Running head: INVERSE CONJUNCTION FALLACY

To appear in *Journal of Memory and Language 2006*

The Inverse Conjunction Fallacy

Martin L. Jönsson	James A. Hampton
Department of Philosophy	Department of Psychology
Lund University	City University, London

Corresponding author:

James A. Hampton

Psychology Department, City University

Northampton Square, London EC1V OHB, UK

hampton@city.ac.uk

PHONE +44 2070408520

FAX +44 20708581

Abstract

If people believe that some property is true of all members of a class such as sofas, then they should also believe that the same property is true of all members of a conjunctively defined subset of that class such as uncomfortable handmade sofas. A series of experiments demonstrated a failure to observe this constraint, leading to what is termed the *inverse conjunction fallacy*. Not only did people often express a belief in the more general statement but not in the more specific, but also when they accepted both beliefs, they were inclined to give greater confidence to the more general. It is argued that this effect underlies a number of other demonstrations of fallacious reasoning, particularly in category-based induction. Alternative accounts of the phenomenon are evaluated, and it is concluded that the effect is best interpreted in terms of intensional reasoning (Tversky & Kahneman, 1983).

KEYWORDS: fallacy, conjunction, concepts, beliefs, intensional reasoning

Fallacies in category-based reasoning

There have been many demonstrations of how people's thinking appears to violate logical or statistical laws. For example, Tversky and Kahneman (1983) famously showed that when people are told that someone called Bill is a rather dull individual, then under a variety of circumstances they will judge that a conjunctive statement (e.g. "Bill is an accountant who plays jazz for a hobby") is more likely to be true than one of its conjuncts (e.g. "Bill plays jazz for a hobby") even though this should be impossible by the conjunction law of probability theory. Tversky and Kahneman argued that the basis of this conjunction fallacy is the use of *intensional* reasoning (reasoning based on the similarity of descriptions of classes). People consider which description is most appropriate for Bill, and choose that as the most likely. Of course from the perspective of *extensional* reasoning, (reasoning based on considerations of class membership), such as the axioms underlying probability theory, the answer based on intensional reasoning is incorrect.

The effects of intensional reasoning are not limited to subjective probability estimates. Similar fallacies also occur when people reason about semantic categories, as the following four examples illustrate. [1] First, Hampton (1982) demonstrated intransitivities in people's classification judgments about everyday objects. For instance, people agreed that "A car headlight is a kind of lamp" and "A lamp is a kind of furniture", but then denied that "A car headlight is a kind of furniture". If these statements are understood as expressing beliefs about class inclusion (i.e. "All lamps are furniture"), then this pattern of belief violates the transitivity of the set inclusion relation of ordinary set theory. Second, in the context of category-based induction, Osherson, Smith, Wilkie, López, and Shafir (1990) demonstrated that people's intuitions about argument strength are often not in accordance with classical logic. For example, argument (1) was judged to be stronger than argument (2):

(1) All robins have sesamoid bones; therefore all <u>birds</u> have sesamoid bones

(2) All robins have sesamoid bones; therefore all <u>ostriches</u> have sesamoid bones.

If all birds have a property then of course necessarily all ostriches must do so too, so (2) must be at least as strong as (1), if not stronger. (Argument strength here is taken to be something akin to the perceived conditional probability of the conclusion given the premise). To show this, compare (2) with the rephrasing of (1) given in (1a):

(1a) All robins have sesamoid bones; therefore all ostriches, and all other birds, have sesamoid bones.

Since (1a) requires an additional proposition to be true in its conclusion, (1a) clearly cannot be stronger than (2). Yet (1) and (1a) are logically equivalent.

The two remaining examples were reported by Sloman (1993; 1998) in the context of category-based induction. In his premise specificity effect Sloman demonstrated that people are prone to judge arguments with more specific premises such as (3) to be stronger than arguments with more general premises such as (4)

(3) All birds have ulnars, therefore all robins have ulnars

(4) All animals have ulnars, therefore all robins have ulnars

even though ordinary logic treats them as both perfectly strong, the class of robins being included in the class of birds which itself is included in the class of animals. In his inclusion similarity effect, Sloman showed that people are prone to judge arguments with typical conclusion categories such as (5) as stronger than arguments with less typical conclusion categories such as (6), even though, again, since mammals and reptiles are both included in the class of animals, classical logic treats them both as perfectly strong.

(5) All animals use norepinephrine as a neurotransmitter, therefore, all mammals use norepinephrine as a neurotransmitter

(6) All animals use norepinephrine as a neurotransmitter, therefore, all reptiles use norepinephrine as a neurotransmitter

What these phenomena all have in common is that people are using similarity between concepts as the basis of their judgments, and as a result they are ignoring considerations based on sets or class inclusion. Just as with the conjunction fallacy, Hampton (1982) argued that intransitivity in categorization arises from the use of similarity to make category membership judgments, together with the fact that there is greater similarity between (for example) headlights and lamps, and between lamps and furniture than there is between headlights and furniture. Similarity is not a transitive relation, and so categorization based on similarity may on occasion also be intransitive. Likewise both Osherson et al. (1990) and Sloman (1993, 1998) proposed models to account for their induction effects in which similarity between premise and conclusion categories plays a major role.

To this list of fallacies we now add another, which we consider may reveal more directly the thought processes that lead to the latter three phenomena. According to common logical intuition, if a property is true of all members of a class, then it should also be true of any subset of that class. Hence, if one agrees to the proposition “All humans are rational animals” one should be equally prepared to agree to the proposition “All humans born in the United States are rational animals”. More formally, in ordinary predicate logic, a statement of the form (7) (expressed verbally in 7a)

(7) $\forall x((P(x)\wedge Q(x))\supset R(x))$

(7a) *For all x, if x is p and x is q then x is r*

can be formally deduced from a statement of the form (8 / 8a)

(8) $\forall x(P(x)\supset R(x))$

(8a) *For all x, if x is p then x is r*

These deductions have substantial intuitive appeal (providing that *p* and *q* are not disjoint sets). However, it will be the main aim of this article to demonstrate that people are often unprepared to agree to the more specific statement even though they agree to the general one. Since the statements in question relate to actual beliefs about known properties (rather than hypothetical beliefs about blank predicates), we argue that this fallacy provides a more direct demonstration of this style of conceptual reasoning, so that it can provide an important clue to understanding similarity based class inclusion fallacies more generally.

The Inverse Conjunction Fallacy

Our interest in exploring the possibility of this new fallacy came from a phenomenon reported by Connolly, Fodor, Gleitman, L.R., and Gleitman, H. (2003). Connolly et al. claimed that many current theories of concept combination (Cohen & Murphy, 1984; Hampton, 1991; Murphy, 1988) embody a “default to the stereotype” strategy. A person “defaults to the stereotype” when he or she assumes that the representation corresponding to the meaning of a Modifier-Noun (MN) expression (e.g. “red apple”) is similar to the representation corresponding to the meaning of the relevant noun (N) expression (e.g. “apple”) in all respects that are orthogonal to the modifier. So if “is crunchy” is orthogonal to “is red” (in apples), red apples are represented as being just as crunchy as other apples are. Connolly et al. then showed that statements of the form “MN are P” are generally judged less likely to be true than matched statements of the form “N are P”, particularly when the modifier is atypical of the noun. Connolly et al. took this result as a demonstration that the default to stereotype strategy is not used in conceptual combination. We have since replicated their result and believe it to be a robust finding, even though we are skeptical of the conclusions that they draw from it (Jönsson & Hampton, 2005).

Connolly et al.’s (2003) experimental result suggested to us that people might be prone to entertain inconsistent thoughts in a way that has not hitherto been demonstrated. For if people are prone to assert that “Sofas have backrests” is more likely to be true than “Uncomfortable handmade sofas have backrests” they might also go so far as to agree with the universally quantified sentence “All sofas have backrests” while denying that “All uncomfortable handmade sofas have backrests”, which would be logically inconsistent (as noted above). Note that no inconsistency threatens if unquantified generic statements are used. Because generic statements carry a weaker implicit quantification such as “typically” or “most” (see Krifka et al., 1995), the generic belief that “Sofas have backrests” is quite compatible with believing that some subclass of sofas do not have backrests. Hence, in order to demonstrate inconsistency in beliefs, it is crucial that the test sentences are explicitly universally quantified.

In addition to providing evidence of a new form of similarity based fallacy, the proposed test is also critical to deciding between two accounts of Connolly et al.’s (2003) finding. They interpreted their result as showing that prototypical property information for a modified noun phrase is not inherited by default from the noun. In fact they argued that prototypes of concepts do not combine in any consistent compositional way – a conclusion that has been taken as a strong argument against concepts being prototypes (Fodor, 1998). Instead, concepts have to combine extensionally – something is in the class MN, just if it is both in the class M and in the class N. If people combine the concepts extensionally in this way, then it should be clear that any property that is universally true of the members of either of the M or N sets should be universally true of the members of the combined concept.

