Opinion

# The Sapir-Whorf hypothesis and inference under uncertainty

Terry Regier* and Yang Xu

**The Sapir-Whorf hypothesis holds that human thought is shaped by language, leading speakers of different languages to think differently. This hypothesis has sparked both enthusiasm and controversy, but despite its prominence it has only occasionally been addressed in computational terms. Recent developments support a view of the Sapir-Whorf hypothesis in terms of probabilistic inference. This view may resolve some of the controversy surrounding the Sapir-Whorf hypothesis, and may help to normalize the hypothesis by linking it to established principles that also explain other phenomena. On this view, effects of language on nonlinguistic cognition or perception reflect standard principles of inference under uncertainty.** © 2017 Wiley Periodicals, Inc.

**How to cite this article:**
*WIREs Cogn Sci* 2017, 8:e1440. doi: 10.1002/wcs.1440

## INTRODUCTION

The Sapir-Whorf hypothesis holds that the semantic categories of one's native language influence thought, and that as a result speakers of different languages think differently. This idea has captured the imaginations of many, and has inspired a large literature. However the hypothesis is also controversial, for at least two reasons, one theoretical and the other empirical. Theoretically, the hypothesis is controversial because it appears to challenge the widely-held belief that human thought rests on a universal cognitive foundation. Such a universalist assumption is essential to many theories, and consistent with many findings, so a challenge to it can appear intellectually nihilistic. Empirically, the hypothesis is controversial because although there are findings that support it, some such findings have an inconsistent record of replication. This empirically mixed picture means that it is not always clear whether the hypothesis is firmly supported. In both these respects, then, the Sapir-Whorf hypothesis occupies a troubled and contested position.

Considering the Sapir-Whorf hypothesis through the lens of probabilistic inference has the potential to resolve both of these issues. Recent work has cast the hypothesis in probabilistic terms, and has accounted for data supporting the hypothesis while retaining the assumption of a universal groundwork for cognition. An important insight provided by this perspective is that cognitive *uncertainty* may play a central role in modulating effects of language on cognition. Although it is natural to ask in simple yes-or-no terms whether the Sapir-Whorf hypothesis is supported, it may be more profitable to instead think in terms of a continuum: from language not affecting cognition, to affecting it somewhat, to affecting it strongly. According to the perspective explored here, uncertainty will determine where along that continuum a given situation falls. For example, you may be thinking of an object and find that some of its details are not mentally available to you, whether because of fading memory, fatigue, or some other factor inducing uncertainty in your mental representation. In such circumstances, the mental uncertainty essentially opens the door to language to fill in some of the missing elements, and there should be a relatively strong effect of language. In contrast, when relevant nonlinguistic information is comparatively certain, when object details are already clearly mentally available, there is little missing information for language to supply, so there should be little or no effect of language.

*Correspondence to: terry.regier@berkeley.edu

Department of Linguistics, Cognitive Science Program, University of California, Berkeley, CA, USA

Conflict of interest: The authors have declared no conflicts of interest for this article.

  

Essentially, uncertainty may be thought of as providing a kind of 'cognitive control knob' that sweeps continuously from no effect of language on cognition, to stronger such effects. In this way, the inclusion of uncertainty as a proposed mediating force has the potential to explain some of the inconsistency of the empirical record.

We lay out this argument primarily with respect to color cognition, because relevant research connecting the Sapir-Whorf hypothesis with probabilistic inference has been focused there. In what follows, we first briefly review the relevant empirical terrain, namely tests of the Sapir-Whorf hypothesis that highlight the sources of controversy referenced above, focusing on the domain of color. We then review the relevant theoretical terrain, namely probabilistic models that have treated related phenomena involving the integration of multiple cues, highlighting the central role of cognitive or perceptual uncertainty in these models. Finally, we review recent work that has brought these two strands of research together by exploring the Sapir-Whorf hypothesis in terms of probabilistic inference.

## THE SAPIR-WHORF HYPOTHESIS, WITH EMPHASIS ON COLOR

The Sapir-Whorf hypothesis is captured in the following passage from Sapir[1]:

> Human beings do not live in the objective world alone, nor alone in the world of social activity as ordinarily understood, but are very much at the mercy of the particular language which has become the medium of expression for their society. It is quite an illusion to imagine that one adjusts to reality essentially without the use of language and that language is merely an incidental means of solving particular problems of communication or reflection. The fact of the matter is that the 'real world' is to a large extent unconsciously built up on the language habits of the group. (p. 209)

and the following passage from Whorf[2]:

> We dissect nature along lines laid down by our native languages. The categories and types that we isolate from the world of phenomena we do not find there because they stare every observer in the face; on the contrary, the world is presented in a kaleidoscopic flux of impressions which has to be organized by our minds—and this means largely by the linguistic systems in our minds. (p. 213)

Even those antagonistic to this proposal have acknowledged that its 'implication is heavy,'[3] and the hypothesis has attracted attention for many years, oscillating in character from enthusiastic to dismissive and back again.[4] The empirical literature concerning the hypothesis has explored the effects of language on cognition about color, number, spatial frames of reference, object individuation, and the understanding of false belief, among other domains. Here we provide a brief review of this literature specifically in the domain of color, and specifically with respect to the sources of controversy sketched above. More broad-ranging reviews of the literature as a whole are available elsewhere.[5–9]

