Familiarity and Plausibility in Conceptual Combination: Reply to Gagné and Spalding (2006)

By: Gregory L. Murphy and Edward J. Wisniewski

Murphy, G.L., & Wisniewski, E.J. (2006). Familiarity and plausibility in conceptual combination: Reply to Gagné and Spalding (2006). *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32*(6): 1431–1442. DOI:10.1037/0278-7393.32.6.1431.

Abstract:

Wisniewski and Murphy (2005) suggested that the apparent effects of relation frequency in Gagné and Shoben's (1997) conceptual combination experiments could be explained by differences between the familiarity and plausibility of their stimuli (noun-noun phrases). However, Gagné and Spalding (2006) argued that our measures of plausibility and frequency are both sensitive to relation frequency. They also suggested that the stimuli were mostly novel such that differences in familiarity could not explain the Gagné and Shoben findings. We focus on the theoretical rationale for the plausibility and familiarity variables, arguing that the original interpretation of our findings is correct.

Article:

We welcome the chance to reply to Gagné and Spalding's (2006) response to our critique of the original Gagné and Shoben (1997) experiments. Although we still do not see eye to eye with their view, we believe that this exchange makes clearer how the two views differ. We focus on the theoretical rationale behind the two main variables that we raised in our earlier critique, familiarity of the phrases and plausibility of their interpretations. Gagné and Spalding provide a different interpretation of the significance and meaning of the correlations we found between these variables and their results, and we will argue that our original interpretations are correct.

FREQUENCY

In our original article (Wisniewski & Murphy, 2005), we asked subjects to evaluate whether they had ever seen or heard the Gagné and Shoben phrases before. The results showed that the phrases in their HH and HL conditions (those easiest to understand) in Experiment 1 were familiar 44% of the time, and the phrases in the LH condition (those hardest to understand) were familiar 25% of the time. To provide a less subjective estimate of familiarity, we also used Google to confirm these differences. In their reply, Gagné and Spalding (p. xx) argue that the number of Google hits was very low: “the majority of the phrases used in Gagné and Shoben were indeed novel with the frequency being zero or close to zero… This casts doubt on Wisniewski and Murphy's (2005, p. 170) suggestion” that subjects were only retrieving familiar combinations.

It is difficult to say from Google or text sources how often in an absolute sense people encounter these phrases in everyday life, since some of these sources are seldom accessed, and everyday conversation is not included. In contrast, our familiarity judgments were provided by undergraduates who were similar to those of Gagné and Shoben's in terms of education, age, and being native speakers of English. Sources such as Google are much better at identifying relative frequency, which was the basis of our argument.

However, if we accept Gagné and Spalding's suggestion that the frequencies are very low, then we believe that the *opposite* conclusion should be drawn. Because frequency has effects at roughly a log scale, frequency differences for low-frequency items are much larger than the equivalent differences for high-frequency items. The first few trials of practice in a new task or on a new item have the largest effect, across a wide variety of cognitive and motor tasks (e.g., Newell & Rosenbloom, 1981). In the domain of lexical decision, small differences at the lowest frequencies have a huge effect (e.g., see Murray & Forster, 2004, Table 1). Thus, if our subjects had encountered the HH or HL phrases a few times in their lives and the LH phrases not at all, this would be the ideal situation for frequency effects to reveal themselves. In particular, the effect of increasing exposures from 0 to 1 would be greater than an equal difference at higher frequencies. It is for this reason that we did not ask subjects for frequency estimates but simply asked them whether they had seen the item before.