The alternative account of Connolly et al.’s (2003) result is the one that we favor (Jönsson and Hampton, 2005). A key part of our account is that, in keeping with psychological models of concept combination (Hampton, 1987; Murphy, 1988), properties are inherited by a complex concept in proportion to their importance or “definingness” for each of the constituent concepts. Thus the weight of the feature “has a back rest” for uncomfortable handmade sofas will be an average of its weight for sofas and its weight for uncomfortable handmade objects in general. (There are a number of exceptions to this rule – see Hampton, 1987, for evidence and details). Since uncomfortable handmade objects do not generally have backrests, the feature will carry less weight for the modified concept “uncomfortable handmade sofa” than for the unmodified concept “sofa”. (We also assume that there are no background knowledge or conceptual consistency effects involved here –

see Murphy, 1988, 2002). According to this account, the addition of universal quantifiers to the statements is likely to change people's judgments very little. It is our contention that people rarely think conceptually in terms of class inclusion or class intersection, as demonstrated in the examples of fallacious reasoning already given. Universally quantified sentences and generic sentences are likely to be treated similarly unless the context very clearly supports extensional reasoning.

There are therefore two clearly different predictions about the outcome of our test. Extensional accounts of concept combination predict that people should consider the MN statement "All uncomfortable handmade sofas have backrests" to be no less likely to be true than the N statement "All sofas have backrests", on the grounds that the latter entails the former. Alternatively, intensional models of concept combination predict that the modified noun statement will be considered less likely to be true, simply because the property is a feature of only one of the conjoined concepts, and not one that is considered necessarily true.

We term this possible effect the inverse conjunction fallacy, in explicit acknowledgement of its close relation to Tversky and Kahneman's (1983) well-known conjunction fallacy. There are actually three different ways in which one can exhibit the effect. The most direct contravention of logical constraints is to say "Yes" to the truth of a statement like "All sofas have backrests" while saying "No" to "All uncomfortable handmade sofas have backrests". Accordingly we will call this pattern of responding Yes/No. It is clearly inconsistent given that uncomfortable handmade sofas are a subclass of sofas.

The remaining two ways of being inconsistent relate to relative degrees of confidence in agreeing with or rejecting the two statements. If one agrees that both statements are true, it would still be inconsistent to express *greater confidence* in unmodified N statements than in modified MN statements. We will call this pattern of responding Yes/Yes. To illustrate this, suppose that we model confidence in the truth of a statement in terms of the proportion of plausible possible worlds in which the statement would be true. So if my confidence in winning a bet were 20% that would be equivalent to my imagining 5 equally likely outcomes (five possible worlds), in 4 of which I lose, and in one of which I win. (See Lewis, 1986, or Stalnaker, 1984, for classical treatments of knowledge and belief in terms of possible worlds). Given this interpretation of confidence, believing that all N are P with greater confidence than that all MN are P is logically inconsistent since it would imply at least one possible world in which all N are P but not all MN are P. (The same inconsistency arises if the possible worlds are not assumed to be equally likely.)

The third way to display the inverse conjunction fallacy is the mirror image of the second way. Believing that not all N are P with lower confidence than believing that not all MN are P (the No/No pattern) is once again inconsistent. It would be equivalent to believing that the statement “Some MN is not P” is more likely than the statement “Some N is not P”. But clearly whenever the former is true, then so is the latter.

To clarify the relation between the two conjunction fallacies, Tversky and Kahneman’s (1983) original conjunction fallacy occurs when people think it more likely that an individual is a member of a conjunction than a member of one of the conjuncts. It is therefore primarily a fallacy concerning the likelihood of an instance belonging in a set. Our inverse version of the fallacy is that people think it more likely that a property is universally true of one of the conjuncts than of a conjunction. Hence it is about the likelihood of a property being true of a set. The inversion arises as a result of the switch from consideration of members (extensions) to consideration of properties (intensions).

The first two experiments set out to investigate whether people are actually prone to make inverse conjunction fallacies. Student participants were tested in two designs. The between-subjects design (Experiment 1) divided the two versions of a statement between two groups of participants whereas the within-subjects design (Experiment 2) involved giving both versions of a particular statement to the same individuals. The experiments were run in parallel with random allocation of participants to each experiment. It was predicted that if participants were at all sensitive to the potential inconsistency, then the fallacy would be more frequently found in the between-subjects design, where each statement was only seen in one form. Results were analyzed in terms of the Yes/No, Yes/Yes and No/No versions of the possible fallacy. However it should be noted that since we deliberately chose statements that had high credibility to start with (e.g. all ravens are black, all sofas have backrests), the number of No/No fallacies was expected to be fairly low, since they required that people disbelieve these statements.

Experiment 1

Method

Participants. Twenty-one undergraduates at City University, London volunteered to participate in the experiment. A small number of participants in the experiments reported here were not native speakers of English, but self-reported their level of English as competent or fluent bilingual. No participant took part in more than one of the experiments reported.

Procedure. Each participant was given one of two booklets with instructions and 36 sentences. The words “Yes / No” and the numbers 1 through 10 appeared to the right of each sentence. Participants circled the word yes or no for each sentence to indicate if it was true or not. They then indicated their confidence by circling a number between 1 (= very unconfident) and 10 (= very confident). The booklet took about 10 minutes to complete.

Materials. The 36 sentences in each booklet consisted of 28 target and 8 filler sentences. Order of sentences (in their alternate forms) was the same in both booklets. Target sentences occurred in one of two versions, differing only in whether the subject noun (e.g. sofa) was modified or not. Unmodified sentences N were simple explicitly universally quantified sentences (All sofas have backrests). Modified sentences MN contained two modifiers prefixed to the head noun (All uncomfortable handmade sofas have backrests). Modifiers were chosen to be atypical of the head noun class (they did not occur in property norms, Cree and McRae, 2003), yet still consistent with both head noun and predicate. Predicates were chosen so that the resulting N sentence should be plausibly true. Many of the sentences were taken with permission directly from Connolly et al.'s (2003) materials since they provided a good fit with the above restrictions. Fillers were either analytically true (All triangles have three corners) or highly plausible (All large explosions are dangerous) in order to encourage full use of the confidence scale.

Design. The sentence pairs were divided into two sets. Set A were in unmodified form in booklet 1, and modified form in booklet 2, while Set B were the other way round. In this way each participant saw just one version of each target sentence, and the two versions of each sentence were rated by different groups of participants. Each participant judged 14 modified and 14 unmodified sentences. Four filler sentences were included at the start to avoid warm-up effects and 4 more filler sentences were distributed among the target sentences in the same position in each booklet. Filler sentences were generally given high confidence "yes" responses as expected.

Results

A Yes/No inverse conjunction fallacy would be seen if participants gave more yes responses to the unmodified than to the modified sentences. Where participants gave no response (25 data points, or 4%), the data were treated as missing. Overall, unmodified sentences received 21% more yes responses (*M* = .72, *SD* = .21) than did modified sentences (*M* = .51, *SD* = .26). Analyses of variance (ANOVA) by subjects (F1) and by items (F2) were run with proportion of "yes" responses as dependent variable, and with booklet and sentence type as factors. Only the main effect of sentence type was significant (*F1*(1,19) = 17.7, *F2*(1,26) = 43.3, Min F' (1,33) = 12.6, p = .001). Overall, the effect was seen in 17 of 21 participants, and in 23 of 28 sentence pairs.

Confidence data were analyzed separately for yes and for no responses. Mean confidence in a yes for N sentences (*M* = 7.8, *SD* = 1.2) was significantly greater than that for MN sentences (*M* = 6.3, *SD* = 2.1, *F1*(1, 18) = 17.2, *F2*(1, 27) = 20.25, Min F'(1, 41) = 9.3, p < .005). For "no" responses there was no significant difference between confidence for N sentences (*M = 6.5, SD = 1.6)* and for MN sentences (*M = 5.8, SD = 1.9)*. There was therefore evidence for the Yes/Yes form, but not for the No/No form of the fallacy.

In sum, nearly three quarters of unmodified sentences but only half the modified sentences were endorsed as universally true, and further, where both sentences were endorsed as true, greater confidence was expressed in the unmodified sentences. These two results indicate a strong tendency for participants to commit the inverse conjunction fallacy. In order to see whether people would continue to make fallacious responses when faced with both versions of the same sentence, Experiment 2 employed a within-subjects manipulation of sentences. It is arguable for example that once having judged that "All MN are P" is false, people would then show reluctance to agree that "All N are P" is true.

Experiment 2

Method

Participants. Twenty-three students at City University, London volunteered to participate.

Materials and Procedure. Exactly the same 36 sentences were used as in Experiment 1, and the same procedure was followed.