Color is a classic testing ground for the Sapir-Whorf hypothesis. It is an accessible domain, relatively easily described and measured, and languages differ in the ways they divide it into categories. Early findings in this domain by Brown and Lenneberg were taken to support the Sapir-Whorf hypothesis.[10] Subsequently however Berlin and Kay noted that there are universal tendencies in color naming patterns across unrelated languages,[11] suggesting a universal cognitive or perceptual basis beneath these linguistic regularities—inconsistent with any reading of the Sapir-Whorf hypothesis that would deny the existence of such a universal basis. This universalist interpretation was further supported by Rosch-Heider's finding of similarities in color cognition across speakers of languages with dissimilar color naming systems,[12,13] and for many years (the late 1960s through the late 1990s) this view dominated. During this period, investigation of the Sapir-Whorf hypothesis itself sometimes assumed a universal basis underlying the effect of language,[14] and more generally the emphasis of the time was on universals rather than variation.

The pendulum began to swing back to greater interest in the Sapir-Whorf hypothesis in the late 1990s and 2000s[15–21] and in the color domain this was initiated in large part by the work of Roberson and colleagues.[15–17] They attempted to replicate Rosch-Heider's earlier finding of similarities in color cognition across speakers of different languages—and failed to do so. Instead, they found differences in color memory that were congruent with the rather different color naming systems of the two languages they examined: Berinmo, a language of Papua New Guinea, and English.[15,17] For example, in a color memory task, participants showed best performance for pairs of colors that were named differently in their native language, yielding language-congruent differences in color memory. A later study obtained similar findings

  

comparing speakers of English and Himba, a language of Namibia.[19] These findings, together with subsequent similar findings from other groups,[20,21] are widely taken as support for the Sapir-Whorf hypothesis in the domain of color.

At the same time, the two tensions that have surrounded the Sapir-Whorf hypothesis generally have continued to do so in the color domain. With respect to universality, during the period in which greater evidence began to accumulate suggesting that language can affect color cognition, evidence also accumulated supporting cross-language universal tendencies in color naming.[22–25] These parallel developments underscored the need for a way to explain effects of language on cognition (for which we have evidence) while retaining a universal cognitive or perceptual foundation for color categorization (for which we also have evidence). With respect to replication, while several studies have found effects of language on color memory and perception, others have failed to do so,[26,27] as reflected in the title of a recent article: 'Whorfian effects on colour memory are not reliable.'[27] Thus, despite the substantial evidence supporting the Sapir-Whorf hypothesis in the color domain, the picture remains unsettled, both theoretically and empirically.

## PROBABILISTIC INFERENCE

A way to resolve this unsettled state of affairs is suggested by an apparently unrelated phenomenon: how humans judge the size of an object by integrating cues from vision and touch. Ernst and Banks[28] noted:

> When a person looks at an object while exploring it with their hand, vision and touch both provide information for estimating the properties of the object. Vision frequently dominates the integrated visual-haptic percept, […] but in some circumstances the percept is clearly affected by haptics […]. (p. 429)

Broadly speaking, this empirically mixed situation mirrors the one we have seen concerning effects of native language on cognition and perception: sometimes effects of language are found, but sometimes not. This parallel motivates a closer look at the proposed resolution to this situation in the case of vision and touch, to determine whether the same general principles may also apply to effects of language. Box 1 presents a schematic overview of these principles, and a roadmap for the remainder of the paper.

### BOX 1

#### OVERVIEW

The general idea behind the research we review is illustrated in Figure 1. A stimulus $S$ is observed, and it leaves behind two cues $c_1$ and $c_2$ in the observer's mind; the observer then combines these two cues to obtain an estimate $\hat{S}$ of the original stimulus $S$. We consider two variants of this general schema.

#### Perceptual Cue Integration

Both cues are perceptual, and must be fused to yield an integrated percept. For example, visual and haptic cues may be integrated to yield an estimate of object size.

#### Category Adjustment Model

One of the two cues is a representation of the particular object observed, and the other represents the general category in which it falls. The estimate combines information from these two representations. The Sapir-Whorf hypothesis may be cast in these terms when the category is given by language.

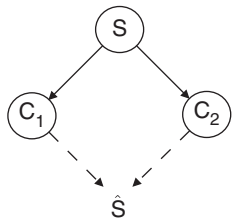


**FIGURE 1** | A stimulus $S$ produces two cues $c_1$, $c_2$. From these cues, we obtain an estimate $\hat{S}$ of the original stimulus.