Table 1: Noun-noun combinations used in Gagné and Shoben's (1997) experiments

Familiar Phrases	Unfamiliar Phrases
cream sauce	cooking treatment
municipal court	floral toy
financial crisis	plastic crisis
student magazine	honey soup
gas crisis	milk virus
plastic toy	marine antique
home remedy	thermal wheel
job conflict	moth signals
song book	sap remedy
morning prayers	rain lake
murder report	olive area
student network	servant scandal
picture story	floral property
summer money	chocolate wreath
flu pills	sugar scales
gas lamp	wood money
wood shavings	chocolate utensils
college magazine	water money

Of course, such judgments are not perfect, but they do have face validity. For example, consider the Gagné and Shoben stimuli in Table 1. Two-thirds or more of the undergraduates we asked believed that they had heard or seen the phrases on the left. In contrast, 20% or fewer of the undergraduates believed that they had heard or seen the phrases on the right. We invite readers to judge for themselves whether they have encountered the phrases on the left or those on the right more. We note that every phrase on the left had a lower response time in Gagné and Shoben's experiments than every phrase on the right.

Gagné and Spalding report that there is a nonsignificant correlation between familiarity and relation frequency in phrases, which contradicts our “claim that relation availability and familiarity were confounded” (p. xx). However, this is stretching our actual claim, that the three conditions that Gagné and Shoben used differed systematically in their frequency (which we

documented), into a much stronger claim that there is a general relationship between frequency and their theoretical variable. We don't know if there is such a relationship, in part because we are not attempting to explain the relation strength variable.

Gagné and Spalding suggested that our participants based their familiarity judgments of a phrase on how *quickly* an interpretation came to mind. To provide evidence for this view, they performed a priming experiment that manipulated the availability of the interpretation. A target phrase was immediately preceded by a prime phrase that had a similar or different relation connecting the nouns. The participant's task was to judge whether the phrase was familiar. Participants were about 6% more likely to judge a phrase as familiar when it was preceded by a prime with a similar interpretation. To our mind, these findings do not provide strong evidence that familiarity judgments more generally are based on how quickly an interpretation comes to mind. First, the absolute size of the effect is small given the immediate priming. If the effect is only 6% when the word is primed by the preceding phrase, how could it account for the much larger differences that we discovered in Gagné and Shoben's conditions, when there was no such prime, and related phrases would have occurred minutes, hours, or days earlier? Second, the study does not provide evidence for Gagné and Spalding's argument that ease of processing is the causal variable. If participants have just read a prime phrase with a similar interpretation as the target phrase, the former could well seem familiar regardless of how quickly its interpretation comes to mind. We believe that readers will agree that the familiar phrases of Table 1, like *cream sauce* and *municipal court*, are familiar not simply because they are understood quickly, but because they are very common in everyday experience and remain familiar even long after they have been initially understood. Similarly, we are convinced that *sap remedy* is unfamiliar not simply because it is slow to understand, but in part because it is completely novel.

PLAUSIBILITY

In our earlier article, we also argued that Gagné and Shoben's conditions were confounded with plausibility, and we provided plausibility ratings that showed similar patterns to their RT results. But Gagné and Spalding object that plausibility judgments are too similar to their sensicality judgments to stand as an independent explanation of them. They argue that "plausibility… is what theories of conceptual combination need to explain. Plausibility is not a nuisance variable to be controlled, nor is it an explanatory variable, on the pain of circularity" (p. xx). They also object to the notion that plausibility is a primitive property of combinations that does not require explanation.

We agree that plausibility is not a primitive characteristic of combinations—we had not intended to suggest that it was. However, we disagree that our plausibility measure is too similar to Gagné and Shoben's sensicality judgments and hence is circular. Gagné and Shoben selected test phrases that they *assumed were sensible,* including an equal number of filler phrases that they assumed were nonsensible (p. 75-6). What they expected to vary were the *response times* to judge the test phrases as sensible. Response time was predicted to be related to the frequency of the modifier relation that formed the basis of a phrase's interpretation. In contrast, our instructions given to subjects did not define plausibility as "whether an item had a sensible interpretation," as Gagné and Spalding say (p. xx). Our instructions did not concern the interpretations of the phrases, but emphasized the question of how well the *objects* described by the phrases were consistent with everyday life and did not ask about the phrases' interpretation.