Design. In Experiment 2 the two sets of sentence pairs from Experiment 1 were used to create two different booklets as replications of a within-subjects design. We decided to keep the length of the booklets the same between experiments, so that there would be no increase in the amount of attention required. Each booklet therefore consisted of a set of just 14 target sentence pairs plus the same 8 fillers as before. The first half of the booklet contained 7 target sentences in their unmodified versions and 7 in modified versions, while the second half contained the alternate versions of the same 14 sentences. In this way participants judged both versions of each target sentence in the same booklet, with an average distance between the two of 18 sentences.

Results

Two of the participants failed to provide scores for half or more of the items and were excluded from the study. There were 10 participants left for one booklet and 11 for the other. In addition 16 pairs of responses (5%) were treated as missing because participants did not respond to one or other sentence. The first key result of interest was the frequency with which participants said yes to the unmodified version of a sentence and no to the modified version (the Yes/No fallacy). Seventeen of the 21 participants made 2 or more responses of this kind to the 14 sentence pairs that they saw. In addition, 24 of the 28 sentence pairs had at least 1 Yes/No response, and 17 pairs had 2 or more from the 10 to 11 participants rating them. Table 1 shows the breakdown of all 278 response pairs according to whether the N and MN versions of a sentence pair were given yes or no responses. Of the 210 occasions where the N version of a sentence was judged true, 56 (27%) had the MN version judged false. If participants had respected the logic of class inclusion, then there would have been no Yes/No responses at all in Table 1. The finding of 56 Yes/No responses was therefore remarkable. To test that the results were not owing to random responding due to lack of attention, the rate of Yes/No responding (20%) was compared to that of No/Yes responding (8%). The difference was significant (*F1*(1,20) = 15.2, *F2*(1,27) = 15.2, *Min F'*(1,46) = 7.6, p < .01). As in Experiment 1, the overall proportion of yes responses was significantly greater for the N (.76) than for the MN sentences (.63), a difference of 13%.

Yes/Yes and No/No responses were also analyzed for an inverse conjunction fallacy based on confidence. Recall that comparing N to MN sentences, it is inconsistent to have greater confidence in a Yes or to have lower confidence in a No for the N version of a sentence pair. For the 154 Yes/Yes response combinations, 97 were rated with more confidence in the unmodified N form and only 25 with more confidence in the modified MN form. Of participants, 19 showed this effect, and only 1 showed the opposite effect while for items, 24 showed the effect and only 4 showed the opposite (*F1*(1,20) = 36.0, *F2*(1, 27) = 30.2, *Min F'*(1,47) = 16.4, p < .001). For the 47 No/No response combinations, there was little evidence of fallacious responding, since 24 of the responses involved rejecting the unmodified form with more confidence than the modified form, and only 7 showed the reverse effect that we identified as indicating a fallacy. In fact, for both Yes/Yes and No/No response combinations, there tended to be more confidence expressed for the unmodified sentences, regardless of whether they were accepted or rejected.

Summing together all three forms of the fallacy, participants produced an inverse conjunction fallacy of one sort or another on just over half (56%) of the sentence pairs they considered. Surprisingly, the order in which the modified and unmodified version of each sentence was presented had no significant effect on a participant's tendency to give a fallacious response. In fact, participants tended to give slightly more Yes/No responses when the modified version was presented first (23%) than the other way around (17%). Contrary to expectation, having decided that *not* all uncomfortable handmade sofas have backrests, people continued to agree that all sofas *do* have backrests. Comparing the rate of "yes" responses with Experiment 1, it was clear that while agreement with the N sentences was about the same (.72 vs. .76) agreement with the MN sentences increased from Experiment 1 to Experiment 2 (.51 vs. .63). So if the repeated judgments in Experiment 2 had any effect it was towards increasing acceptance of MN sentences rather than reducing agreement with N. However ANOVA with experiment and sentence type as factors showed no significant interaction effect (*F1*(1,40) = 2.20, *p* = .14, *F2*(1,27) = 3.89, *p* < .06), and the change in agreement with MN across experiments was significant across items but not across subjects.[2]

The fallacy we have observed is only a fallacy if our participants accepted that the MN class was a subset of the N class – for example that all uncomfortable handmade sofas are sofas. The reader is invited to consult the Appendix, where it may be confirmed that it is implausible to suppose for example that dirty German lambs are not lambs, or that thin polyester shirts are not shirts. Sofas do not cease to be sofas when they are uncomfortable and handmade. Nonetheless, concepts denoted by modified noun phrases are not always members of the unmodified noun class. First there is the well-known case of privative expressions like "fake dollars" which are not dollars. Furthermore, Hampton (1982, Experiment 1) showed that people did not treat phrases like school furniture or office furniture as proper subsets of furniture, but accepted that some office furniture is not furniture. There are also biological concepts where older folk nomenclature is at odds with scientific knowledge, thus leading to anomalous naming (for example the silk oak, the tan-oak and the poison oak are none of them true oaks, although they may have acorns). In Experiment 3 we therefore set out to test the presumption that in the case of our materials people do see the modified noun phrase as referring to a subset of the unmodified noun.

Experiment 3

Method

Participants. Twenty-one undergraduates at City University, London participated in the experiment in exchange for course credit.

Procedure. Each participant was given a booklet with instructions and 66 sentences. The words "yes/no" appeared to the right of each sentence. Participants circled the word yes or no for each sentence to indicate if it was true or not. The booklet took about 10 minutes to complete.

Materials. The 66 sentences in each booklet consisted of 28 target and 38 filler sentences. All sentences were of the form "All MN's are N's". The target sentences (e.g. "All uncomfortable handmade sofas are sofas") were generated from the target sentences used in Experiments 1 and 2. Ten of the filler sentences were intended to be true based on ordinary modifier noun combinations with typical modifiers (*All crunchy red apples are apples, All long curly hair is hair*). These fillers were later used as catch trials in order to identify participants who might have become over-cautious, being unwilling to affirm even clearly analytical statements, or who were responding erratically for some other reason. These fillers also provided clear examples of "yes" responses, so that participants would not feel it necessary to say yes to any of our target sentences. The remaining 28 filler sentences were intended to be false and used privative adjectives as modifiers. Privatives belong to a class of modifier that explicitly contradicts the set relation that we were interested in confirming. For example counterfeit dollars are not dollars, and imitation leather is not leather. By including these two kinds of fillers in the list, we provided participants with clear examples for yes and no responses so that they would be less influenced by the demand characteristics of the list of target sentences (which we predicted would be all true).

Design. Ten of the filler sentences were included at the start of the booklet in order to avoid warm-up effects. The rest of the fillers were randomly distributed among the target sentences. The target sentences appeared in random order.

Results and Discussion

Taking the data of all 21 participants, the mean percentage of yes responses to the target subset sentences was 90%, to the true filler sentences 94% and to the false privative fillers 23%. All the true filler sentences had at least 19 out of 21 yes responses. Taking this level of agreement as a criterion of full acceptance, 22 of the 28 target sentences were clearly judged to be true. Of the 6 target sentences with less good agreement, the two worst were "All futuristic fruit wagons are wagons" and "All Appalachian stake-out shacks are shacks", which each received only 12 out of 21 yes responses. On examination, it turned out that all of the "no" responses given to the true fillers came from just 5 of the 21 participants. It is therefore quite possible that these 5 participants were either being too cautious, or were inattentive to the task. If the responses of these 5 participants were omitted, then 94% of responses to target sentences were yes, 17 of the target sentences had 100% yes responses, and just 4 had more than 2 no responses. At the same time the false privative fillers rate of yes responses actually decreased from 23% to 20%, indicating that the 5 excluded participants were not just biased towards saying no, but were also generally responding in an idiosyncratic fashion.

In conclusion, the great majority of our target sentences were endorsed as true. The 6 sentences with a lower acceptance rate were noted, and account was taken of them in the analyses of experiments to be reported below. Although it would have been better to replace these items for subsequent experiments, the results of Experiment 3 were not available in time to enable us to do this. We therefore resorted to a post-hoc check to confirm that later results were not owing to these 6 items.

Looking back at the results of Experiments 1 and 2, there was no significant correlation across target sentences between the number of participants answering yes to a subset relation in Experiment 3 and the degree to which the sentence showed the Yes/No fallacy in those experiments (r(26) = -0.01 and 0.11 respectively). Excluding the 6 items in question left the size and significance of the modifier effects in Experiment 1 and 2 unaffected. For Experiment 1 the mean probability of yes responses was .72 for N sentences and .53 for MN (compared to .72 and .51 before exclusion). For Experiment 2 the probabilities were unchanged at .76 for N and .63 for MN sentences. We thus conclude that our assumption was correct and the effect does not stem from the fact that participants did not believe that the MN were a subclass of the N.