### Perceptual Cue Integration

Perceptual cue integration is often cast in terms of probabilistic inference,[28–32] which provides a normative standard to which human behavior can be compared. According to this standard, when combining information from different sources, the optimal way to combine them is to average the sources together, weighted by the certainty of each source of information. For example, in the vision-and-touch context, Ernst and Banks[28] proposed that humans combine visual and haptic cues to object size in line with this principle, and provided empirical evidence supporting that proposal. They

  

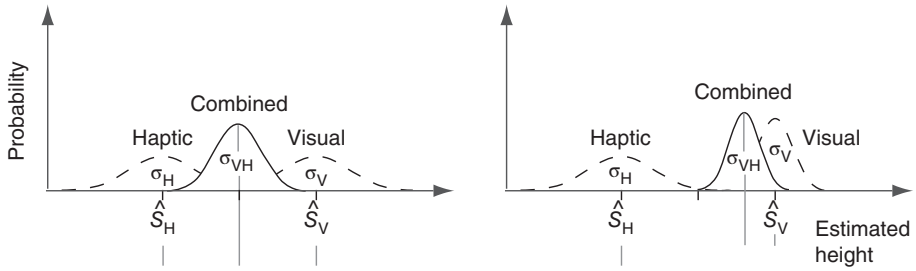


**FIGURE 2** | Probabilistic cue integration, weighted by cue certainty. Probability densities for haptic (*H*) and visual (*V*) cues are shown in dashed outline, and their combination (*VH*) is shown in solid outline. When the haptic and visual cues are equally certain (left panel), the combination is centered evenly between them. When the visual cue is more certain (right panel), the combination is centered nearer to the visual than the haptic cue. (Adapted with permission from Ref 28. Copyright 2002 Nature Publishing Group)

argued that the reason vision often dominates in the integrated percept is that visual cues tend to be more certain than haptic ones, and cues should be weighted by certainty. This certainty-weighting notion is illustrated in Figure 2, and Box 2 shows how this idea follows from general principles.

## Category Adjustment Model

The category adjustment model is a class of probabilistic model originally advanced by Huttenlocher, Hedges, and colleagues to explain category effects on memory.[33–36] According to this model, experiences are mentally encoded in memory at two levels: (1) a fine-grained memory of a *particular* object or event experienced, and (2) a designation of the general *category* in which the object or event fell, both represented as probability distributions. For example, consider the simple task of seeing a dot located somewhere inside a circle, and then reconstructing from memory where the dot was.[34] If the circle is implicitly categorically divided into four quadrants defined by the horizontal and vertical axes, then the memory of the dot's location would be composed of a fine-grained and unbiased but inexact memory of the actual location (the particular), and the quadrant of the circle within which it fell (the category). To reconstruct the dot's location from memory, one would probabilistically combine evidence from these two representations following the general principles we have just seen for probabilistic cue integration, biasing the reconstructed memory toward the center of the category, such that a more uncertain memory for the particular would yield a greater effect of the category. The category adjustment model can be seen as a form of probabilistic cue integration in which one of the cues is a category, rather than a perceptual cue from another modality.

To illustrate these ideas further, and to highlight especially the role of uncertainty, consider the task of remembering the date when a certain event occurred. This task was studied by Huttenlocher and colleagues[33] by asking students at a university to recall when specific movies were shown on campus. The university in question divided the academic year into fall, winter, and spring quarters, so these quarters were presumed to be the relevant temporal categories. Consistent with a category adjustment model, they found that students exhibited category-consistent biases in their memories of when a given movie was shown. Specifically, students tended to remember movies as occurring nearer the middle of the academic quarter than they had actually occurred. Importantly, this bias was greater for movies further in the past—consistent with a gradual fading of memory for the specific date (increasing uncertainty for the fine-grained particular) leading to greater influence of the academic quarter (the category). The same principles have also been used to explain category effects on judgments of the location of objects in the world,[37] object size,[35,38] shades of gray, and line length,[35] and on the perception of vowels in speech.[39] The mathematical details of the various models sometimes differ, but the general principles are the same, and sometimes even the details are the same. For example, in Feldman et al.'s[39] model of category effects on vowel perception, the critical notion of certainty-weighting is captured in an equation (their Eq. (6), p. 758) that is identical to Eq. (6) of Box 2.

# THE SAPIR-WHORF HYPOTHESIS AND PROBABILISTIC INFERENCE

The principles we have just reviewed have the potential to resolve some of the controversy surrounding the Sapir-Whorf hypothesis. That hypothesis holds that native-language categories shape cognition, and the principles we have reviewed suggest a specific way in which this might happen. Importantly, they

  

BOX 2

## WEIGHTING BY CERTAINTY IN PROBABILISTIC CUE INTEGRATION

Ernst and Banks'[28] formulation of cue integration may be derived as follows. Assume that $c_1$ and $c_2$ are perceptual cues that were produced by some stimulus $S$. Assume that each of these three variables ($S$, $c_1$, $c_2$) is specified by a probability distribution along a single continuum (e.g., size), such that the width of each distribution captures cognitive/perceptual uncertainty about the value of the corresponding variable on that continuum. Assume that the cues $c_1$ and $c_2$ are independent and normally distributed with means $\mu_1$, $\mu_2$ and variances $\sigma_1^2, \sigma_2^2$. Finally, assume that there is no *a priori* preference for any particular value for $S$, such that the prior $p(S)$ is uniform. Given these representations and assumptions, cue integration can be cast as the problem of inferring what the stimulus $S$ must have been, given the cues $c_1$ and $c_2$:

$$p(S|c_1,c_2) \propto p(c_1,c_2|S)p(S) \quad \text{[Bayes' rule]} \tag{1}$$

$$= p(c_1|c_2,S)p(c_2|S)p(S) \tag{2}$$

$$= p(c_1|S)p(c_2|S)p(S) \quad \text{[because } c_1 \text{ is independent of } c_2 \text{ given } S\text{]} \tag{3}$$

$$= p(c_1|S)p(c_2|S) \quad \text{[because } p(S) \text{ is uniform]} \tag{4}$$

$$= N(\mu_1,\sigma_1^2)N(\mu_2,\sigma_2^2) \quad \text{[because } c_1, c_2 \text{ are normally distributed]} \tag{5}$$