The instructions said that subjects would read “phrases that describe things,” and then consistently referred to the plausibility or weirdness of those “things,” rather than to the plausibility (or sensicality) of the phrases. Furthermore, our study found that subjects judged some of the objects referred to by Gagné and Shoben's phrases to be very implausible whereas Gagné and Shoben assumed that all the test phrases were sensible.

To explain what plausibility ratings measure, we first consider what the variable of plausibility refers to. Plausibility is a well-known variable that has been discussed and used for many years in psycholinguistics. For example, implausible sentences are harder to interpret correctly than plausible ones (see Clark & Clark, 1977). Sentence processing research has shown that the plausibility of specific sentence parses has a strong effect on comprehension, although the mechanism by which it works is still controversial (Ferreira & Clifton, 1986; Garnsey, Pearlmutter, Myers, & Lotocky, 1997).. In these cases, *plausibility* refers to a property of semantic representations—being coherent and consistent with other knowledge. Plausibility has not been considered to be a circular property of the ease of understanding sentences, because the it derives from our knowledge of the world, not from linguistic forms.

Plausibility has long played a role in accounts of conceptual combination as well. In the first major study of conceptual combination in psychology, Gleitman and Gleitman (1970) investigated the grammatical influences on people's interpretations of three-word phrases like *black-bird house*. Their results could not be accounted for by the grammatical factors alone: “Two major sources of variation still remain. The *semantic plausibility* and *prior familiarity* of subparts of the stimuli” (p. 125). Gleitman and Gleitman (1970, p. 178) also provided a schematic explanation of what makes some of these interpretations plausible or not, namely the availability of a relationship between the two words that is consistent with prior knowledge:

> It is hardly surprising that *house-bird* is semantically `easier' even though no more familiar than *house-foot*… For some generalized foot, it is difficult to imagine why it should be related to a house. Is it perhaps the foot one always puts in the house first? Or the stay-at-home foot?… Improbable. (p. 178)

Since then, many authors have referred to plausibility (or consistency with prior knowledge) as a constraint on interpreting conceptual combination, including Cohen and Murphy (1982), Costello and Keane (2000), Hampton (1987, p 65-66), Murphy (1988), Smith, Osherson, Rips, and Keane (1988, p. 525), and Wisniewski (1997, p. 174).

We have covered this history to make a number of points. First, our critique of the Gagné and Shoben materials was not an ad hoc proposal, but relied on a variable that has been viewed as important to conceptual combination for 30 years. Second, if Gagné and Spalding wish to rule out plausibility as a variable, they cannot simply do so by stating that it is an alternative measure of sensicality judgments and hence not an explanatory variable. They must discuss the mass of previous literature that has referred to it and show that such explanations can be replaced by their own account. Third, given this history, if researchers wish to show that a new variable influences conceptual combination, we believe they must manipulate it without manipulating plausibility. We are not married to plausibility ratings as the only measure of this variable; other ways of measuring consistency with prior knowledge are certainly possible. Since Gagné and Shoben's

conditions were apparently confounded with plausibility, the evidence for their proposal is accordingly weakened.

Gagné and Spalding also note that our plausibility judgments were highly correlated with familiarity judgments and suggest that this correlation is "particularly surprising given that they were intended to represent independent explanatory factors" (page xx). However, we did not state or believe that these variables are uncorrelated. Familiar phrases typically refer to common objects and events, which are therefore consistent with our prior knowledge and judged as plausible. At the same time, people may judge novel phrases to be plausible because the inferred referent is consistent with prior knowledge. For example, we have never heard the phrase *avocado box* but we infer that it could refer to a plausible thing, given our prior knowledge that boxes can contain many different things (including fruits). Thus, familiarity and plausibility judgments may be correlated but also conceptually distinct. Gagné and Shoben's stimuli contained a mix of familiar phrases (e.g., *municipal court*) that were necessarily judged as plausible and novel phrases (e.g., *sap remedy*) that referred to implausible things, resulting in a correlation of these variables. A different selection of stimuli could reduce that correlation, if desired.