Experiment 4

An alternative explanation for the fallacy to which we turned next is that people may not interpret the word “All” as mapping onto the universal quantifier in logic. We may frequently use sentences starting with “All” in common parlance as indicating a strongly generic rather than a strictly universal quantification. Experiment 4 set out to manipulate the verbal form of the quantifier in order to explore this possibility.

Our first variation was to add the word “always”, as in “All sofas always have backrests”). According to an account of the results suggested by Johannes Persson (personal communication, January, 2006), the target sentences may be interpreted as containing not one, but two quantifiers. One refers to variation within the class of objects (i.e. every object) and the other refers to variations in an object over time. The unmodified sentences may be interpreted just in terms of the object class, whereas the modified sentences may be interpreted also in terms of the implicit temporal quantifier. For example “All uncomfortable handmade sofas have backrests” might be understood as “All sofas have backrests *when they are uncomfortable and handmade*”. If the modifier introduces a second temporal dimension over which quantification has to apply, and which is not considered for the unmodified sentence, then the Yes/No response is no longer inconsistent. Adding the word “always” should counteract this possibility. In addition to this possible account, the addition of “always” could be seen as simply serving to strengthen the universal nature of the quantifier.

The other two variations on the quantifier were directed at encouraging participants to attend to the set inclusion relation by using explicitly extensional terms. In one condition “All” was replaced by “Every single”, which was intended to draw attention to individual objects in the extension. In the other condition we used the quantifier “100%”, following Gigerenzer’s finding that fallacies in probabilistic reasoning can be reduced when frequencies are used in place of likelihoods (Gigerenzer & Hoffrage, 1995, for a review see Lagnado & Sloman, 2004).

We predicted that if the fallacy is at least in part to be explained by vagueness in the meaning of the word “All” in common language, then one or more of our variations would lead to a significant reduction in, or even elimination of the fallacy.

Experiment 4

Method

Participants. Sixty-nine undergraduates at City University, London participated in the experiment in exchange for course credit.

Design and Materials. Participants were allocated randomly to one of 4 conditions. The All condition was a replication of Experiment 1 and used exactly the same sentences. The three remaining conditions differed only in terms of the quantifiers used, giving the All-Always, the Every-Single and the 100% conditions. Each condition involved a pair of booklets constructed in the same way as for Experiment 1, but with the appropriate change in quantifier. Within each booklet, half the target sentences were modified and half unmodified. As a small improvement on the design, the original 8 true fillers used in Experiment 1 were replaced with 12 fillers half of which were expected to be true and half false (see Appendix). In addition, to mirror the target sentences, half the fillers had modified subject nouns and half unmodified. Ten copies of each of the 8 booklets were distributed at random to participants who completed them in class or individually. Between 7 and 10 participants completed and returned each booklet, giving between 15 and 19 participants per condition.

Results

Individual means for proportion of yes responses given to modified and unmodified sentences in each condition are shown in the top panel of Figure 1, together with 95% confidence intervals. Overall, the modified sentences were judged true 60% of the time, while the unmodified sentences were judged true 72% of the time. To test for the Yes/No fallacy, a 2-way ANOVA was run with quantifier type (4 levels between subjects and within items) and sentence type (modified vs unmodified, within subjects and within items) as factors. Only the main effect of sentence type was significant on both analyses (*F1*(1,65) = 26.3, *F2*(1,27) = 13.673, *Min F'*(1,56) = 9.0, *p* < .005). The effect of quantifier type was significant across items (*F2*(3,81) = 4.2, p < .01) but not across subjects (*F1*<1). There was no significant interaction (*F1* and *F2* <1). Planned contrasts for each quantifier condition between modified and unmodified sentences were significant by items for all conditions, and by subjects for all except for the All-always condition (p = .09). *Min F'* statistics were also calculated for this contrast in each condition, and *Min F'* was only significant for the Every Single condition (*Min F'*(1,42) = 4.87. p < .05). Averaged across conditions the fallacy effect was seen in 78% of all items and in 66% of all participants.

Following the finding in Experiment 3 that a small set of 6 sentence pairs may not always be interpreted as having a subset relation, the analysis was re-run with these pairs excluded. The main effect of sentence type was still significant (*F1*(1,65) = 18.2, *F2*(1,21) = 7.7, Min F' (1,40) = 5.41, p < .05), and the effect still occurred in 17/22 items (77%).

Mean confidence ratings data were calculated separately for yes and for no responses in order to examine the Yes/Yes and No/No versions of the fallacy. The lower panel in Figure 1 shows mean confidence for yes and no responses for the four conditions. Error bars show 95% confidence intervals.

Mean confidence across conditions for yes responses was 8.3 on a 10 point scale for unmodified N sentences, and 6.8 for modified MN sentences, a difference of 1.5, confirming the Yes/Yes fallacy. A 2-way ANOVA of the confidence given to yes responses confirmed a significant effect of modifier (*F1*(1, 64) = 67.9, *F2*(1, 26) = 54.6, *Min F'* (1, 67) = 30.2, p < .001). The effect of quantifier condition was significant only by items, and not by subjects (p = .056). A post hoc comparison of confidence across conditions showed that the All condition had greater confidence ratings than either Every Single or 100% (p < .05 on both subjects and items analyses). Strengthening the quantifier thus seems to have reduced confidence, but not the likelihood of saying yes. Finally, the modifier effect on confidence was analyzed for each condition separately. The contrast was significant on a Min F' analysis for all conditions (p < .005), except for the All condition, where Min F’ was marginal (Min F’ (1,40) = 3.67, p = .06).

For no responses, mean confidence was 6.8 for N and 5.8 for MN sentences. Since this difference was in the direction that is consistent with logical reasoning, there was no evidence for the No/No version of the fallacy.

As before, the results indicated a strong tendency for participants to commit the inverse conjunction fallacy. Further, since no interaction between type of sentence and type of quantifier was found, this tendency cannot be attributed to some peculiarity of the meaning of ‘all’. More specifically, the data rules out an explanation in terms of an implicit temporal quantifier since the “All Always” condition didn’t deviate significantly from the other quantifiers. There was some observed difference in the size of the effect across conditions (see Figure 2), with All and Every Single generating a 0.15 drop in agreement between unmodified and modified sentences, and All-Always and 100% generating a 0.10 drop. Our design was however clearly not powerful enough to detect a significant interaction of this size, so it is safer to conclude for the present that the Inverse Conjunction Fallacy is not strongly affected by different means of expressing universal quantification. Whatever difficulties there may be in people interpreting “All” as really meaning all seem to apply also to the other forms of quantifier. Both the likelihood of saying yes, and the confidence with which a yes response was given showed a preference for unmodified forms of the sentences.

Since the effect appears to be so robust, our final two experiments were directed at testing the boundary conditions of the effect further. Changing the quantifier in a between-subjects design had little detectable effect. But what if the within-subjects design of Experiment 2 were adapted so that the relation between the two sentences of a pair became more obvious? Would participants then resort to more logical extensional reasoning and avoid making the fallacy? Experiment 2 presented each participant with the modified and the unmodified version of each sentence, with a gap of about 18 sentences in between. In

Experiments 5 and 6 we presented the two versions of each sentence next to each other. In the first of these experiments, participants were asked to judge whether one sentence or the other was more likely to be true, or whether they were equally true. In the second (Experiment 6), the participants simply saw each version of a sentence one above the other as a pair, and judged whether each was true in turn.

An explicit comparison between the likelihood of the two sentences, or judging them as a pair, should highlight the purely formal relationship between the two sentences and the subset relation that it signifies. Hence, we predicted less occurrence of fallacious responding in these two experiments.

Experiment 5

Method

Participants. Eighteen undergraduates at City University, London participated in the experiment in exchange for course credit.

Procedure. Each participant was given one of two booklets with instructions and 50 sentence-pairs. The options "A is more likely", "Equally Likely" and "B is more likely" appeared to the right of each sentence-pair. The first sentence of each sentence pair was labeled "A", the second sentence was labeled "B". Participants circled the option that they thought was the most appropriate for each sentence pair. The booklet took about 10 minutes to complete.

Materials. The 50 sentence pairs in each booklet consisted of 28 target and 22 filler pairs. The order of the pairs was the same in both booklets, and the sentences that made up each sentence pair was also the same for both booklets. Each target pair contained the two versions of one of the sentences that had been used in Experiment 1 (e.g. ‘All lambs are friendly" together with "All dirty German lambs are friendly"). The filler pairs had the same form as the target pairs, with one modified and one unmodified sentence. Half of the filler-pairs used privatives as in Experiment 3 so that the unmodified sentence was intuitively more plausible than the modified one (“All guns are dangerous” and “All plastic replica guns are dangerous”) and half used knowledge-based effects relating the modifier to the predicate to make the modified sentence intuitively more plausible (“All bags are flammable” and “All dry paper bags are flammable”). We took care to make the more likely filler statements more or less analytic in order for the universal quantification to be appropriate.