The product of these two normal distributions is itself proportional to a normal distribution with:

$$\mu_{12} = \frac{\sigma_2^2}{\sigma_1^2+\sigma_2^2}\mu_1 + \frac{\sigma_1^2}{\sigma_1^2+\sigma_2^2}\mu_2 \quad \text{and} \quad \sigma_{12}^2 = \frac{\sigma_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2} \tag{6}$$

This yields weighting by certainty: in the combined mean $\mu_{12}$, each cue's mean is weighted proportionally to the other cue's variance, so the lower variance (more certain) cue receives more weight in the combination. The variance $\sigma_{12}^2$ of the combination is no greater (and generally less) than the variance of either cue alone. We have laid this reasoning out in some detail here because it is relevant to the probabilistic framing of the Sapir-Whorf hypothesis, as we will see below.

do so while retaining the assumption of a universal basis for cognition and perception. For example, cue integration models explain how a haptic cue might bias the influence of a visual one, while positing a single undistorted perceptual space underlying both cues and the resulting integrated percept (recall Figure 2). Similarly, in the movies example above, bias in memory reflects the institutionally-defined categories of the academic quarter system—and in this model, as in cue integration models, the underlying mental timeline is not itself biased or distorted in any way. Instead, the memory bias arises from an interaction between a proposed universal mental timeline and conventionally defined temporal categories represented relative to it. Different categories, for instance academic semesters rather than quarters, should yield different bias patterns, again without affecting the underlying universal timeline. If we simply replace the institutionally-defined categories of academic quarters and semesters with categories supplied by native language, such an account has the potential to accommodate effects of language on cognition without calling into question the important idea of a universal cognitive foundation—thus resolving one source of controversy.

The other source of controversy we have considered concerns replicability: although there are findings that indicate effects of language on cognition, they do not always replicate reliably. Considering these phenomena through the lens of a category adjustment model has the potential to resolve this tension as well. As we have seen, in such models, the effect of the category is strongest when there is substantial uncertainty in the fine-grained representation of the particular instance seen. For example, if you wish to recall a particular hue of green that you have

  

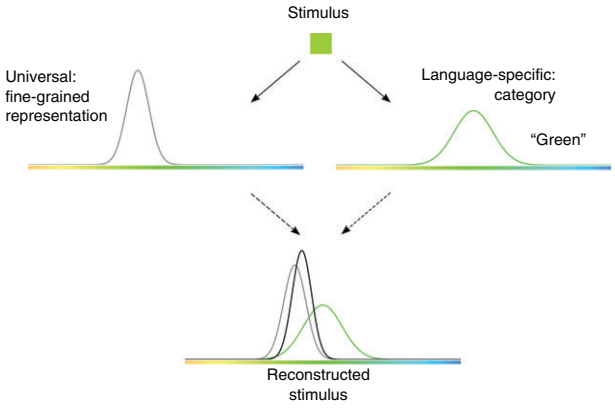


**FIGURE 3** | A category adjustment model applied to the Sapir-Whorf hypothesis in the domain of color. An observed stimulus is encoded in two ways: (1) a fine-grained representation of the stimulus itself, shown as a (gray) distribution over stimulus space centered at the stimulus' location in that space, and (2) the language-specific category (e.g., English 'green') in which the stimulus falls, shown as a separate (green) distribution over the same space, centered at the category prototype. The stimulus is reconstructed by combining these two sources of information through probabilistic inference, resulting in a reconstruction of the stimulus (black distribution) that is biased toward the category prototype. The amount of bias is determined by the uncertainty of the fine-grained representation. (Reprinted from Ref 49.)

seen, but your memory of that particular hue is uncertain, on this account your recall of it would be biased toward the center of the English linguistic color category in which it fell: green. In contrast, if there is little uncertainty in the fine-grained representation, the effect of the category would be very small, and in the limiting case the category effect may not be empirically detectable—which could explain some of the failures to replicate. Thus, the central role of uncertainty in modulating category effects in the category adjustment model offers a possible resolution of the second source of tension concerning the Sapir-Whorf hypothesis.