Can plausibility be replaced by or explained by relation availability, as Gagné and Spalding propose? We do not agree that plausibility of an interpretation is viciously circular, reflecting other variables that determine the phrase's interpretation. Rather, plausibility is based on our knowledge of what is usual and explicable in the world, independently of whatever phrase or sentence might be used to describe it. The plausibility lies in the event or object—not in words.

We can illustrate this claim with two examples from Gagné and Shoben's LH condition (the difficult one). Consider *cooking treatment*. We find it difficult to think of any plausible relation here. A treatment for cooking? Cooking is not usually treated. A medical treatment that requires the patient to cook? This is a most unusual treatment, and it is difficult to think what condition it treats. The problem, we believe, is not that we were slowed in deriving the correct relation because the modifier does not usually take that relation (as Gagné and Spalding suggest), but that even after many seconds of thinking about the phrase, we are not sure what relation links these two words in a way that is consistent with world knowledge (as Gleitman and Gleitman described in their example quoted earlier). The initial availability of the relation cannot explain this difficulty. In contrast, *servant language* has a very clear interpretation, we believe. But that interpretation is peculiar. There in fact is no language used by servants in particular, except in some metaphorical sense in which "Yes, ma'am" or "Dinner is served" is thought of as a language. So, either the interpretation is not consistent with everyday knowledge, or else *language* must be construed in an unfamiliar way to discover a plausible relation. Again, even when we have plenty of time to think about what the intended relation is, the phrase is peculiar because it is not consistent with our understanding of the world. Perhaps such items explain the fact that Gagné and Shoben's subjects judged the phrases in the LH condition as nonsensical as often as 22% of the time (Experiment 3, unequal relation items). (Recall that Gagné and Shoben selected all of their phrases to have sensible interpretations.)

Finally, Gagné and Spalding point out that plausibility judgments may be rather similar to the speeded sensicality judgments that were Gagné and Shoben's dependent measure; therefore,

plausibility “cannot carry the theoretical weight in this particular instance” (p. xx), because it is not independent of the measure. However, this argument puts the cart before the horse. The tested phrases differed in how consistent they were with people's prior knowledge, which we have argued is reflected in our plausibility judgments. It seems obvious that this factor will influence sensicality judgments. That is not a problem with our instructions for the plausibility judgments, but it is instead a problem for the use of sensicality judgments if one is using them to obtain evidence for a different variable. Careful control of world knowledge (by plausibility judgments or another appropriate method) is necessary if one wants to argue that a different factor is influencing sensicality judgments. To put it more strongly, we believe that plausibility would influence almost *any* task that requires people to interpret conceptual combinations, because people must check their interpretations against world knowledge. So, our explanation is not specifically tailored to the sensicality judgment task.

In short, we believe that some of these phrases do not have an obvious relation that seems correct, and others do have an obvious relation, but the resulting interpretation is not consistent with everyday knowledge. Such phrases are difficult to interpret. That claim is completely independent of CARIN's claim that modifiers are associated with preferred relations, rather than being circularly based on it. Indeed, as we pointed out in our previous article, plausibility explained some of the significant effects in Gagné and Shoben's paper that were inconsistent with CARIN's predictions. Clearly, it could not do that if plausibility were a consequence of relation frequency.

CAN CARIN EXPLAIN PLAUSIBILITY IN TERMS OF RELATION FREQUENCY?
Finally, we ask whether CARIN could truly explain plausibility by modifier preference. The empirical basis for CARIN is that when phrases use a relation that is less preferred (by the modifier), processing is slowed, due to competition from the other relations. However, why is the unpreferred relation *ever* used? Presumably, it is not just random chance, because Gagné and Shoben listed some phrases as having an unpreferred relation as the best interpretation. For example, although *mountain* normally prefers to be used to modify location, in *mountain magazine*, they judged that it referred to a magazine *about* mountains.