Design. Order of the sentences within each pair in the first booklet was reversed in the second booklet. Eight filler sentence-pairs were included at the start to avoid warm-up effects. The rest of the filler pairs were distributed randomly among the target pairs.

Results

Combining the results across the two booklets, the filler items were judged as predicted, with 75% of responses favoring the predicted member of each pair. For target sentences, participants judged the unmodified statement as the most likely about half the time (46%), they judged the two sentences as equally likely half the time (50%), and they very rarely judged the modified statement as the more likely (4% of the answers). There was only one item (saxophones) where the modified statement was judged more likely overall than the unmodified one. Proportions of responses were unaffected if the 6 pairs with weaker subset relations from Experiment 3 were excluded. The response distribution across participants showed strong individual variation. Roughly half the participants used a majority of “Equally likely” responses, and the other half mostly preferred to say the unmodified sentence was more likely. Proportion of “Equally likely” responses varied across participants from 7% to 93% of responses. No participants succeeded in completely avoiding any judgments in favor of the unmodified sentence, even when the 6 weaker subset pairs were ignored.

So despite the close comparison required of the two versions of each sentence, the inverse conjunction fallacy was again frequently observed. The “best” 2 participants judged the sentences equally likely on over 80% of pairs, but the large majority was just as happy to judge the unmodified sentence as more likely than the modified.

For the final experiment, we made another attempt to find conditions in which the subset relation would become obvious to our participants, and so discourage them from giving the fallacious response. In Experiment 5, a judgment that the unmodified sentence was more likely to be true could be consistent with the belief that both sentences were false. Although we have argued above that it is still logically inconsistent to think that it is more likely that the modified sentence is false than that the unmodified sentence is false, it is possible that asking for comparative judgments of possibly false sentences weakens the interpretation that would be placed on the quantifier “All”. The discovery that two of the filler sentences occurring early in the booklet were also clearly false, as opposed to just unlikely to some degree, (“All birds fly” and “All stones are used for construction”) increased our concern that we were not directly measuring the belief that the unmodified sentence was true, while the modified sentence was false. Experiment 6 therefore used a different procedure. We still placed the two versions of each sentence next to each other, but we didn’t require a choice between them. Instead, we asked, for each version of the sentence whether it was true or false. This way of probing beliefs (in terms of yes/no) should make the logical characteristics of the situation yet more salient. We also took the opportunity to replace the two filler pairs that we considered to be false (see Appendix).

Experiment 6

Method

Participants. Twenty undergraduates at City University, London participated in the experiment in exchange for course credit.

Procedure. Each participant was given one of two booklets with instructions and 50 pairs of sentences. The options “True” and “False” appeared to the right of each sentence. Participants circled the option that they thought was the most appropriate for each sentence.

The booklet took about 15 minutes to complete.

Materials. The 50 sentence pairs in each booklet consisted of the same 28 target and 22 filler pairs as in Experiment 5 (with the exception of the replacement of 2 fillers).

Design. Four filler sentence-pairs were included at the start to avoid warm-up effects and the rest were distributed randomly among the target pairs. Order within each pair was reversed across the two booklets.

Results

Three responses (0.6%) were missing and were excluded from the analysis. One participant was excluded because she always gave different responses to the two sentences of each pair, and so was judged to have misunderstood the task. For the remaining 19 participants, the frequency of response pairs to unmodified and modified sentences for each pair is shown in Table 2. Only 12% of response pairs were of the type we have labeled a fallacy, saying yes to All N are P and no to All MN are P. That corresponded to 17% of sentence pairs where the N sentence was considered true. This rate of responding was still significantly above that of giving the reverse response combination No/Yes (4%), so we consider that the Yes/No responses were not just random errors in selecting a yes or no response (*F1*(1,18) = 13.0, *F2*(1,27) = 9.1, *Min F'*(1,45) = 5.35, $p < .05$). The rate of making the fallacy was reduced further however if the 6 items with weaker subset relations from Experiment 3 were excluded. Rate of Yes/No responses fell to 9% which was no longer significantly more than the 4% rate of No/Yes on the more conservative *Min F'* test (*F1*(1,18) = 5.2, $p < .05$, *F2*(1,21) = 3.7, $p = .07$, *Min F'*(1,38) = 2.16, $p = .15$). Analysis of order of sentences showed absolutely no difference between pairs where the modified or the unmodified sentence was listed first.

The aim of Experiments 5 and 6 was to test the boundary conditions of our phenomenon. If placing two sentences side-by-side has the effect of drawing attention to the logical relation between them, and hence encouraging logical extensional reasoning, then we predicted that fallacious responding should decrease. The present experiment, unlike Experiment 5, succeeded in greatly lowering the rate of Yes/No judgments. Participants judged the sentences either to be both true or to be both false 84% of the time, compared with only 50% judging them equally likely in Experiment 5. Even so, there was only 1 participant who gave zero fallacious responses – even when the 6 sentence pairs with weaker subset relations were excluded.

General discussion

Overall, our experiments have demonstrated that people have a strong tendency to accept unmodified universally quantified sentences with greater frequency and/or greater confidence than the equivalent modified versions. The analysis of confidence ratings suggested that confidence was also generally higher for unmodified sentences – either for accepting them or for rejecting them. In the case of acceptance, this led to a logical inconsistency, and in the case of rejection it did not.

The inverse conjunction fallacy is closely related to phenomena found in category-based induction studies reported and discussed by Osherson et al. (1990) and Sloman (1993, 1998), and described in the introduction. It is therefore instructive to draw parallels between them before considering possible explanations of the inverse conjunction fallacy itself.

Induction fallacies and the inverse conjunction fallacy

The induction phenomena relate to how the similarity between premise and conclusion categories influences the perceived argument strength for (logically) perfectly strong arguments. Osherson et al.'s (1990) *inclusion fallacy* for example (that an argument from Robins to Birds is stronger than an argument from Robins to Ostriches) arises because the step from Birds to Ostriches is not taken as being perfectly strong. If it were, then the argument from Robins to Birds to Ostriches would be as strong as that from Robins to Birds. That it is not perceived to be perfectly strong indicates a failure to generalize from all birds to all subsets of birds. (Hampton and Cannon, 2004, demonstrated that inductive arguments from one category member to another were seen as stronger the more typical the conclusion item).

In the *inclusion similarity* phenomenon (Sloman, 1993, 1998) arguments from a superordinate category to a subclass are seen as stronger if the particular subclass is more typical of the category, while in the *premise specificity* phenomenon the same arguments are seen as stronger if the general category is more specific. In each case two arguments that are perfectly strong are treated as being of different strength when the premise-conclusion similarity differs between them. So once again, facts that are stipulated to be true of all members of a category are not generalized reliably to all members of subsets of that category.

These three induction phenomena may well rely on the same mechanisms as the inverse conjunction fallacy to the extent that, like the inverse conjunction fallacy, they involve a failure to accept the logical entailment of class inclusion – that if a general class has a property, then all subclasses, no matter how atypical or how distant taxonomically, must also have the property. There are also some important differences between the demonstration of the inverse conjunction fallacy and the induction effects that suggest that the former may reflect a more fundamental process.

First, whereas the induction effects require participants to accept the truth of a blank premise and then judge the likelihood of the conclusion, the inverse conjunction fallacy leaves participants free to accept or deny either of the two sentences. People are therefore shown to entertain beliefs deviating from logical norms in a wider context than just within the context of an explicit argument involving possibly counterfactual suppositions. (López, Atran, Coley, Medin, and Smith, 1997, reported that category-based inductive arguments may be rejected because respondents are unwilling to accept the truth of the premise). Second, the induction phenomena use blank predicates whereas the inverse conjunction fallacy uses predicates familiar to the participants. Hence, the inconsistency illustrated by the inverse conjunction fallacy may be considered more fundamental, since no conditional or hypothetical reasoning is involved. Third, the fact that the same head noun (e.g. *sofa* or *raven*) is seen in each sentence pair, renders the inclusion relation between the sentences much more explicit in the inverse conjunction fallacy, than when terms from a taxonomy are used as in Sloman's examples (e.g. *plant* versus *moss)*. In fact Sloman (1998, Expt 4) found that if the inclusion relation is explicitly added to the arguments in his induction task, the effects disappear.

In spite of these differences between Sloman's phenomena and ours, we consider that his explanation of why people treat logically equivalent arguments differently is substantially correct. It seems plausible that the explanation for all of these phenomena should be in terms of intensional reasoning using abstract representations of the relevant categories, with consequent neglect of the set relations holding between category members. We turn now to considering different accounts of the results in more detail.