Figure 3 presents an overview of a category adjustment model instantiating these ideas in the domain of color. Several studies have explored ideas related to the one illustrated here. The notion that effects of language on cognition may arise from the interplay of verbal codes with perceptual representations is well-represented in the literature[14,16,40–42] and several studies have appealed to variants of the category adjustment model in this connection, in the color domain and others.[43–45] Recently, a number of studies have tested such ideas directly by comparing empirical data with the output of computational category adjustment models.[46–49]

An especially relevant recent study is that of Bae et al.[48] They investigated the relation of English color naming and color memory in U.S. undergraduates, and sought to account for their empirical findings using a category adjustment model. In their study, participants were shown a color, and attempted to identify that color on a color wheel that was presented either simultaneously with the target color, or after a delay. Other participants indicated the extensions of basic English color terms relative to the same set of colors. Consistent with the principles of the category adjustment model, they found that participants' reconstructions of colors exhibited bias patterns that reflected the categories named by English color terms, such that responses were 'biased away from category boundaries and toward category centers' (p. 744). Importantly, they also found that such category-congruent bias was stronger when participants responded after a delay than when they responded simultaneously with stimulus presentation. Delay also led to greater variance in responses, suggesting greater uncertainty in the representation of the particular color seen. Thus, these findings are consistent with the general certainty-weighting prediction of the category adjustment model. Finally, they compared their empirical reconstruction data to simulated responses by a computational category adjustment model based on the English color naming data they had collected, and obtained a good fit, as illustrated in Figure 4.

These findings support the principles of the category adjustment model in the color domain, when the relevant categories are English color terms. However, because they are based only on English, they do not directly test the core claim of the Sapir-Whorf hypothesis, which is that different category systems in different languages will lead to corresponding differences in cognition.

A recent study by Cibelli et al.[49] addressed that claim directly. They considered the cross-language color naming and memory data of Roberson and colleagues[17,19] that we reviewed above, through the lens of a computational category adjustment model. The central notion of certainty-weighting in that model was instantiated using Equation 6 of Box 2 above, as in Ernst and Banks'[28] model of cue integration and Feldman et al.'s[39] model of category effects in vowel perception.

The left panel of Figure 5 compares the color naming systems of English and Berinmo against a standard color naming grid, and also highlights the ranges of colors for which Roberson and colleagues collected color memory data. Roberson and colleagues employed a two-alternative forced choice

  

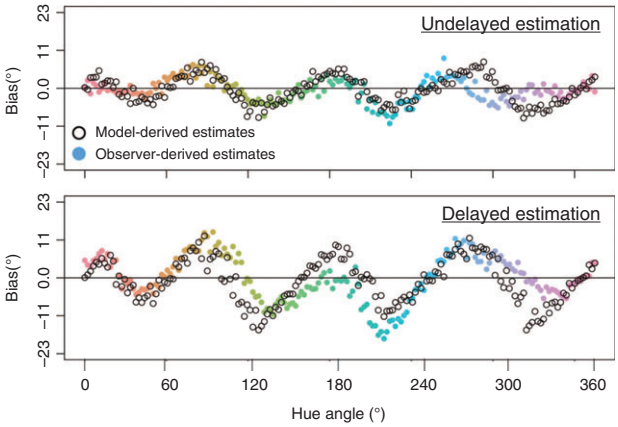


**FIGURE 4** | A category adjustment model accounts for bias patterns in color memory.[48] In each panel, the horizontal axis denotes the hue of a target color that participants saw, and the vertical axis denotes bias in memory: positive values indicate that the color in question was remembered as being a color further to the left along the horizontal axis than it actually was, and negative values indicate that the color was remembered as being a color further to the right. Colored dots denote empirical data, and black circles denote estimates provided by a category adjustment model based on English color naming. Comparison of the upper and lower panels reveals that delay (and thus greater uncertainty) produces greater category-congruent bias. (Reprinted from Ref 48.)

(2AFC) task: participants were shown a color, then shown that color again together with a different distractor color, and asked which was the color they had seen originally. The pairs of colors were chosen to lie either within or across a category boundary in the participants' native language, and as can be seen, these boundaries fall in different places for the two languages. Cibelli et al. created two category adjustment models, one with the category component determined by Berinmo color naming data, and the other with the category component determined by English color naming data. Because category adjustment models bias memories for particular stimuli toward the prototypes of native-language categories, colors on opposite sides of a native-language category boundary are pulled in opposite directions, making such cross-category pairs easier to discriminate in memory. The right panel of Figure 5 shows empirical performance on the 2AFC color memory task by Berinmo and English speakers for various pairs of colors,[17] compared in each case with the simulated responses of the native-language category adjustment model. It can be seen that the empirical results differ substantially across languages, that the best performance in each case is for color pairs that cross a native-language boundary, and that the native-language category adjustment models provide a reasonable fit to the memory data from each language. Similar results were also found when simulating findings from a study that compared color naming and cognition in English and Himba,[19] and in simulating effects of stimulus presentation order across all three languages.[50]

The findings we have reviewed in this section show that standard principles of probabilistic inference can account for data that have been taken to support the Sapir-Whorf hypothesis. In doing so, they underscore the important role that uncertainty appears to play in mediating such category effects. We have suggested that cognitive or perceptual uncertainty may help to explain why effects of language on cognition are sometimes found and sometimes not, and there are hints in the literature that are consistent with that suggestion. For example, a recent study[51] failed to find consistent category effects in a color discrimination task with participants who had had extensive prior experience with a related color perception task. In contrast, that study and another[52] did find category effects in participants who had had no or much less such prior experience. A possible interpretation of these results is that extensive prior experience—effectively a form of perceptual training—enabled participants to encode colors with greater perceptual certainty, which would account for the absence of a category effect in the experienced participants. (We thank Gary Lupyan for drawing our attention to this connection.)

