X and Y of the type “X co-occurs with Y”; “X is associated with Y”, and so on (see Appendix A for the complete list of statements). The statements were presented in an individually randomized order to the participants; only one statement was visible at a time; and going back to previous statements was not possible. The participants judged how strongly they agreed or disagreed that a certain statement expressed a causal relation between X and Y. Judgments were collected on a 7-point scale with options: "I strongly disagree" (-3), "I disagree", "I slightly disagree", "I neither agree nor disagree" (0), "I slightly agree", "I agree", "I strongly agree" (3). Based on participants’ ratings, we selected three types of statements for our main experiments: statements with a neutral causal framing (“X co-occurs with Y”), with a weak causal framing (“X is associated with Y"), and with a strong causal framing ("X leads to Y" and "X causes Y") (Table 1).

Table 1: Wordings that were perceived to express weak, neutral, and strong causal framing of the relationship between an explanans (X) and an explanandum (Y)

Causal Framing	**Framing of the hypothesis**
Weak	X is associated with Y
Neutral	X co-occurs with Y
Strong	X causes Y[1]
Strong	X leads to Y

Prior Credibility

We identified the prior credibility of different hypotheses by asking a new sample of N = 42 participants (mean age 30.7 years, *SD* = 7.5, 16 male) from America (n = 29) and other countries to rate a list of 24 statements (Appendix A). Participants judged how strongly they disagreed or agreed that a certain hypothesis was credible. For all hypotheses, we used the phrasing "... co-occurs with..." to avoid the influence of causal framing. Based on participants' ratings (see Appendix A), we selected four statements to use in our main experiments: two were highly credible, the other two were highly incredible (Table 2).

Table 2: The four hypotheses rated as least credible and as most credible.

Credibility	**Hypothesis**
Low	Eating pizza co-occurs with immunity to flu.
Low	Drinking apple juice co-occurs with anorexia.
High	Well-being co-occurs with frequent smiling
High	Consuming anabolic steroids co-occurs with physical strength.

[1]According to the ratings observed in the pre-study, "X causes Y" and "Y leads to Y" express causal relations to an equal extent.

Generalizability

We conducted a pre-study in order to determine how the description of the sample used in a scientific study influenced the perceived generalizability of the study's results. This pre-study included two questionnaires, which were administered to two different groups of participants. One questionnaire presented descriptions of the samples used in scientific studies, which varied with regard to the *number* of people involved. The other questionnaire presented sample descriptions that varied with regard to the type of people in the sample. The statements were presented in an individually randomized order to the participants. Only one statement was visible at a time, and going back to previous statements was not possible.

Forty-two participants (mean age 33.5 years, *SD* = 10.8, 27 male) from America (n = 38) and other countries were presented with a list of six brief statements about a sample of a particular number of participants, e.g. "The study investigates five people"; "The study investigates 500 people" (see Appendix A for the complete list of items). We found that the perceived generalizability of a study increased with the number of people in the sample of the study.

A new group of N = 41 participants (mean age 33.0 years, *SD* = 9.7, 26 male) from America (n = 36) and other countries was presented with a list of nine brief statements about samples of particular types of people, e.g. "The study investigates a group of people who sit in a park"; "The study investigates a group of people who work at a university" (see Appendix A for the complete list of items). However, focusing on the *number* instead of the *type* of people in the sample allowed for a neater distinction between narrowly and widely generalizable results. Therefore we characterized generalizability as a function of the number of participants in the main vignettes of the experiment (see Table 3).

Table 3: Ratings of the generalizability of studies in the pre-tests, dependent of the number of people in the sample.

Generalizability	**Description**
Narrow	The study investigates five people.
Medium	The study investigates 240 people.
Wide	The study investigates 10,000 people.

Vignettes of the Main Experiment

All experiments were performed, using a 2x2 (within-subject) design with explanatory power as dependent variable and prior credibility of the hypothesis being one of the independent variables. The other independent variable was either causal framing, generalizability, or statistical relevance of the reported research study.

Participants were presented with four short reports about fictitious research studies. Two of these reports involved highly credible hypotheses, the other two reports involved incredible hypotheses. Two reports showed a high level of the other independent variable (causal framing/generalizability/statistical relevance), while the other two reports showed a low level of that variable. To account for the possible influence of the content of a particular report, the allocation of low and high levels of that variable was counterbalanced to the credibility conditions across the items, leading to two versions of each experiment.

Each vignette in our experiments followed the same format, including a headline and five sentences. The headline stated the hypothesis, the first sentence introduced the study, the second sentence described the sample size, the third sentence reported the results of the study, and the fourth sentence reported factors controlled by the researchers. The final sentence presented a brief conclusion, essentially restating the hypothesis.

We now present a sample vignette for a study that investigates the link between anabolic steroids and physical strength. For details of the vignettes in the individual experiments, see Appendices B-D.

Consuming anabolic steroids leads to physical strength

A recent study by university researchers investigated the link between consuming anabolic steroids and physical strength. The researchers studied 240 persons. The level of physical strength was higher among participants who regularly consumed anabolic steroids than among the participants who did not regularly consume anabolic steroids. Family health history, age, and sex, which were controlled by the researchers, could not explain these results. The study therefore supports the hypothesis that consuming anabolic steroids leads to physical strength.

In all experiments, we varied the level of prior credibility of a hypothesis. In Experiment 1 and 2, we also varied the causal framing and interchanged "leads to" with "causes" and "is associated with", while we kept generalizability at its control value (N=240) and did not provide information about statistical relevance. In Experiment 3, we varied the sample size (=generalizability) and controlled for causal framing by using the predicate "co-occurs with" in the headline and the conclusion. Finally, in Experiment 4 and 5, we varied the levels of statistical relevance (=the frequency of a causal effect in the treatment and in the control group) while controlling for causal framing ("X co-occurs with Y") and generalizability (N=240). Participants were asked to rate our dependent variable: the explanatory power of the stated hypothesis for the results of the study.

Experiment 1 and 2. Credibility x Causal Framing

Two-hundred-three participants (mean age 34.7 years, *SD* = 10.5; 121 male) from America (n= 130), India (n = 67) and other countries completed Experiment 1 for a small monetary payment. A new sample of two-hundred-eight participants (mean age 34.56 years, *SD* = 9.97; 124 male) from America (n = 154), India (n = 43), and other countries completed Experiment 2 for a small monetary payment.

Design and Material

In both experiments, participants were presented with four short reports about fictitious research studies along the lines of the above vignette. Across vignettes, we manipulated the causal framing of the relationship between hypothesis and evidence as well as the choice of the hypothesis (credible vs. incredible). Generalizability was controlled for by setting it to its medium value (240 participants). Two of the four reports involved highly credible hypotheses, the other two reports involved incredible hypotheses. Similarly, two of these reports used weak causal framing (Experiment 1 and 2: "X is associated with Y") while the other two reports used strong causal framing (Experiment 1: "X leads to Y", Experiment 2: "X causes Y"). In other words, Experiment 1 used implicit causal language and Experiment 2 used explicit causal language, while the experiments were, for the rest, identical with respect to design, materials, and procedure.

To account for the possible influence of the content of a particular report, we counterbalanced the allocation of weak and strong causal framing conditions to the credibility conditions across the items, and created two versions of the experiments: Version A and B (see Appendix B for details). The order of reports was individually randomized for each participant.

Procedure

Participants judged each report in terms of the explanatory power of the hypothesis it described. Specifically, participants considered the statement: "The researchers' hypothesis explains the results of the study", and expressed their judgments on a 7-point scale with the extremes (-3) "I strongly disagree" and (3) "I strongly agree", and the center pole (0) "I neither disagree nor agree".