Explaining the inverse conjunction fallacy

Similarity-based accounts. Tversky and Kahneman (1983) suggested that people break the conjunction rule because they use a "representativeness" heuristic. People judge how representative the individual would be of the simple or conjunctive categories. So our unimaginative friend Bill would be more representative of an accountant who plays jazz for a hobby than of just anyone who plays jazz for a hobby. Rather than considering the problem extensionally in terms of inclusion relations among the sets involved, people solve it intensionally, in terms of the similarity of the concepts.

Accounts of the three induction effects described above also rely on intensional reasoning – the computation of similarity between premise and conclusion categories. It seems very plausible therefore to adopt a similar approach to explaining the inverse conjunction fallacy. People clearly ignore extensional considerations when they make the fallacy, and the fact that the effect involves the use of *atypical* modifiers is also indicative that similarity is at issue. (Connolly et al., 2003, showed that the reduction in likelihood of unquantified MN sentences compared to N sentences was greater if the modifier was atypical, a result confirmed by Jönsson & Hampton, 2005). A similarity-based model of conceptual combination such as Hampton's (1987) composite prototype model should therefore provide a useful framework for understanding the result. Taking the modified noun phrase MN as a conceptual combination, then the model suggests that it will inherit the attributes of each concept in approximately equal measure (subject to the resolution of possible conflicts and the incorporation of possible knowledge of particulars – see Rips, 1995). We would therefore expect attributes of the head noun to be incorporated into the conceptual combination but with reduced weight (since they are only true of one half of the concept pair). Our first account is therefore straightforwardly that people (at times erroneously) evaluate the truth of universally quantified statements, not by considering whether any counterexamples may exist, but rather by considering the weight or centrality of an attribute as part of the concept representation. Attributes tend to have reduced centrality in modified concepts, and so people are less likely to agree to their universal truth.

Restricted counterexample search accounts. An alternative account makes use of a notion central to mental models theory, that we evaluate the truth of a statement via the search for a model in which the statement may be false (Johnson-Laird, 1983). Unless a statement is *analytically* true (as is 2 + 3 = 5), then universally quantified truth must involve the contingent absence of counterexamples. That is to say for example that "All ravens are black" is true simply because of the non-existence of any non-black ravens. From this perspective, an explanation should therefore address why people might fail to find counterexamples to the unmodified N statements, but find counterexamples for the MN statements. We consider three possibilities.

The first possibility is that the modifier focuses the attention on some circumstance that is not available when evaluating the statement in its unmodified form. Thus the effective search space for N statements may be much smaller than that for MN statements. If asked if all coins are made of metal, this question may be taken implicitly to refer to types of coins currently in circulation as legal currency. At a pinch it might include recent but now obsolete examples such as French francs or German marks. The possibility that old Egyptian coins may have been made from some other form of precious mineral, or that there may be a tribe in the Himalaya that use coins carved from the tusks of ibex may then be overlooked. It is not until the modified statement is encountered that the possibility of a broader notion of the unmodified concept class – of a more widely defined set of category members – comes to mind. The roots of this behavior may lie in the vagueness of our concepts (Keefe & Smith, 1997). If we are unclear (as we often are) just what should count as a sofa or a shack, then such terms can be taken with a broader or a narrower sense. It would then be logically consistent to agree to the N statement (in a narrower sense it is true that all ravens are black) but to disagree with MN (in a broader sense of "raven" the class may include foreign species of unknown color). Somewhat paradoxically then, consideration of a narrower class (jungle ravens are more specific than ravens) may lead to a widening of the category of relevant exemplars to be considered.

A second possibility is that the modifier provides a retrieval cue for knowledge-based reasoning that may serve to undermine the truth of the statement. This possibility may occur even if there is no direct contradiction of the predicate implied in the modifier (i.e. we rule out privative cases like "All broken clocks tell the time" or "All dead pigeons can fly"). For example, for the statement "All young jungle ravens are black", each modifier yields an additional consideration that might be helpful in finding counterexamples. From "jungle" a person may reason from knowledge of biological theory thus: "Coloration often serves as camouflage. Ravens living in jungles may therefore have evolved to be some other color. Color change is also plausible since color is an attribute that can change easily without entailing deep changes in other attributes," (see Johnson & Keil, 2000). Or from "young" they might reason based on analogy with remembered examples: "Swans change color as they grow older, so young ravens may do also". In other words, the modifiers may spark off either theory-based or experience-based reasoning that would not normally be used when evaluating the N sentence.

The third possibility is that the MN concepts engage *modal* rather than existential interpretations of the sentences (Rips, 2001). A modal interpretation directs the search for counterexamples beyond the universe of existing objects and into the realm of what may be possible. Consider again "All uncomfortable handmade sofas have back rests". All sofas we have ever seen have back rests, but if a sofa is handmade, it would be possible for someone to make it differently, and if it had no back rest it would be likely to be uncomfortable. The modifier may switch the search for counterexamples from actual to possible cases.

All of these accounts suggest that the modifier provides access to a wider range of possible counterexamples. A problem with these accounts is that if someone is aiming for logical consistency then after expansion to a wider context, or after a deeper consideration of possibilities, it is strange that they would not then keep in mind the same counterexamples when subsequently accepting N sentences. We would have expected a much reduced inverse conjunction fallacy in Experiment 2 for those sentence pairs where the MN was judged first, and the N second, but we found no evidence for this. Similarly, if one of these accounts is correct the inverse conjunction effect should disappear in a forced choice setting such as that in Experiment 5. Processing the two sentences together in order to compare them should make available the same search space and the same possible counterexamples. Yet nearly half the time, participants still rated the unmodified sentence more likely than the modified sentence.

A fall-back position might be to appeal to some pragmatic consideration – that when evaluating the N sentences in Experiment 2, the participants reverted to the default context because the narrower category context was more pragmatically appropriate. For example it could be argued that having said that not all old Egyptian coins are made of metal, one could then reasonably say that it is still true that all coins (i.e. all normal everyday coins) are. One may be implicitly contrasting the atypical subset with the rest of the category in this case. But again, it seems implausible to think that in the forced choice setting of Experiment 5, where the two sentences are interpreted together in order to come to a decision on which is more likely, the counterexamples found for one of the sentences should be ignored when assessing the other.

Another reason for doubting the generality of the counterexample search accounts is that they do not explain the range of phenomena found in category-based induction. If you are *told* that all birds have ulnars, then you have no good reason to feel it more likely that robins do than that ostriches do. The use of blank predicates in these tasks takes away the responsibility for judging the truth of the unmodified sentences from the participant, so that the question of failing to notice counterexamples does not arise. If a single explanation is to be found to cover both induction effects and the inverse conjunction fallacy, then the intensional reasoning account is to be preferred.

Existential quantifiers and Russell's theory of definite descriptions. A third competing explanation goes as follows. A statement like “All jungle ravens are black” conveys the presupposition that there are such things as jungle ravens. This presupposition might be treated as something independent of the truth conditions of the statement, but it might also be treated as part of these truth conditions. Bertrand Russell provided a solution for the problem of reference to non-existents (referential failure) in sentences such as “The present king of France is bald” with his theory of definite descriptions (Russell, 1905, 1919). He proposed that the interpretation of such a sentence entailed the conjunction of three propositions –

(9) There exists at least one thing x such that x is the King of France

(10) There exists at most one thing x such that x is the King of France

(11) x is bald

Given that (9) is false, then the conjunction of (9), (10) and (11) is false, and there is no need to evaluate (11). In a similar vein, “All jungle ravens are black” might be interpreted as “There exist things that are jungle ravens, and all such things are black”. Certainly then, if one had doubts about whether there are any such things as jungle ravens, one’s confidence in the conjunctive statement will be low. It would therefore be consistent to deny that “All jungle ravens are black” while affirming that “All ravens are black” if the denial was based on believing that jungle ravens do not exist. This explanation fits well with there being no change in response to N sentences when the modified MN sentences occurred before the unmodified ones in Experiment 2. Since the reason for not believing “All MN are P” is that one believes that there are no MNs, exposure to “All MN are P” should not influence your confidence in “All N are P”.

It follows from this account that people should produce more inverse conjunction fallacies if there is reason to doubt the existence of the MN class. In our materials, there are clearly better reasons to doubt the existence of some MN (jungle ravens for example) than others (uncomfortable handmade sofas, or thin polyester shirts). Accordingly we had three independent judges rate whether or not they thought that the MN of the target sentences did actually exist. They all agreed that 15 did exist, and that 4 did not, and there was disagreement about the remaining 9. Therefore the account offered here would not apply to the majority of our sentences. Correlation between number of judges agreeing that the MN did not exist and the size of the fallacy effect across items in our initial experiments was -.18 for Experiment 1 and +0.18 for Experiment 2, neither of which were significant. In Experiment 4, there was no difference at all in the mean difference in acceptance of unmodified and modified sentences for those 21 pairs that were considered by most judges to exist and those 7 pairs considered not to exist (12.7% vs 12.8%). Nor were the results of

Experiment 5 affected (46% vs 47% preferring the unmodified sentence in each case). Finally the items of questionable existence actually showed less fallacious responding (6%) in Experiment 6 than the rest (14%). There was therefore no evidence that referential failure played a role in any of our results.