## Number, and Other Domains

An important remaining question is whether these ideas will extend to other semantic domains, for example number. Language is implicated in some forms of numerical cognition and not others[53]:

> Infants and [nonhuman] animals appear to represent only the first three numbers exactly. Beyond this range, they can approximate 'numerosity,' with a fuzziness that increases linearly with the size of the numbers involved (Weber's law). … Exact arithmetic would require language, whereas approximation would not. (p. 499)

Thus the ability to represent numbers *approximately* is often taken to be universal, whereas the ability to represent large numbers *exactly*, such as the number 54, is often taken to rely on a linguistic counting system, which exists alongside the approximate number system. This view is supported by evidence documenting an inability to accurately represent exact high numerosity in speakers of languages with counting systems that lack terms for such exact high numbers[53,54]

  

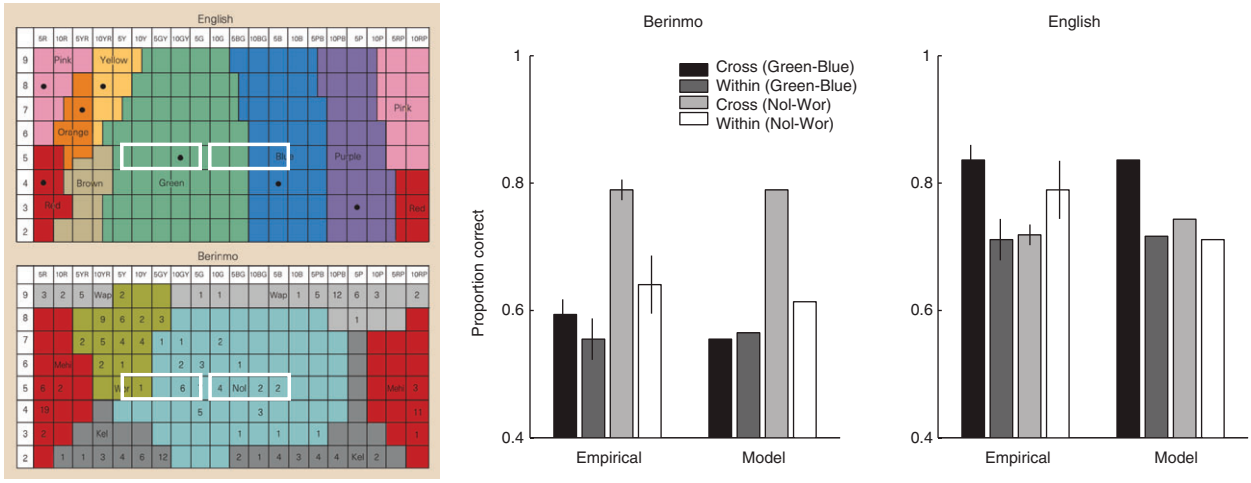


**FIGURE 5** | Left: English (top panel) and Berinmo (bottom panel) color naming systems, both mapped against a standard color naming grid in which lightness varies by row and hue varies by column. Each false-colored region represents the extension of a named color category. The white rectangles denote ranges of colors for which Roberson et al.[17] collected color memory data; these are the same ranges for the two languages. (Adapted with permission from Ref 15. Copyright 1999 Nature Publishing Group). Right: Empirical color memory performance by Berinmo and English speakers for various color pairs,[17] compared with the fit of category adjustment models based on native-language color naming.[49] Model fits are range-matched to the empirical data. (Reprinted from Ref 49.)

and by a similar inability in speakers of English whose verbal representations have been temporarily suppressed through a verbal interference task.[55] In such cases, people appear to fall back on the approximate number system for representing high numerosities.

These findings appear to be consistent, at least in broad outline, with the category adjustment model. Consider Figure 3 again, but this time imagine that the stimulus is a large number, for example, 54, that the universal nonlinguistic representation (on the left) is an approximate representation of that quantity, and that the language-specific category (on the right) is the English number word 'fifty-four.' The probability distribution corresponding to the universal approximate representation will be wide, encoding the uncertainty associated with that representation. In contrast, the probability distribution associated with the number word 'fifty-four' will be a degenerate distribution with probability mass only at the number 54, and thus with no uncertainty. These two distributions will then be combined using the same certainty-weighting principles applied to color. Because the linguistic representation for number here has perfect certainty, it will dominate absolutely, and the resulting combined representation will be identical to it and will show no trace of the universal approximate representation; this is a mirror image of the hypothetical case considered earlier in which perfect certainty of the fine-grained nonlinguistic representation yielded no effect of language. However, when linguistic numerical codes are not available, either because of the nature of the language's counting system or because of verbal interference, it is natural to represent the (nonexistent) linguistic code with a uniform distribution—in which case the combination would be identical to the approximate representation. It remains to be seen whether this account of language and number will survive more detailed examination, but some general findings do appear to be compatible with it.

We do not yet know to what extent this idea will apply to domains beyond color and number, nor how much of the evidence from such domains it may account for, and we highlight that as an important question for future research.