Analysis and Results

Separate two-way ANOVAs were calculated for Experiment 1 and 2, with the factors Credibility (low, high) and Causal Framing (weak, strong). ANOVA of Experiment 1 (implicit causal language) revealed a main effect of Credibility, F (1, 202) = 84.5; $p < .001$; $\eta_{part}^2 = 0.30$. There was no main effect of Causal Framing ($p = .37$), and no interaction ($p = .08$). Pair-wise comparisons showed that incredible hypotheses were rated significantly lower than credible hypotheses, independently of the value of Causal Framing (incredible hypotheses: $M = 0.26$; $SEM = 0.10$;

credible hypotheses: M = 1.14; SEM = 0.09; t-test: $t(202) = -9.2$; $p < 0.001$; d = 0.67). See Figure 1. The results of Experiment 1 therefore indicate that the prior credibility of a hypothesis was a strong predictor of judgments of explanatory power. Instead, framing a hypothesis with implicit causal language did not have effects on explanatory judgment.

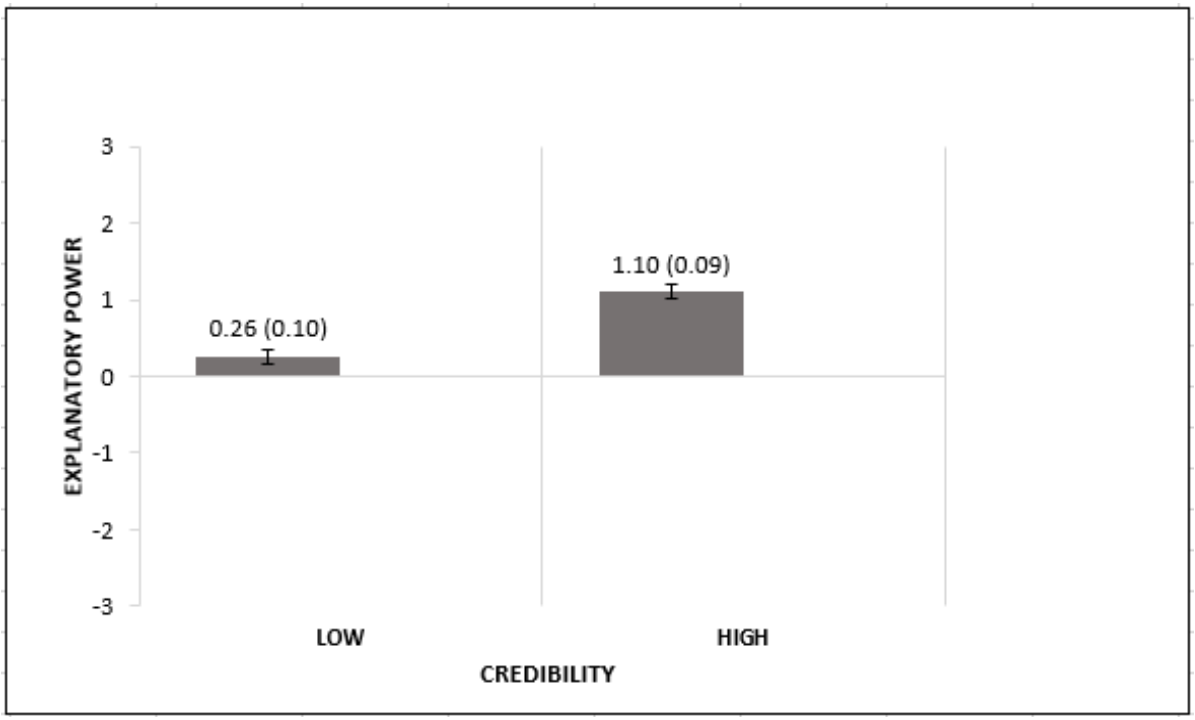

Figure 1. The graph shows explanatory power ratings for credible and incredible statements in Experiment 1. Ratings were significantly higher for credible as opposed to incredible statements. Error bars show standard errors of the mean and are also expressed numerically, in parentheses next to the mean value.

ANOVA of Experiment 2 (explicit causal language) revealed main effects of Credibility (F (1, 207) = 286.9; p <.001; η_{part}^2 = 0.58) and Causal Framing, F (1, 207) = 31.0; p <.001; η_{part}^2 = 0.13, as well as a significant interaction Credibility x Causal Framing, F (1, 207) = 37.6; p <.001; η_{part}^2 = 0.15. Figure 2 shows the effect sizes and the interaction between both factors as well as the relevant descriptives.

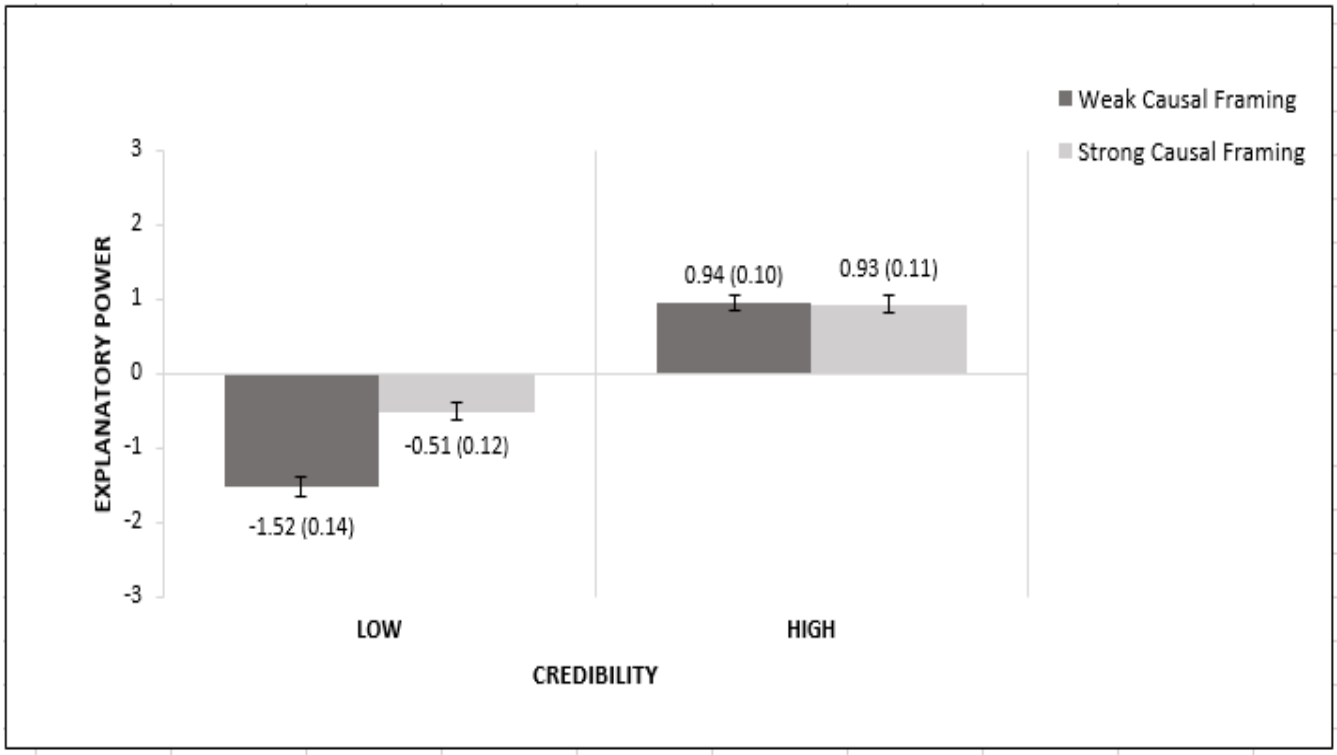

Figure 2. The graph shows how explanatory power ratings vary with regard to Credibility and Causal Framing (as presented in Experiment 2). Ratings were significantly higher for statements with high compared to low Credibility, and for statements with strong compared to weak Causal Framing. The graph shows the (significant) interaction between both factors. Error bars show standard errors of the mean and are also expressed numerically, in parentheses next to the mean value.