Pragmatic interpretation of "All". Our final account concerns the interpretation of the quantification of the statements. Exaggeration is a common device in everyday language use (we do it all the time, and we mean literally all the time), so people may be taking a pragmatic position in which "all" is not taken to mean "all without exception". Once quantification becomes "almost all", then there are much weaker constraints on whether one can believe N while disbelieving MN. However, Experiment 4 made evident that the inverse conjunction fallacy is not restricted to 'all'. In fact, we observed no significant differences between "all", "all-always", "every single" and "100%" when it comes to participants tendency to perform the fallacy. In none of the conditions did the incidence of the effect drop to zero.

More importantly, saying that there is a certain looseness with how quantifiers such as 'all' are interpreted, provides at most a redescription of the phenomenon without really explaining why the particular pattern of responses is obtained, or why it should be reduced in Experiment 6. So in order to provide an explanation of the inverse conjunction fallacy this account has to be supplemented in some way. The account in terms of intensional reasoning explains neatly why changing the apparent force of quantifiers should have little effect. Unless the subset relation is made highly salient, as when the sentences are placed side-by-side and each is judged true or false in turn, people turn to their intensional representations and avoid considerations of sets and subsets.

Conclusions

The explanation of the Inverse Conjunction Fallacy that we favor in the end is that people are poor at judging the truth of universally quantified statements because the human knowledge system seems to have no generally reliable way of accessing the information needed for telling whether such statements are true or false, (outside of the realm of concepts with stipulated definitions). Proving a contingent universally quantified statement true in logic requires an exhaustive search for counterexamples and a subsequent failure to find any. In the case of a set with a known finite number of members, then this can be easily done. We can know that all months of the year have at least 28 days, or that all past US presidents are male. For everyday concept classes however we do not have this type of set. Not possessing an easily searchable exemplar space, we rely instead on the strength with which the predicate is represented as part of our concept. When the concept is modified with an atypical modifier, then the strength of the properties composing the concept prototype is reduced, and so we are less likely to endorse the predicate as true. We are subsequently prone to deviate from the logical norm for reasoning about category properties, unless the subset relation is made highly salient by placing the two sentences side-by-side (Experiment 6). Even then, if we are asked to say which is *more* true as in Experiment 5, we still prefer the general unmodified sentence with its higher feature strength.

References

Cohen, B. & Murphy, G.L. (1984). Models of concepts. *Cognitive Science, 8*, 27-58.

Connolly, A.C., Fodor, J.A., Gleitman, L.R., & Gleitman, H. (2003). Why stereotypes don't even make good defaults. Unpublished manuscript, University of Pennsylvania.

Cree, G.S. & McRae, K. (2003). Analysing the factors underlying the structure and computation of the meaning of chipmunk, cherry, chisel, cheese, and cello (and many other such concrete nouns). *Journal of Experimental Psychology: General, 132*, 163-201.

Fodor, J.A. (1998). *Concepts: Where Cognitive Science Went Wrong*. Oxford: Clarendon Press.

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review, 102*, 684-704.

Hampton, J.A. (1982). A demonstration of intransitivity in natural categories. *Cognition, 12*, 151-164.

Hampton, J.A. (1987). Inheritance of attributes in natural concept conjunctions. *Memory & Cognition, 15*, 55-71.

Hampton, J. A. (1991). The combination of prototype concepts. In P.J.Schwanenflugel (Ed.), *The Psychology of Word Meanings* (pp. 91-116). Hillsdale, NJ: Erlbaum.

Hampton, J.A. & Cannon, I. (2004). Category-based induction: An effect of conclusion typicality. *Memory & Cognition, 32,* 235-243.

Harman, G. (2002). Internal critique of logic and practical reasoning. In D. M. Gabbay and H. J. Ohlbach (Eds.) *Studies in Logic and Practical Reasoning*.. Amsterdam, Elsevier Science. 1.

Johnson-Laird, P.N. (1983). *Mental Models: Towards a Cognitive Science of Language, Inference and Consciousness*. Cambridge: Cambridge University Press.

Johnson, C., & Keil, F.C. (2000). Theoretical Centrality vs Typicality in Conceptual

Combinations. In Keil, F.C. and Wilson, R.A. (Eds.) *Explanation and Cognition*, (p.327-360) Cambridge, MIT Press.

Jönsson, M.L. & Hampton, J.A. (2005). Effects of noun modification on the plausibility of attribute information . Paper presented at the Annual Meeting of the Psychonomic Society, Vancouver, November.

Keefe, R. & Smith, P. (1997). Theories of vagueness. In R.Keefe & P. Smith (Eds.), *Vagueness: A Reader* (pp. 1-57). Cambridge: MIT Press.

Krifka, M., Pelletier, F.J., Carlson, G.N., ter Meulen, A., Link, G., & Chierchia, G. (1995). Genericity: An Introduction. In G.N.Carlson and F.J.Pelleter (eds.) *The Generic Book* (pp 1-124). Chicago: University of Chicago Press.

Lagnado, D.A., Sloman, S.A. (2004). Inside and Outside Probability Judgment. In D.J.Koehler and N.Harvey (Eds.) *Blackwell Handbook of Judgment and Decision Making*, pp. 157-176. Oxford, UK: Blackwell Publishing.

Lewis, D. (1986) *On the Plurality of Worlds*. Oxford: Blackwells.

López, A., Atran, S., Coley, J.D., Medin, D.L., & Smith, E.E. (1997). The tree of life: Universal and cultural features of folkbiological taxonomies and inductions. *Cognitive Psychology, 32,* 251-295.

Murphy, G.L. (1988). Comprehending complex concepts. *Cognitive Science, 12,* 529-562.

Murphy, G.L. (2002) *The Big Book of Concepts*. Cambridge, MA: MIT Press.

Osherson, D.N., Smith, E.E., Wilkie, O., López, A., & Shafir, E. (1990). Category-based induction. *Psychological Review, 97,* 185-200..

Rips, L.J. (1995). The current status of research on concept combination. *Mind and Language, 10,* 72-104.

Rips, L.J. (2001). Necessity and natural categories. *Psychological Bulletin, 127*, 827-852.

Russell, B. (1905). On denoting. *Mind, 14,* 479-493.

Russell, B. (1919). *Introduction to Mathematical Philosophy*. New York: Clarion Books/ Simon and Schuster.

Sloman, S.A. (1993). Feature-based induction. *Cognitive Psychology, 25,* 231-280.

Sloman, S.A. (1998). Categorical inference is not a tree: The myth of inheritance hierarchies. *Cognitive Psychology, 35,* 1-33.

Stalnaker, R. (1984). *Inquiry*. Cambridge, Mass.: MIT Press

Tversky, A. & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review, 90,* 293-315.

Appendix

1) Target Sentences Used in all Experiments (with appropriate changes by condition).

Unmodified Subject	Modified Subject	Predicate
All candles	All expensive purple candles	are made of wax
All caterpillars	All Canadian poisonous caterpillars	have many legs
All coins	All old Egyptian coins	are made of metal
All coyotes	All old white coyotes	howl
All crocodiles	All albino African crocodiles	are dangerous
All ducks	All baby Peruvian ducks	have webbed feet
All hamsters	All dark-skinned savannah hamsters	are furry
All kites	All silk weather kites	have strings
All lambs	All dirty German lambs	are friendly
All nectarines	All genetically manipulated giant nectarines	grow on trees
All ostriches	All Paleolithic European ostriches	have long necks
All pearls	All oval South Sea pearls	are hard
All penguins	All solitary migrant penguins	are black and white

All pigs	All wild Samoan pigs	can be turned into pork chops
All ravens	All young jungle ravens	are black
All refrigerators	All inexpensive commercial refrigerators	can be used for storing food
All rhubarb	All homegrown Albanian rhubarb	is grown for food
All saxophones	All expensive hand-made saxophones	are made of brass
All shacks	All Appalachian stake-out shacks	are made for storage
All shirts	All thin polyester shirts	can be worn for warmth
All sinks	All round antique sinks	can retain water
All sofas	All uncomfortable handmade sofas	have back rests
All squirrels	All black Nicaraguan squirrels	eat nuts
All storks	All domestic hybrid storks	have long legs
All thimbles	All Belgian painted thimbles	are worn for protection
All tissues	All medical tissues	are made of paper
All tortoises	All South American fighting tortoises	are slow
All wagons	All futuristic fruit wagons	are used by pulling them

2) True Filler sentences used in Experiments 1 and 2

All triangles have three corners.

All computers are electronic.

All cars have wheels.

All large explosions are dangerous.

All fish can swim.

All birds are animals.

All bachelors are unmarried.

All humans breathe air.