## CONCLUSIONS

Viewing the Sapir-Whorf hypothesis through the lens of probabilistic inference—and of the category adjustment model in particular—may be useful in several ways. This perspective echoes ideas that Whorf himself appears to have held, as interpreted by Kay and Kempton[14]:

> Whorf […] suggests that he conceives of experience as having two tiers: one, a kind of rock bottom, inescapable seeing-things-as-they-are (or at least as human beings cannot help but see them), and a second, in which [the semantic structures of a particular language] cause us to classify things in ways that could be otherwise (and are otherwise for speakers of a different language). (p. 76)

  

These two tiers map naturally onto the two halves of the category adjustment model: the universal fine-grained representation (a 'rock bottom, inescapable' representation), and the language-specific category (which may vary from language to language). Kay and Kempton argued further that these two tiers interact, such that 'there do appear to be incursions of linguistic categorization into apparently nonlinguistic processes of thinking' (p. 77). The category adjustment model suggests a precise specification of how such incursions may happen, and under what principles.

We have argued that viewing the Sapir-Whorf hypothesis in these terms has the potential to resolve some of the controversy and tension surrounding it. One source of tension is the question of universality: on at least some readings, the Sapir-Whorf hypothesis is incompatible with the important and widely-held assumption of a universal foundation for cognition. The category adjustment model—and the 'two tiers' notion more generally—helps to resolve this tension by showing how one may engage the hypothesis seriously while retaining the assumption of a universal underlying representation. The other primary source of tension is the question of replication. We have suggested that considering effects of language on cognition in terms of probabilistic inference has the potential to resolve this issue as well, by highlighting the important role of cognitive uncertainty in such effects. On this view, uncertainty can operate as a cognitive control knob, moving from situations in which there is no effect of language-specific categories (the hypothetical case in which the fine-grained universal representation is perfectly certain), to situations in which there is an effect *only* of the language-specific categories (the case of exact high numerosity), with various mixtures in between. By focusing on uncertainty, this view also captures an important level of generality: it is in principle applicable to effects of language either on cognition or on perception, because the same argument holds whether the uncertainty in question is assumed to be cognitive or perceptual in origin. While we have seen some initial evidence consistent with this view, more is needed, to test more comprehensively the various elements of this proposal, such as the modulating role of uncertainty, the universality of an underlying representation—as well as its generality across domains.

Most generally, approaching the Sapir-Whorf hypothesis in these terms has the potential to *normalize* the hypothesis, such that it need not be seen as an intellectually threatening idea with an ill-understood empirical basis, but may instead be seen as a reflection of general principles that also explain other aspects of cognition and perception.

## ACKNOWLEDGMENTS

Emily Cibelli, Joseph Austerweil, and Tom Griffiths contributed to the development of these ideas. Preparation of this article was supported by the National Science Foundation under grant SBE-1041707. We thank Paul Kay and Charles Kemp for helpful discussions, Steven Piantadosi for the connection with perceptual cue integration, and Kensy Cooperrider, Susanne Gahl, and Gary Lupyan for their comments. Any errors are our own. We dedicate this paper to the memory of Janellen Huttenlocher.

## REFERENCES

1. Sapir E. The status of linguistics as a science. *Language* 1929, 5:207–214.
2. Whorf BL. Science and linguistics. In: Carroll JB, ed. *Language, Thought, and Reality: Selected Writings of Benjamin Lee Whorf*. Cambridge, MA: MIT Press; 1956, 207–219.
3. Pinker S. *The Language Instinct: How the Mind Creates Language*. New York, NY: W. Morrow and Co.; 1994.
4. Levinson SC. Foreword. In: Carroll JB, Levinson SC, Lee P, eds. *Language, Thought, and Reality*. 2nd ed. Cambridge, MA: MIT Press; 2012, vii–xxiii.
5. Gentner D, Goldin-Meadow S. *Language in Mind: Advances in the Study of Language and Thought*. Cambridge, MA: MIT Press; 2003.
6. Boroditsky L. Linguistic relativity. In: Nadel L, ed. *Encyclopedia of Cognitive Science*. London, UK: Nature Publishing Group; 2003, 917–921.
7. Malt B, Wolff P, eds. *Words and the Mind: How Words Capture Human Experience*. New York, NY: Oxford University Press; 2010.
8. Wolff P, Holmes KJ. Linguistic relativity. *WIREs Cogn Sci* 2011, 2:253–265.
9. Gleitman L, Papafragou A. Relations between language and thought. In: Reisberg D, ed. The Oxford Handbook of Cognitive Psychology. New York, NY: Oxford University Press; 2013. doi:10.1093/oxfordhb/9780195376746.013.0032.