The results of Experiment 2 therefore confirm that the prior credibility of a hypothesis is a strong predictor of judgments of the hypothesis' explanatory power. Incredible hypotheses received negative explanatory power ratings, credible hypotheses receive positive ratings. The results also showed that explicit causal framing can increase ratings of explanatory power, but only for incredible hypotheses. While this effect may lead explanatory judgment astray, in most practical cases of explanatory reasoning, people are interested in the explanatory power of hypotheses which they find, at least to a certain extent, credible. As Figure 2 shows, there was no effect of causal framing on explanatory power in this important case.

All in all, the observed patterns in both experiments confirm that the prior credibility of a hypothesis plays a gate-keeping-role in explanatory reasoning: only credible causal hypotheses

qualify as explanatorily valuable. By contrast, implicit or explicit causal framing plays a small to negligible role in influencing judgments of explanatory power.

Experiment 3: Credibility x Generalizability

Participants

Two-hundred-seven participants (mean age 33.4 years, *SD* = 9.1; 123 male) from America (n = 156), India (n = 37) and other countries completed Experiment 3 for a small monetary payment.

Design and Material

The experiment resembled Experiment 1 and 2. Four vignettes, each of which included a headline and five sentences, presented credible and incredible hypotheses. The relation between hypothesis and evidence was expressed by using the causally neutral wording "X co-occurs with Y". The critical manipulation concerned the sample descriptions used in the vignettes, which expressed either narrow or wide generalizability of the study's result. For narrowly generalizable results, the second sentence of a report indicated that the sample of the study encompassed around 5 people (e.g. "The researchers studied 6 people"). For widely generalizable results, the sample included about 10,000 people (*wide* generalizability condition, e.g. "The researchers studied 9891 people").

To control for the possible influence of the content of a particular report, we counterbalanced the allocation of narrow and wide generalizability conditions to the credibility conditions across the items, and created two versions of the experiments (see Appendix C for detailed information). The order in which reports were presented to the participants was individually randomized for each participant.

Procedure

Participants were asked to carefully assess each report with regard to Explanatory Power. Participants' ratings were collected on 7-point scales, with the extreme poles (-3) "I strongly disagree" and (3) "I strongly agree", and the center pole (0) "I neither disagree nor agree".

Analysis and Results

The ratings were analyzed with a two-way ANOVA with the factors Credibility (low, high) and Generalizability (narrow, wide). ANOVA revealed significant main effects of Credibility, F (1, 206) = 83.830; p <.001; η_{part}^2 = 0.289; and Generalizability, F (1, 206) = 29.593; p < .001; η_{part}^2 = 0.126, and no interaction Credibility x Generalizability (p = .085, n.s.).

As with Experiment 1 and 2, credible hypotheses achieved significantly higher ratings than incredible hypotheses (incredible hypotheses: M = -0.01; SEM = 0.10; credible hypotheses: M = 0.95; SEM = 0.08; t-test: t(206) = -9.2; p < .001; d = 0.72). Furthermore, reports with wide generalizability achieved significantly higher ratings compared to reports with narrow generalizability (narrow: M = 0.21; SEM = 0.10; credible hypotheses: M = 0.73; SEM = 0.08; t-test: t(206) = -5.4; p < .001; d = 0.40). Figures 3 and 4 show the main effects for both variables.

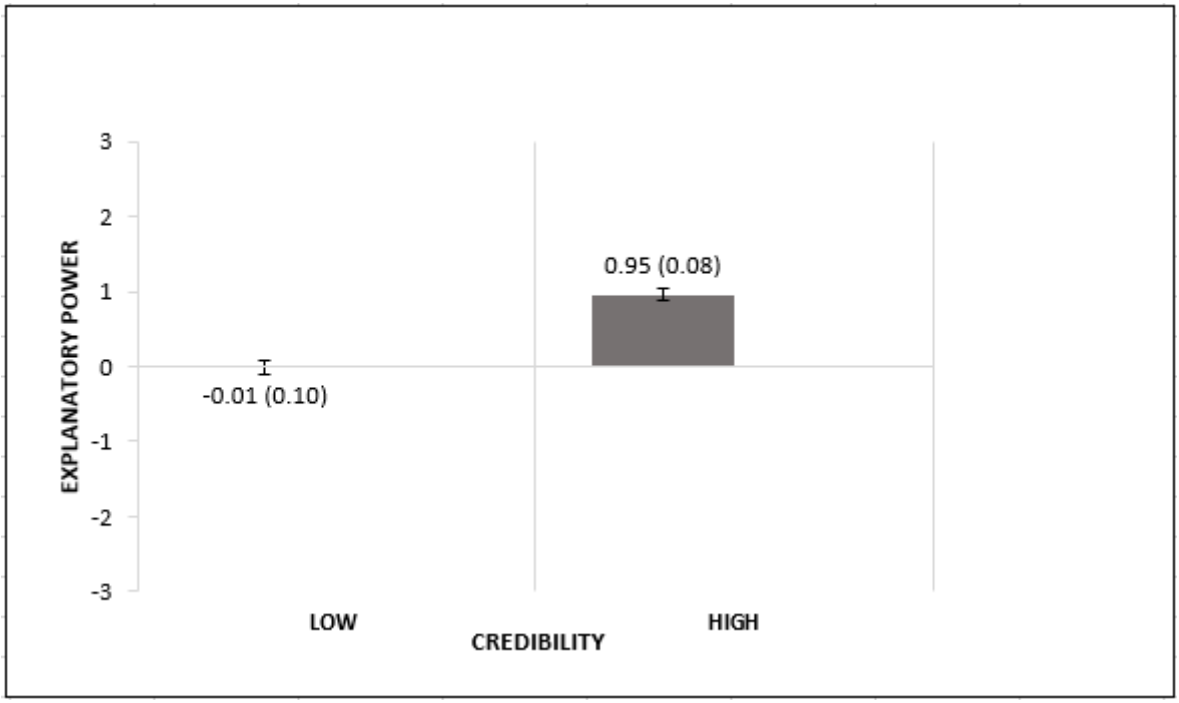

Figure 3. The graph shows how explanatory power ratings vary with regard to Credibility. Ratings were significantly higher for statements with high compared to low Credibility. The graph shows the main effect for this factor. Error bars show standard errors of the mean and are also expressed numerically, in parentheses next to the mean value.