In its quantitative modeling, CARIN simply accepts the “correct” relation as an input—it does not decide which relation is the correct one (and, in fact, no computational model can do that at this time). But when people hear or read this phrase, they must decide for themselves which relation is correct, and, according to CARIN, sometimes they decide that the less frequent relation is correct. Why? Why not always choose the preferred relation for that modifier? The reason is obvious: Because the preferred relation leads to a less plausible interpretation. *Mountain magazine* could mean a magazine located in/at a mountain, but that is a strange thing. Magazines do not have any particular domicile, and the present location of a magazine does not seem an important aspect of it worth drawing attention to. However, there are magazines about many different topics, and it is possible that a magazine could focus on mountains. In short, we agree with Gagné and Shoben's interpretation of this phrase. But their model must admit of some other variable that overrules the preferred relation.[1]

If we are right in our interpretation, then CARIN cannot use relation frequency to explain plausibility effects as Gagné and Spalding suggest, because then it would have no way to account

for people systematically choosing an unpreferred relation for some combinations. That is, if people decide the plausibility of phrases involving *mountain* by using relation frequency, then *mountain magazine* would be understood as a magazine that inhabits a mountain, because that is the most frequent relation. But Gagné and Shoben themselves have given a different interpretation and have based their predictions on that interpretation.

GAGNÉ AND SPALDING'S MULTIPLE REGRESSION ANALYSES

Gagné and Spalding found that after a number of predictors are entered into a regression model, the addition of modifier relation strength still accounts for variance in Gagné and Shoben's (1997) RT results. However, this additional variance is generally quite small, for example, only 8% and 3%, in Experiments 1 and 3, when the predictors include subjective frequency. When the predictors include plausibility, the variance accounted for is only 6% and 4%. Further, given the limitations of multiple regression, one cannot determine the unique variance accounted for by a predictor unless it is uncorrelated with the other predictors. A controlled experiment (as we recommended) would resolve these issues.

Gagné and Spalding criticized the use of relation frequency in our analyses, suggesting that modifier relation *strength* was more valid, as it reflects the competition among relations assumed by CARIN. However, Gagné and Shoben (1997) also used a dichotomized relation frequency measure in their analyses of variance and the rank of the relation frequency in two of their regression analyses (p 76). It is possible that relation strength is a better predictor, explaining the differences between the regression analyses. However, relation strength is also more highly correlated with the frequency and plausibility judgments (see Gagné & Spalding's Table 1), perhaps reducing the ability to measure their effects. Furthermore, the main result of transforming relation frequency into relation strength is to spread out the scores of the less frequent relations and reduce the differences among the higher-scoring relations.[2] We are not sure why this represents "the notion of competition among relations" (p. xx). Thus, we are not clear on exactly why the regressions give different results or how to interpret these differences. Although the new regression does provide some evidence for the use of relation strength, the problems of correlated variables reinforce our recommendation that a new experiment should be performed in which frequency are plausibility are controlled as much as possible.

CONCLUSION

What would convince us that relation frequencies are involved in people's interpretations of conceptual combinations? We will simply summarize our desiderata for a truly convincing empirical demonstration of the importance of relation frequency (see our earlier article for more detail). First, the judgments about the acceptability and interpretations of the phrases should be done by subjects, and not by researchers alone. Second, relation frequency should not be measured within a small set of phrases, but should be based on a larger corpus that reflects the language use of the subject population. Third, we have grave concerns about the basic relations from Levi's (1978) linguistic analysis that were the basis for past predictions, because phrases that have the same relation often appear to have very different meanings (e.g., *for* can mean to prevent or to aid; see Downing, 1977). A more data-driven analysis of relations would be of great benefit to the field. Finally, as we have argued throughout this interchange, it is important to hold constant other variables that might be correlated with relation frequency. We do not believe that it is impossible to derive a set of stimuli in which familiarity and plausibility are not

correlated with relation strength. (So, we do not argue that relation frequency is in fact just the same thing as one of these other variables, as Gagné and Spalding suggest. They are theoretically quite distinct.) Any theory that hopes to propose a new and interesting variable as partly explaining conceptual combination must first rule out the old and less interesting variables in its empirical test.