3) Filler sentences used in Experiment 3

True Fillers

All dusty illustrated books are books

All crunchy red apples are apples

All half-full water bottles are bottles

All long curly hair is hair

All square traditional tables are tables

All great brick walls are walls

All sturdy triangular containers are containers

All smart buttoned shirts are shirts

All short bent carrots are carrots

All hand-painted grey cups are cups

False (privative) fillers

All handwritten counterfeit leases are leases

All plastic replica guns are guns

All long false noses are noses

All beautiful unreal fish are fish

All obviously fake tickets are tickets

All petite ceramic trees are trees

All bogus company shares are shares

All miniature toy elephants are elephants

All unconvincing mock lions are lions

All large-scale simulated operations are operations

All alluring pseudo-sciences are sciences

All carved wooden birds are birds

All shiny imitation leather is leather

All friendly imaginary persons are persons

All small phony alarm-clocks are alarm-clocks

All implausible fictitious animals are animals

All green artificial flowers are flowers

All long forged prescriptions are prescriptions

All fabricated historical knowledge is knowledge

All unemployed ex-mayors are mayors

All corrupt past chiefs of police are chiefs of police

All tasty chocolate rabbits are rabbits

All impressive steel flowers are flowers

All friendly pretend pirates are pirates

All hard plastic lemons are lemons

All ugly bronze kittens are kittens

All frightening hallucinatory spiders are spiders

All former England captains are captains

4) Filler sentences used in Experiment 4 (with quantifier appropriate to condition)

All advanced portable computers are electronic

All large furry mammals have hearts

All birds are animals

All small red triangles have three corners

All chairs can be used to sit on

All trained bottlenosed dolphins are intelligent

All bachelors are unmarried

All humans breathe with gills

All scientists are good at abstract reasoning

All long black cars have propellers

All small silent explosions are dangerous

All actors are strong

5) Filler sentences used for Experiments 5 and 6

Unmodified is more true	*Modified is more true*
All beautiful unreal fish use gills to breathe	All dry paper bags are flammable
All carved wooden birds fly*	All fresh green cucumbers are crunchy.
All friendly pretend pirates are criminals	All funny children’s books are illustrated
All frightening hallucinatory spiders spin webs	All large starved alligators are dangerous
All green artificial flowers are organic	All long gold chains are valuable
All hard plastic lemons are edible	All new luxurious cars are expensive
All miniature toy elephants drink water	All old sewer rats carry diseases
All plastic replica guns are dangerous	All ripe apples are sweet
All tasty chocolate rabbits run fast	All square medium-sized stones are used for construction*
All ugly bronze kittens purr	All sweet fruit pies are eaten for desert
All unconvincing mock lions are carnivores	All thin glass bottles are fragile

* replaced in Experiment 6 by

All carved wooden birds hatch from eggs

All large quarried stones are heavy

Author notes

The authors acknowledge the financial support of City University, London, and Erik and Gurli Hultengrens's Foundation for Philosophy (Sweden). We wish particularly to thank John Barry, Herbert Clark, Zachary Estes, Daniel Heussen, Johannes Persson, Steven Sloman, Julia Tanney, Lena Wahlberg and Annika Wallin for fruitful discussions of the research and comments on the manuscript. The two authors contributed equally to the research. Address enquiries to James A. Hampton, Psychology Department, City University, Northampton Square, London EC1V OHB, UK. Email: hampton@city.ac.uk.

Footnotes

1. The term “fallacy” is not intended to imply that it is irrational to violate these laws, or even that people generally try to act in accordance with them. These effects are fallacies only in terms of the laws of logic and statistics. If no further motivation for acting in accordance with these laws exists, it could very well be rational to break them. As Gilbert Harman (2002) suggests, if you are starving and have just discovered an inconsistency in your beliefs, the rational thing to do might not be to try to resolve the inconsistency but to find something to eat (unless of course you have reason to suppose that the inconsistency in your beliefs is the cause of your starvation in the first place).
2. Participants were randomly assigned to Experiments 1 or 2, which were run at the same time, and in the same group testing situation, so a comparison between the experiments is fully justified.

Table 1

Total frequencies and percentages of combined answers to sentence pairs in Experiment 2

		Modified sentence MN	
		Yes	No
Unmodified Sentence N	Yes	154	56
		55%	20%
	No	21	47
		8%	17%

Table 2

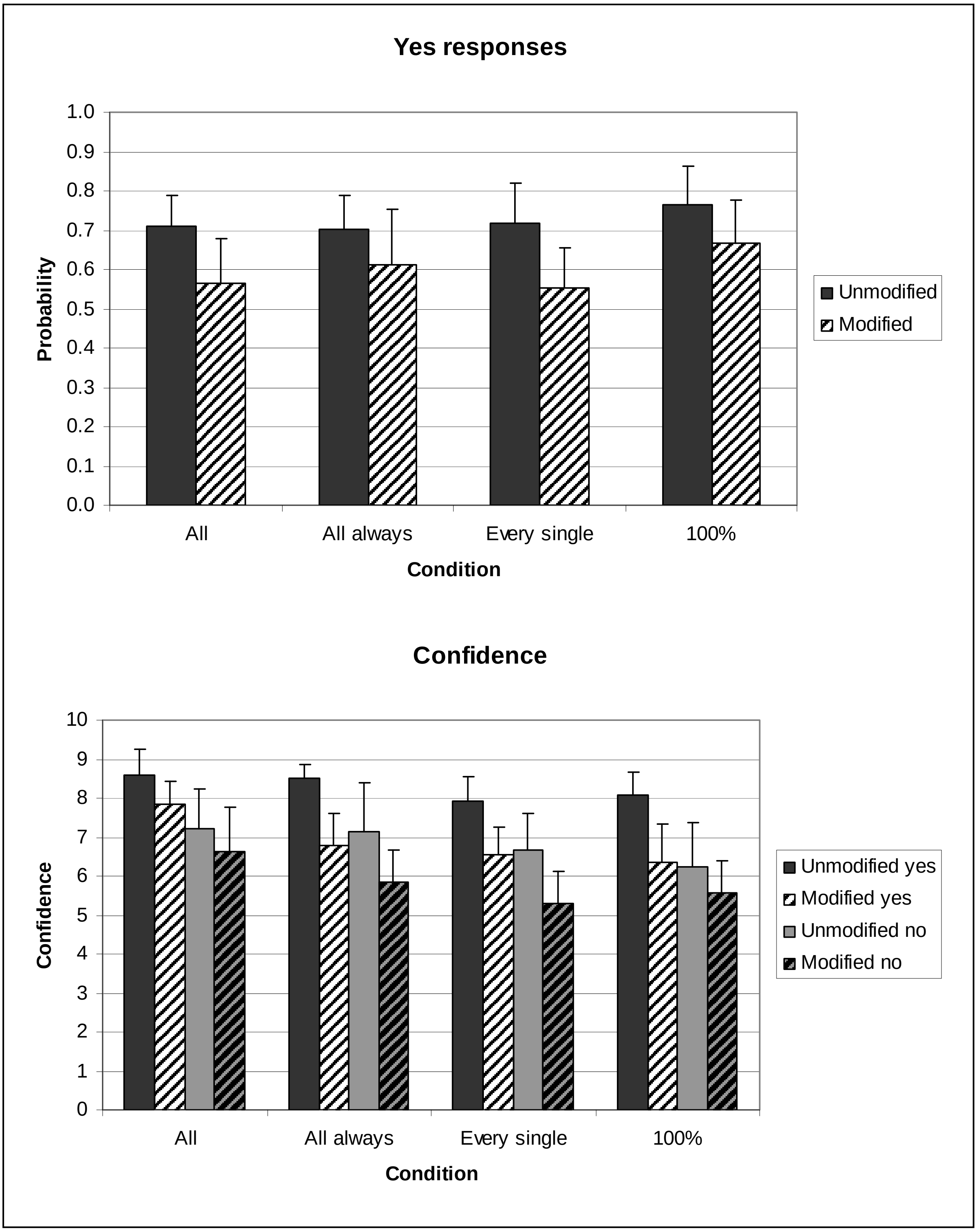
Total frequencies of combined answers to sentence pairs in Experiment 6

		Modified sentence MN	
		Yes	No
Unmodified Sentence N	Yes	318	64
		60%	12%
	No	21	126
		4%	24%

Figure captions

Figure 1

Proportion of yes responses for unmodified and modified sentences (top panel), and mean confidence ratings (bottom panel) for yes and no responses separately in Experiment 4. Error Bars show 95% Confidence Intervals.


Yes responses
Probability
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
All
All always
Every single
100%
Condition
Unmodified
Modified
Confidence
Confidence
10
9
8
7
6
5
4
3
2
1
0
All
All always
Every single
100%
Condition
Unmodified yes
Modified yes
Unmodified no
Modified no