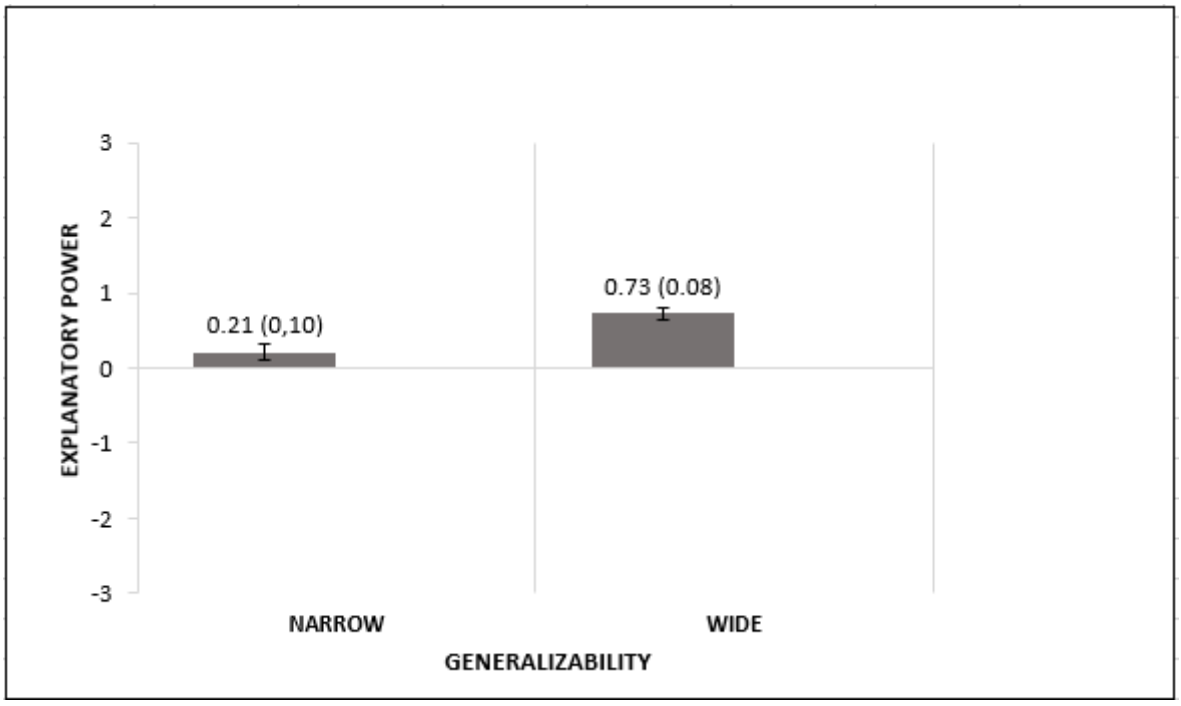

Figure 4. The graph shows how explanatory power ratings vary with regard to Generalizability. Ratings were significantly higher for statements with high compared to low Generalizability. The graph shows the main effect for this factor. Error bars show standard errors of the mean and are also expressed numerically, in parentheses next to the mean value.

Experiment 4 and 5: Credibility x Statistical Relevance

Experiment 4 and 5 examined in what way probabilistic information influences explanatory judgments and how statistical information is taken into account for credible versus incredible hypotheses. Experiment 4 presented the statistical information numerically, Experiment 5 presented it visually.

Participants

Two-hundred-three participants (mean age 34.7 years, *SD* = 9.5; 122 male) from America (n = 168), India (n = 15), and other countries completed Experiment 4 for a small monetary payment. A new sample of N = 208 participants (mean age: 36.0 years, *SD* = 19.7; 133 male), from America (n = 122), India (n = 69), and other countries completed Experiment 5 for a small monetary payment.

Design and Material

The experiments resembled the previous ones. The four vignettes presented credible and incredible hypotheses. The sample descriptions in the vignettes were chosen such that both generalizability and causality were perceived as "neutral", according to the results of our pre-study. This meant that we opted for a medium-sized population sample (like in Experiment 1 and 2) and the wording “X co-occurs with Y” (like in Experiment 3). The novel manipulation was implemented in the part of the vignette where the results of the study are reported. This part now included statistical information. In a case of weak statistical relevance, the frequency of the property of interest was almost equal in the treatment and control group, e.g.: "Among the participants who regularly consumed anabolic steroids, *26 out of 120 (= 22%)* exhibited an exceptional level of physical strength. Among the participants who did not regularly consume anabolic steroids, *24 out of 120 (= 20%)* exhibited an exceptional level of physical strength". For strong statistical relevance, there was a notable difference in the frequency of the property of interest, e.g.: "Among the participants who regularly consumed anabolic steroids, *50 out of 120 (= 42%)* exhibited an exceptional level of physical strength. Among the participants who did not regularly consume anabolic steroids, *7 out of 120 (= 6%)* exhibited an exceptional level of physical strength". While Experiment 4 represented the statistical information numerically like in the previous sentences, Experiment 5 stated the same absolute numbers and replaced the accompanying percentages with two pie charts (see Figure 5).

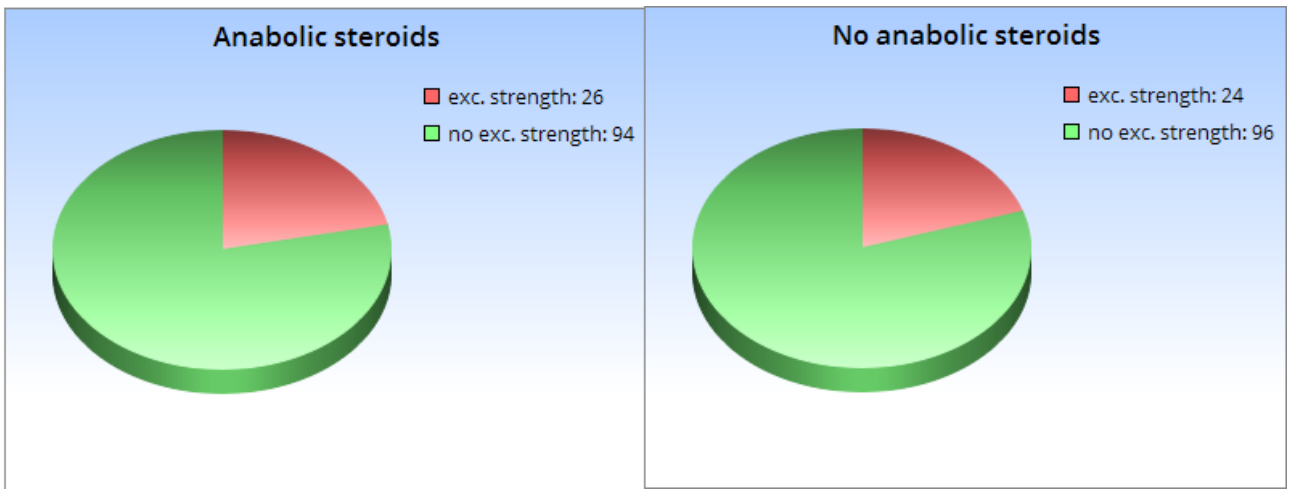

Figure 5. Visual representation of statistical information of the fictitious research groups as provided in Experiment 5.

As in the previous experiments, we counterbalanced the allocation of the weak statistical relevance and strong statistical relevance conditions across the items, and created two versions of each experiment (see Appendix D for detailed information). The order of reports was individually randomized for each participant.

Procedure

Participants were asked to carefully assess each report with regard to Explanatory Power. Again, the ratings of the participants were collected on 7-point scales, with the extreme poles (-3) "I strongly disagree" and (3) "I strongly agree", and the center pole (0) "I neither disagree nor agree".

Analysis and Results

Separate two-way ANOVAs were calculated for Experiment 4 and 5, with the factors Credibility (low, high) and Statistical Relevance (weak, strong). ANOVA of Experiment 4 revealed significant main effects of Credibility (F (1, 202) = 65.3; p <.001; η_{part}^2 = 0.24) and Statistical Relevance (F (1 ,202) = 74.2; p <.001; η_{part}^2 = 0.27) and a significant interaction Credibility x Statistical Relevance (F (1, 202) = 47.7; p <.001; η_{part}^2 = 0.19).

Figure 6 shows the effect sizes and the interaction between both factors as well as the relevant descriptive statistics. Positive levels of explanatory power were only achieved for highly

credible hypotheses and high statistical relevance. The other conditions roughly led to the same explanatory power ratings. This suggests that both factors act as a gate-keeper in explanatory reasoning: if they take their low values, no hypothesis can be rated as explanatorily powerful. On the other hand, if both conditions are satisfied, the effect is very pronounced.