ACKNOWLEDGMENTS

The writing of this article was supported by NIMH grant MH41704 and NSF grant BCS-9975198. Please address correspondence concerning this article to Gregory L. Murphy, Department of Psychology, New York University, 6 Washington Pl. 8th floor, New York, NY 10011.

FOOTNOTES

[1]Gagné and Shoben briefly entertained the possibility that a "relation may be applied to the phrase to see if it provides a plausible interpretation" and that "Plausibility may be based on whether the head noun has the correct properties for the relation suggested by the modifier" (p. 73). However, their experiments and model did not directly address this process. Instead, they explicitly focused on "whether the availability of a… relation influences the ease with which a combined concept can be interpreted" (p. 74) and on competition between these relations.

[2]Gagné and Shoben (1997, pp. 81-2) work through an example of how relation strength differs from relation frequency (proportion). Their examples show that using a ratio measure (the proportion of times the word is in the correct relation divided by the sum of proportions of the top four relations) tends to magnify the larger proportions. They then report that they first raised the proportions to a negative exponential before calculating the ratio. By our calculations (they do not work their examples through this stage), this has the effect of bunching together moderate and large proportions but making low proportions much worse. For example, for the proportions .82, .10, .02, and .01 for the word *mountain*, the strength values appear to be: ~0, .02, .40, and .58 (lower values are better). Clearly, the biggest difference is not between the dominant relation (occurring .82 of the time) and the others, but between the .10 and .02 relations. It is not clear why competition between the different relations should yield this result.

REFERENCES

Clark, H. H., & Clark, E. V. (1977). *Psychology and language*. New York: Harcourt Brace Jovanovich.

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

Costello, F. J., & Keane, M. T. (2000). Efficient creativity: Constraint-guided conceptual combination. *Cognitive Science, 24,* 299–349.

Downing, P. (1977). On the creation and use of English compound nouns. *Language, 53,* 810–842.

Ferreira, F., & Clifton, C. (1986). The independence of syntactic processing. *Journal of Memory and Language, 25,* 348–368.

Gagné, C. L., & Shoben, E. J. (1997). Influence of thematic relations on the comprehension of modifier–noun combinations. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23,* 71–87.

Gagné, C. L., & Spalding, E. J. (2006). Relation availability was not confounded with familiarity or plausibility in Gagne´ and Shoben (1997): Comment on Wisniewski and Murphy (2005). *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32,* 1431–1437.

Garnsey, S. M., Pearlmutter, N. J., Myers, E., & Lotocky, M. A. (1997). The contribution of verb bias and plausibility to the comprehension of temporarily ambiguous sentences. *Journal of Memory and Language, 37,* 58–93.

Gleitman, L. R., & Gleitman, H. (1970). *Phrase and paraphrase: Some innovative uses of language.* New York: Norton.

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

Levi, J. N. (1978). *The syntax and semantics of complex nominals*. New York: Academic Press.

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

Murphy, G. L. (1990). Noun phrase interpretation and conceptual combination. *Journal of Memory and Language, 29,* 259–288.

Murray, W. S., & Forster, K. I. (2004). Serial mechanisms in lexical access: The rank hypothesis. *Psychological Review, 111,* 721–756.

Newell, A., & Rosenbloom, P. S. (1981). Mechanisms of skill acquisition and the law of practice. In J. R. Anderson (Ed.), *Cognitive skills and their acquisition* (pp. 1–55). Hillsdale, NJ: Erlbaum.

Smith, E. E., Osherson, D. N., Rips, L. J., & Keane, M. (1988). Combining prototypes: A selective modification model. *Cognitive Science, 12,* 485–527.

Wisniewski, E. J. (1997). When concepts combine. *Psychonomic Bulletin & Review, 4,* 167–183.

Wisniewski, E. J., & Murphy, G. L. (2005). Frequency of relation type as a determinant of conceptual combination: A reanalysis. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31,* 169–174.