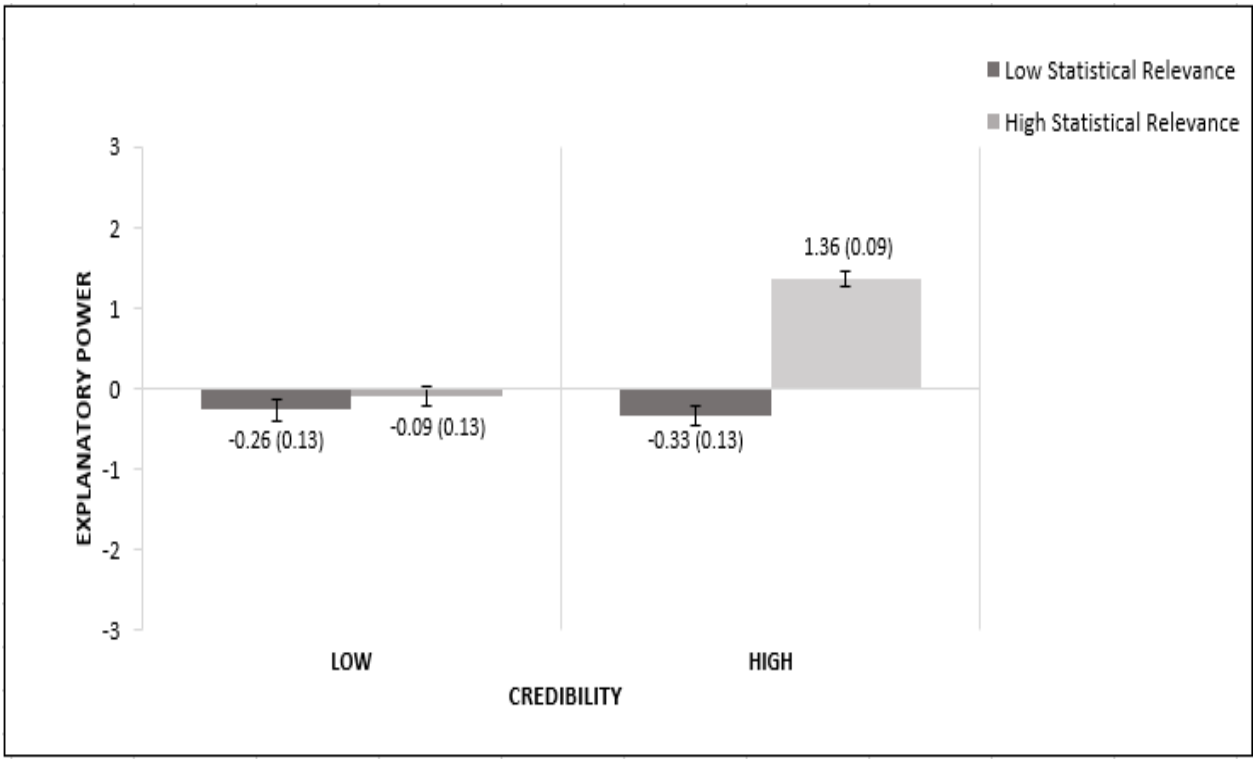

Figure 6. The graph shows how explanatory power ratings vary with regard to Credibility and Statistical Relevance (as presented in Experiment 4). Ratings were significantly higher for statements with high compared to low Credibility, and for statements with high compared to low Statistical Relevance. The graph shows the (significant) interaction between both factors. Error bars show standard errors of the mean and are also expressed numerically, in parentheses next to the mean value.

Similar results were obtained for Experiment 5. ANOVA of Experiment 5 revealed significant main effects of Credibility, F (1, 207) = 38.2; p <.001; η_{part}^2 = 0.16, and Statistical Relevance, F (1 ,207) = 152.5; p <.001; η_{part}^2 = 0.42, and a significant interaction Credibility x Statistical Relevance, F (1, 207) = 47.4; p <.001; η_{part}^2 = 0.10.

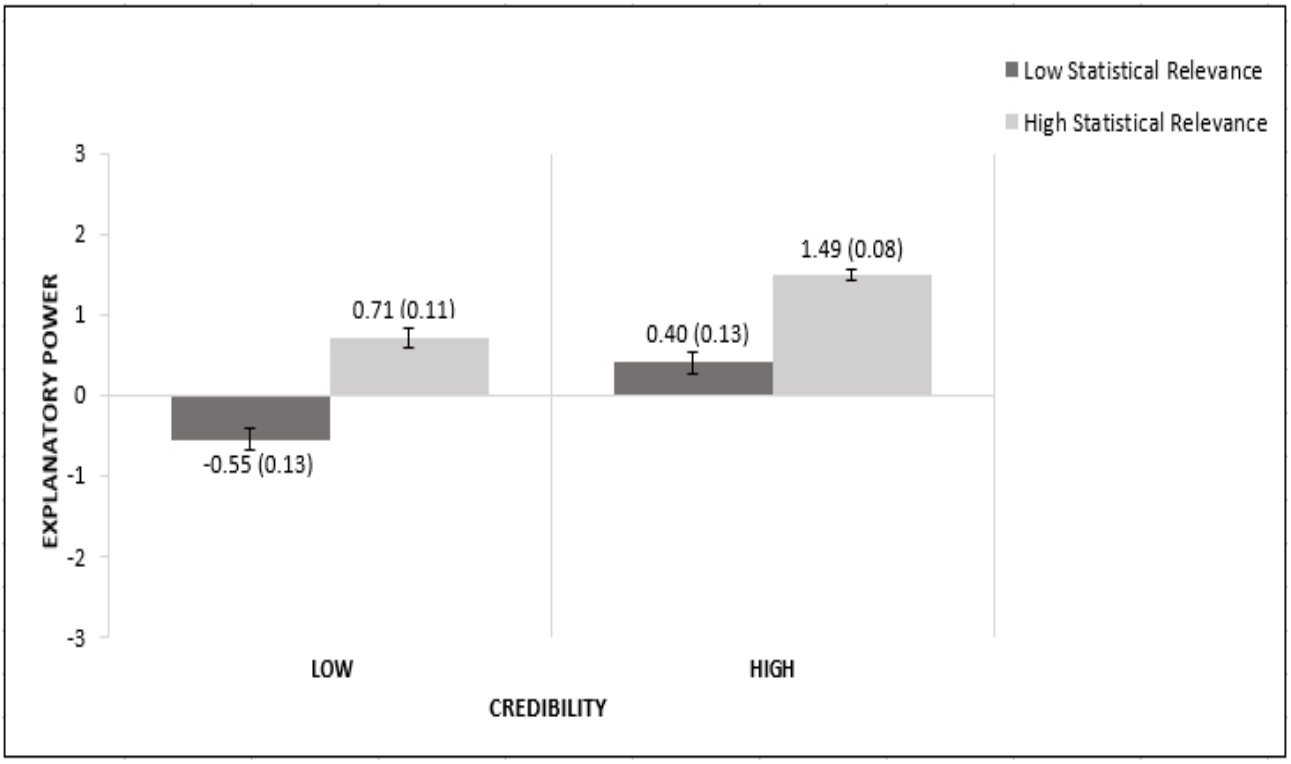

Figure 7. The graph shows how explanatory power ratings vary with regard to Credibility and Statistical Relevance (as presented in Experiment 5). Ratings were significantly higher for statements with high compared to low Credibility, and for statements with high compared to low Statistical Relevance. The graph shows the (significant) interaction between both factors. Error bars show standard errors of the mean and are also expressed numerically, in parentheses next to the mean value.

Figure 7 shows the effect sizes and the interaction between both factors as well as the relevant descriptives. We found a slightly different interaction pattern than in Experiment 4. Again, both variables have to take their high values for a hypothesis to be rated as explanatorily powerful. However, we also see that the gate-keeping role of both variables is weaker than in the case where statistical information was only presented numerically. Either variable taking its high value suffices for a judgment of positive (albeit weak) explanatory power. Like in Experiment 4, the level of explanatory power was by far the highest in the condition where both credibility and statistical relevance were high.

Discussion

We examined the impact of four factors---prior credibility, causal framing, generalizability, and statistical relevance---on judgments of explanatory power. In a series of five experiments, we varied

both the subjective credibility of an explanation and one of the other factors: causal framing, generalizability, and statistical relevance (both with numeric and with visual presentation of the statistics). In Experiments 1 and 2 we found that the impact of causal language on judgments of explanatory power was small to negligible. Experiment 3 showed that generalizable explanations with wider scope positively affected judgments of explanatory power. In Experiments 4 and 5, we found that explanatory power increased with the statistical relevance of the hypothesis for the observed evidence.

Across all experiments, we found that the prior subjective credibility of a hypothesis had a striking effect on how participants assessed explanatory power. In particular, the credibility of an explanatory hypothesis had an important gate-keeping function: the impact of statistical relevance on explanatory power was more significant when credibility was high. On the other hand, the high credibility of a hypothesis controlled for the potentially misleading effect of causal framing on explanatory judgment.

This pattern of findings is consistent with existing psychological research demonstrating that people resist endorsing explanatory hypotheses that appear unnatural and unintuitive, given their background common-sense understanding of the physical and of the social world (Bloom & Weisberg 2007). Our findings are also consistent with the idea that stable background personal ideologies (often referred to as "worldview") can reliably predict whether people are likely to reject well-confirmed scientific hypotheses (Lewandowsky et al., 2013; Colombo, Bucher, & Inbar, 2016). So, scientific hypotheses that are inconsistent with our prior, background, common-sense beliefs or in tension with personal ideologies are likely to be judged as implausible, and may not be endorsed as good explanations unless they are supported by extra-ordinary evidence gathered by some trustworthy source. On the other hand, for hypotheses that fit our prior, background belief or ideology, we often focus on information that, if the candidate explanatory hypothesis is true, would boost its goodness (Klayman & Ha 1987).

This kind of psychological process of biased evidence evaluation and retention bears a similarity to the properties of incremental measures of confirmation called *Matthew properties* (Festa, 2012). According to confirmation measures presenting Matthews properties, an equal degree of statistical relevance leads to higher (incremental) confirmation when the hypothesis is already credible than when it is incredible. The same was observed in our experiment, where the effect of statistical relevance on different dimensions of explanatory power was much more pronounced for credible than for incredible hypotheses. Moreover, the highest ratings of explanatory power, across different experiments, were achieved when, in addition to a credible hypothesis, the report was widely generalizable or statistical relevance was high. Only in those cases, a substantial degree of explanatory power was achieved. This confirms that those factors play a crucial role in explanatory reasoning: a good explanation has to be credible, statistically relevant and widely generalizable. In comparison, the impact of causal framing is negligible.

The interplay we observed between statistical relevance, prior credibility, and explanatory power is also relevant to understanding the relationship between abductive and probabilistic reasoning. In abductive reasoning, explanatory considerations are taken to boost the credibility of a target hypothesis while inducing a sense of understanding (Lipton, 2004). Previous psychological studies investigated the effect on people's assessments of explanatory power of factors like simplicity (Lombrozo, 2007; Bonawitz & Lombrozo, 2012) and coherence (Koslowski *et al*. 2008). Our results advance this body of literature by suggesting that the generalizability of a hypothesis and its statistical relationship to the evidence will boost the acceptability of the hypothesis, especially when the hypothesis has a high prior subjective credibility. High prior credibility may also insulate an explanation from causal framing effects, which may produce a deceptive sense of understanding (Rozenblit & Keil, 2002; Trout, 2002).

Overall, our experiments show that explanatory power is a complex concept, affected by considerations of prior credibility of a (causal) hypothesis, its generalizability and its statistical relevance for the evidence. These factors also figure prominently in (normative) philosophical

theories of explanation. For instance, the D-N model (Hempel, 1965) stresses the generality of the proposed explanation, the causal-mechanical account (Woodward, 2003) requires a credible causal mechanism, and statistical explanations are usually ranked according to their relevance for the observed evidence (Salmon, 1970, Schupbach & Sprenger, 2011).

On the other hand, the multitude of relevant factors in explanatory judgment explains why it has been difficult to come up with a theory of abductive inference that is both normatively compelling and descriptively accurate: after all, it is difficult to fit quite diverse determinants of explanatory judgment into a single unifying framework. In that spirit, we hope that our results will promote an interdisciplinary conversation between empirical evidence and philosophical theorizing, and about the "prospects for a naturalized philosophy of explanation" in particular (Lombrozo 2011, 549; Schupbach, 2015; Colombo, 2016).

References

Aleven, V. A., & Koedinger, K. R. (2002). An effective metacognitive strategy: Learning by doing and explaining with a computer-based Cognitive Tutor. *Cognitive science*, *26*(2), 147-179.

Bloom, P., & Weisberg, D. S. (2007). Childhood origins of adult resistance to science. *science*, *316*(5827), 996-997.

Bonawitz, E.B. & Lombrozo, T. (2012). Occam's rattle: children's use of simplicity and probability to constrain inference. *Developmental Psychology, 48,* 1156-1164.

Brem, S. K., & Rips, L. J. (2000). Explanation and evidence in informal argument. *Cognitive science*, *24*(4), 573-604.

Carey, S. (1985). *Conceptual Change in Childhood.* Plenum, Cambridge, MA

Carnap, R. (1966). *Philosophical foundations of physics* (Vol. 966). M. Gardner (Ed.). New York: Basic Books.

Chi, M.T.H., de Leeuw, N., Chiu, M., & Lavancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science, 18,* 439–477.

Colombo, M. (2016). Experimental Philosophy of Explanation Rising. The case for a plurality of concepts of *explanation*. *Cognitive Science*.

Colombo, M., Postma, M., & Sprenger, J. (2016). Explanatory Judgment, Probability, and Abductive Inference. In Papafragou, A., Grodner, D., Mirman, D., & Trueswell, J.C. (Eds.). *Proceedings of the 38th Annual Conference of the Cognitive Science Society* (pp. 432-437) Austin, TX: Cognitive Science Society.

Colombo, M., Bucher, L., & Inbar, Y. (2016). Explanatory Judgment, Moral Offense, and Value-Free Science. An Empirical Study. *The Review of Philosophy and Psychology*, 7, *743–763.*

Craik, K. J. W. (1943). *The nature of explanation.* Cambridge: Cambridge University Press.

Crupi, V., & Tentori, K. (2012). A second look at the logic of explanatory power (with two novel representation theorems). *Philosophy of Science*, *79*(3), 365-385.

Crupi, V., Tentori, K., & Gonzalez, M. (2007). On Bayesian Measures of Evidential Support: Theoretical and Empirical Issues*. *Philosophy of Science, 74*(2), 229-252.

Douven, I. (2011). Abduction. In E. Zalta (Ed.) *The Stanford Encyclopaedia of Philosophy*. URL = < https://plato.stanford.edu/entries/abduction/ > last accessed January 2017

Dowe, P. (2000). *Physical Causation.* Cambridge: Cambridge University Press.

Earman, J., & Salmon, W (1992). The Confirmation of Scientific Hypotheses. In M.H. Salmon, J. Earman, C. Glymour, J.G. Lennox, P. Machamer et al. (Eds.), *Introduction to the Philosophy of Science*, Englewood Cliff: Prentice Hall, pp. 42–103.

Entman, R. M. (1993). Framing: Toward clarification of a fractured paradigm. *Journal of communication*, *43*(4), 51-58.

Festa, R. (2012). For unto every one that hath shall be given. Matthew properties for incremental confirmation. *Synthese, 184,* 89-100.

Friedman, M.(1974). Explanation and Scientific Understanding. *The Journal of Philosophy 71,* 5-19.

Gilovich, T. (1983). Biased evaluation and persistence in gambling. *Journal of Personality and Social Psychology, 44*, 1110-26.

Greeno, J. G. (1970). Evaluation of statistical hypotheses using information transmitted. *Philosophy of Science, 37,* 279-294.

Hahn, U., & Harris, A. J. (2014). What does it mean to be biased: Motivated reasoning and rationality. *Psychology of learning and motivation*, *61*, 41-102.

Hempel, C.G..(1965). Aspects of Scientific Explanation. In: *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science.* The Free Press, pp. 331-496.

Hempel, C.G. (1945) Studies in the Logic of Confirmation. *Mind*, 54: 97–121.

Hempel, C.G. & Oppenheim, P. (1948). Studies in the Logic of Explanation. *Philosophy of Science* 15: 135-75.

Holyoak, K. J., & Cheng, P. W. (2011). Causal learning and inference as a rational process: The new synthesis. *Annual review of psychology*, *62*, 135-163.

Howson, C. & Urbach, P. (2006). *Scientific Reasoning. The Bayesian Approach*, La Salle (IL): Open Court.

Jeffrey, R. (1969). Statistical Explanation vs. Statistical Inference. In N. Rescher (Ed.), *Essays in honor of Carl G Hempel* (pp. 104–113). Dordrecht: D. Reidel.

Kahneman, D., & Tversky, A. (1982). On the study of statistical intuitions. *Cognition*, *11*(2), 123-141.

Keil F.C., & Wilson R.A.. (2000). *Explanation and Cognition*. Cambridge, MA: MIT Press.

Keil, F. C. (2006). Explanation and understanding. *Annual Review of Psychology*, 57, 227-254.

Khemlani, S. S., Sussman, A. B., & Oppenheimer, D. M. (2011). Harry Potter and the sorcerer's scope: latent scope biases in explanatory reasoning. *Memory & Cognition*, *39*(3), 527-535.

Klayman, J., & Ha, Y. W. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review*, *94*(2), 211-228.

Koslowski, B., Marasia, J., Chelenza, M., & Dublin, R., (2008) Information Becomes Evidence when an Explanation Can Incorporate it into a Causal Framework. *Cognitive Development*, 23: 472–487.

Lagnado, D. A., & Shanks, D. R. (2002). Probability judgement in hierarchical learning: A conflict between predictiveness and coherence. *Cognition, 83*, 81 – 112

Legare, C. H. & Lombrozo, T. (2014). Selective effects of explanation on learning during early childhood. *Journal of Experimental Child Psychology, 126,* 198-212

Lewis, D. (1986). Causal Explanation. In *Philosophical Papers*, Volume 2, New York: Oxford University Press, 214–40.

Lipton, P. (2004). *Inference to the Best Explanation* (second edition). London: Routledge.

Lombrozo, T. (2012). Explanation and abductive inference. In K. J. Holyoak & R. G. Morrison (eds.): *Oxford Handbook of Thinking and Reasoning*, 260–276. Oxford, UK: Oxford University.

Lombrozo, T. (2011). The instrumental value of explanations. *Philosophy Compass, 6,* 539–551.

Lombrozo, T. (2009). Explanation and Categorization: How 'Why?' Informs 'What?' *Cognition, 110,* 248-253.

Lombrozo, T. (2007). Simplicity and probability in causal explanation. *Cognitive Psychology, 55,* 232-257.

Lombrozo, T. (2006). The structure and function of explanations. *Trends in Cognitive Sciences, 10,* 464–470.

Mastropasqua, T., Crupi, V., & Tentori, K. (2010). Broadening the study of inductive reasoning: Confirmation judgments with uncertain evidence. *Memory & cognition*, *38*(7), 941-950.

McGrew, T. (2003). Confirmation, heuristics, and explanatory reasoning. *The British Journal for the Philosophy of Science*, *54*(4), 553-567.

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological review*, *92*(3), 289.

Myrvold, W. C. (2003). A Bayesian account of the virtue of unification. *Philosophy of Science 70,* 399-423.

Nickerson, R. S. (1998). Confirmation bias: a ubiquitous phenomenon in many guises. *Review of General Psychology, 2*, 175-220.

Nicod, J., 1924, *Le problème logique de l'induction*, Paris: Alcan. (Engl. transl. "The Logical Problem of Induction", in *Foundations of Geometry and Induction*, London: Routledge, 2000.)

Pearl, J. (2000). *Causality: Models, reasoning, and inference*. Cambridge: Cambridge University Press.

Preston, J. & Epley, N. (2005). Explanations Versus Applications: The Explanatory Power of Valuable Beliefs. *Psychological Science 10,* 826-832.

Rozenblit, L., & Keil, F. (2002). The misunderstood limits of folk science: an illusion of explanatory depth. *Cognitive Science* 26, 521–562.

Salmon, W. (1989). *Four Decades of Scientific Explanation,* Minneapolis: University of Minnesota Press.

Salmon, W. (1984). *Scientific Explanation and the Causal Structure of the World*. Princeton: Princeton University Press.

Salmon, W. C. (1970). Statistical explanation. In R. G. Colodny (Ed.), *The nature and function of scientific theories* (pp. 173–231). Pittsburgh: University of Pittsburgh Press.

Scheufele, B. T., & Scheufele, D. A. (2010). Of spreading activation, applicability, and schemas. *Doing News Framing Analysis: Empirical and Theoretical Perspectives, New York, Routledge,* 110-134.

Schupbach, J. N. (2015). Experimental Explication. *Philosophy and Phenomenological Research*. doi: 10.1111/phpr.12207

Schupbach, J. N. (2011). Comparing probabilistic measures of explanatory power. *Philosophy of Science*, *78*(5), 813-829.

Schupbach, J. N., & Sprenger, J. (2011). The Logic of Explanatory Power. *Philosophy of Science, 78*(1), 105-127.

Sides, A., Osherson, D., Bonini, N., & Viale, R. (2002). On the reality of the conjunction fallacy. *Memory & Cognition, 30,* 191 – 198.

Spirtes, P., Glymour, C. and Scheines, R. (2000). *Causation, Prediction and Search*, second edition, Cambridge, MA: MIT Press.

Strevens, M. (2008) *Depth: An Account of Scientific Explanation*. Cambridge, MA: Harvard University Press.

Trout, J. D. (2002). Scientific Explanation And The Sense Of Understanding. *Philosophy of Science*, *69*(2), 212-233.

Van Fraassen, B.C.,(1989). *Laws and Symmetry.* Oxford: Oxford University Press.

Walker, C.M., Lombrozo, T., Legare, C., & Gopnik, A. (2014). Explaining prompts children to privilege inductively rich properties. *Cognition, 133*, 343-357.

Woodward, J. (2014). Scientific Explanation. In E.N. Zalta (Ed.) The Stanford Enclycopedia of Philosophy. URL = <https://plato.stanford.edu/entries/scientific-explanation/>

Appendix A (Pre-tests)

A.1 List of statements expressing relationships between an explanans (X) and an explanandum Y) presented in the pre-study on Causal Framing.

Statements	Ratings: Mean (Standard Deviation)
X prevents Y.	-0.59 (1.98)
X co-occurs with y.	0.09 (1.60)
X predicts Y.	0.25 (1.57)
X is associated with Y.	0.70 (1.83)
X promotes Y.	0.80 (1.46)
X is correlated with Y.	0.86 (1.64)
X causes Y.	1.02 (2.01)
X leads to Y.	1.23 (1.67)

A.2 List of hypotheses, presented to the participants (N = 42) of the pre-study on Credibility, and mean ratings (and standard deviation), collected on 7-point scales with the options: "I strongly disagree" (-3), "I disagree", "I slightly disagree", "I neither agree nor disagree" (0), "I slightly agree", "I agree", "I strongly agree" (3).

Hypotheses	Ratings: Mean (Standard Deviation)
Eating pizza co-occurs with immunity to flu.	-1.95 (1.45)
Drinking apple juice co-occurs with anorexia.	-1.86 (1.28)
Breast feeding co-occurs with hair loss in the baby.	-1.83 (1.46)
Vegetarianism co-occurs with aggressiveness.	-1.76 (1.36)
Helpfulness co-occurs with blond hair.	-1.79 (0.23)
Exercising co-occurs with frequent headache.	-1.45 (1.61)
Kleptomania co-occurs with sexual deprivation.	-1.33 (1.53)
Eating crab co-occurs with good eyesight.	-1.12 (1.58)
Attending religious services co-occurs with positive mood.	0.14 (1.69)
Drinking coffee co-occurs with higher blood pressure.	0.43 (1.73)
Vandalism co-occurs with low self-esteem.	0.52 (1.15)
Low interest rates co-occur with a high number of newly built houses.	0.69 (1.65)
Professional success co-occurs with parental income above $ 100,000/year.	0.74 (1.50)

Having breakfast co-occurs with a healthy body mass index.	0.79 (1.22)
Rainy days co-occur with birds breeding.	0.79 (1.62)
Eating hot dogs co-occurs with obesity.	0.83 (1.50)
Drinking whisky co-occurs with liver cancer.	0.90 (1.65)
Smoking cannabis co-occurs with drowsiness.	1.10 (1.27)
Well-being co-occurs with frequent smiling.	1.14 (1.46)
Consuming anabolic steroids co-occurs with physical strength.	1.21 (1.86)

A.3 List of sample description statements presented in the pre-study on Generalizability

1. Sample description based on the *number* of participants of a study

Statements	**Ratings: Mean (Standard Deviation)**
The study investigates 5 people.	-1.88 (1.60)
The study investigates 50 people.	-1.05 (1.82)
The study investigates 100 people.	-0.43 (1.78)
The study investigates 500 people.	0.55 (1.70)
The study investigates 1,000 people.	0.93 (1.80)
The study investigates 10,000 people.	1.24 (2.05)

2. Sample description based on the *type* of participants of a study

Statements	**Ratings: Mean (Standard Deviation)**
The study investigates a group of people who sit in a park.	-0.34 (1.88)
The study investigates a group of people who work at a university.	-0.05 (1.97)
The study investigates a group of people who attend a religious ceremony.	0.07 (1.82)
The study investigates a group of people who have their number in the telephone book.	0.12 (2.19)
The study investigates a group of people who watch a movie in the cinema.	0.22 (1.80)
The study investigates a group of people who wait for their flight at an airport	0.27 (1.91)
The study investigates a group of people who attend a sports event.	0.29 (1.82)
The study investigates a group of people who shop at a mall.	0.49 (1.69)
The study investigates a group of people who are registered on Facebook.	0.85 (1.85)

Appendix B (Experiment 1 and 2)

The table shows the allocation of strong (as opposed to weak) causal framing conditions as implemented by the wording "X leads to Y" (Experiment 1) and "X causes Y" (Experiment 2) to the four hypotheses in the two different versions (A and B) of Experiment 1 and 2. In Experiment 1, n = 103 participants completed version A, the remaining participants (N = 100) completed version B. In Experiment 2, N = 103 completed Version A and N = 108 completed version B.

	Experiment 1 and 2	
Credibility	**Version A**	**Version B**
Low	Eating pizza *is associated with* immunity to flu. **Weak Causal Framing**	Eating pizza *leads to/causes*[2] immunity to flu. **Strong Causal Framing**
Low	Drinking apple juice *leads to/causes* anorexia. **Strong Causal Framing**	Drinking apple juice *is associated with* anorexia. **Weak Causal Framing**
High	Well-being *is associated with* frequent smiling **Weak Causal Framing**	Well-being *leads to/causes* frequent smiling **Strong Causal Framing**
High	Consuming anabolic steroids *leads to/causes* physical strength. **Strong Causal Framing**	Consuming anabolic steroids *is associated with* physical strength. **Weak Causal Framing**

[2]The phrasing "leads to" was used for high/strong causal items in Experiment 1, and was replaced by the explicit causal wording "causes" in Experiment 2.

Appendix C (Experiment 3)

The table shows the allocation of "narrow" and "wide generalizability" conditions to the four hypotheses in the two different versions (A and B) of Experiment 3. N = 104 participants completed Version A, the remaining participants (N = 104) completed Version B.

	Experiment 3	
Credibility	**Version A**	**Version B**
Low	Eating pizza co-occurs with immunity to flu. [...] *The researchers examined 6 persons.* **Narrow generalizability**	Eating pizza co-occurs with immunity to flu. [...] *The researchers examined 10187 persons.* **Wide generalizability**
Low	Drinking apple juice co-occurs with anorexia. [...] *The researchers examined 9891 persons.* **Wide generalizability**	Drinking apple juice co-occurs with anorexia. [...] *The researchers examined 6 persons.* **Narrow generalizability**
High	Well-being co-occurs with frequent smiling. [...] *The researchers examined 10391 persons.* **Wide generalizability**	Well-being co-occurs with frequent smiling. [...] *The researchers examined 5 persons.* **Narrow generalizability**
High	Consuming anabolic steroids co-occurs with physical strength. [...] *The researchers examined 5 persons.* **Narrow generalizability**	Consuming anabolic steroids co-occurs with physical strength. [...] *The researchers examined 9971 persons.* **Wide generalizability**

Appendix D (Experiment 4 and 5)

The table below shows the allocation of "low" and "high statistical relevance" conditions to the four hypotheses in the two different versions (A and B) of Experiment 4 and 5. N = 101 participants completed Experiment 4's version A, the remaining participants (N = 102) completed Experiment 4's version B. In Experiment 5, N = 106 completed Version A and N = 102 completed version B.

	Experiment 4 and 5	
Credibility	**Version A**	**Version B**
Low	Eating pizza co-occurs with immunity to flu. [...] Among the participants who regularly ate pizza, *27 out of 120 (= 23%)* exhibited immunity to flu. Among the participants who did not regularly eat pizza, *25 out of 120 (= 21%)* exhibited immunity to flu. **Low statistical relevance**	Eating pizza co-occurs with immunity to flu. [...] Among the participants who regularly ate pizza, *48 out of 120 (= 40%)* exhibited immunity to flu. Among the participants who did not regularly eat pizza, *6 out of 120 (= 5%)* exhibited immunity to flu. **High statistical relevance**
Low	Drinking apple juice co-occurs with anorexia. [...]Among the participants who regularly drank apple juice, *48 ouWeak support by statisticst of 120 (= 40%)* exhibited anorexia. Among the participants who did not regularly drink apple juice, *6 out of 120 (= 5%)* exhibited anorexia. **High statistical relevance**	Drinking apple juice co-occurs with anorexia. [...]Among the participants who regularly drank apple juice, *26 out of 120 (= 22%)* exhibited anorexia. Among the participants who did not regularly drink apple juice, *24 out of 120 (= 30%)* exhibited anorexia. **Low statistical relevance**
High	Consuming anabolic steroids co-occurs with physical strength. [...] Among the participants who regularly consumed anabolic steroids, 26 *out of 120 (= 22%)* exhibited an exceptional level of physical strength. Among the participants who did not regularly consume anabolic steroids, *24 out of 120 (=*	Consuming anabolic steroids co-occurs with physical strength. [...] Among the participants who regularly consumed anabolic steroids, *50 out of 120 (= 42%)* exhibited an exceptional level of physical strength. Among the participants who did not regularly consume anabolic steroids, *7 out of 120 (= 6%)* exhibited an exceptional level of physical strength. **High statistical relevance**

	20%) exhibited an exceptional level of physical strength. **Low statistical relevance**	
High	Well-being co-occurs with frequent smiling. [...] Among the participants who reported a high level of well-being, *50 out of 120 (= 42%)* smiled frequently. Among the participants who did not report a high level of well-being, *7 out of 120 (= 6%)* smiled frequently. **High statistical relevance**	Well-being co-occurs with frequent smiling. [...] Among the participants who reported a high level of well-being, *27 out of 120 (= 23%)* smiled frequently. Among the participants who did not report a high level of well-being, *25 out of 120 (= 21%)* smiled frequently. **Low statistical relevance**