  

10. Brown W, Lenneberg EH. A study in language and cognition. *J Abnorm Soc Psychol* 1954, 49:454–462.
11. Berlin B, Kay P. *Basic Color Terms: Their Universality and Evolution*. Berkeley, CA: University of California Press; 1969.
12. Heider ER. Universals in color naming and memory. *J Exp Psychol* 1972, 930:10–20.
13. Heider ER, Olivier DC. The structure of the color space in naming and memory for two languages. *Cogn Psychol* 1972, 30:337–354.
14. Kay P, Kempton W. What is the Sapir-Whorf hypothesis? *Am Anthropol* 1984, 86:65–79.
15. Davidoff J, Davies I, Roberson D. Colour categories in a stone-age tribe. *Nature* 1999, 398:203–204.
16. Roberson D, Davidoff J. The categorical perception of colors and facial expressions: The effect of verbal interference. *Mem Cogn* 2000, 28:977–986.
17. Roberson D, Davies I, Davidoff J. Color categories are not universal: Replications and new evidence from a stone-age culture. *J Exp Psychol Gen* 2000, 129:369–398.
18. Pilling M, Wiggett A, Özgen E, Davies IRL. Is color "categorical perception" really perceptual? *Mem Cogn* 2003, 31:538–551.
19. Roberson D, Davidoff J, Davies IRL, Shapiro LR. Color categories: Evidence for the cultural relativity hypothesis. *Cogn Psychol* 2005, 50:378–411.
20. Gilbert A, Regier T, Kay P, Ivry R. Whorf hypothesis is supported in the right visual field but not the left. *Proc Natl Acad Sci USA* 2006, 103:489–494.
21. Winawer J, Witthoft N, Frank MC, Wu L, Wade AR, Boroditsky L. Russian blues reveal effects of language on color discrimination. *Proc Natl Acad Sci USA* 2007, 104:7780–7785.
22. Kay P, Maffi L. Color appearance and the emergence and evolution of basic color lexicons. *Am Anthropol* 1999, 101:743–760. doi:10.1525/aa.1999.101.4.743.
23. Kay P, Regier T. Resolving the question of color naming universals. *Proc Natl Acad Sci USA* 2003, 100:9085–9089.
24. Lindsey DT, Brown AM. Universality of color names. *Proc Natl Acad Sci USA* 2006, 1030:16608–16613. doi:10.1073/pnas.0607708103.
25. Kay P, Regier T. Color naming universals: The case of Berinmo. *Cognition* 2007, 102:289–298.
26. Brown AM, Lindsey DT, Guckes KM. Color names, color categories, and color-cued visual search: Sometimes, color perception is not categorical. *J Vis* 2011, 11:2.
27. Wright O, Davies IRL, Franklin A. Whorfian effects on colour memory are not reliable. *Q J Exp Psychol* 2015, 68:745–758.
28. Ernst MO, Banks MS. Humans integrate visual and haptic information in a statistically optimal fashion. *Nature* 2002, 415:429–433.
29. Yuille AL, Bülthoff HH. Bayesian decision theory and psychophysics. In: Knill DC, Richards W, eds. *Perception as Bayesian Inference*. Cambridge, UK: Cambridge University Press; 1996, 123–162.
30. Jacobs RA. What determines visual cue reliability? *Trends Cogn Sci* 2002, 6:345–350.
31. Knill DC, Pouget A. The Bayesian brain: The role of uncertainty in neural coding and computation. *Trends Neurosci* 2004, 27:712–719.
32. Berniker M, Kording K. Bayesian approaches to sensory integration for motor control. *WIREs Cogn Sci* 2011, 2:419–428.
33. Huttenlocher J, Hedges L, Prohaska V. Hierarchical organization in ordered domains: Estimating the dates of events. *Psychol Rev* 1988, 95:471–484.
34. Huttenlocher J, Hedges LV, Duncan S. Categories and particulars: Prototype effects in Estimating spatial location. *Psychol Rev* 1991, 98:352–376.
35. Huttenlocher J, Hedges LV, Vevea JL. Why do categories affect stimulus judgment? *J Exp Psychol Gen* 2000, 129:220–241.
36. Crawford LE, Huttenlocher J, Hans Engebretson P. Category effects on estimates of stimuli: Perception or reconstruction? *Psychol Sci* 2000, 11:280–284.
37. Holden MP, Newcombe NS, Shipley TF. Location memory in the real world: Category adjustment effects in 3-dimensional space. *Cognition* 2013, 128:45–55.
38. Hemmer P, Steyvers M. A Bayesian account of reconstructive memory. *Topics Cogn Sci* 2009, 1:189–202.
39. Feldman NH, Griffiths TL, Morgan JL. The influence of categories on perception: Explaining the perceptual magnet effect as optimal statistical inference. *Psychol Rev* 2009, 116:752–782.
40. Lupyan G. From chair to 'chair:' A representational shift account of object labeling effects on memory. *J Exp Psychol Gen* 2008, 137:348–369.
41. Lupyan G. Linguistically modulated perception and cognition: The label-feedback hypothesis. *Front Psychol* 2012, 3:54.
42. Hu Z, Hanley JR, Zhang R, Liu Q, Roberson D. A conflict-based model of color categorical perception: Evidence from a priming study. *Psychon Bull Rev* 2014, 21:1214–1223.
43. Goldstone RL. Effects of categorization on color perception. *Psychol Sci* 1995, 6:298–304.
44. Roberson D, Damjanovic L, Pilling M. Categorical perception of facial expressions: Evidence for a "category adjustment" model. *Mem Cogn* 1814–1829, 35:2007.
45. Hemmer P, Persaud K, Kidd C, Piantadosi S. Inferring the Tsimane's use of color categories from recognition memory. In: Noelle DC, Dale R, Warlaumont AS,


  

Yoshimi J, Matlock T, Jennings CD, Maglio PP, eds. *Proceedings of the 37th Annual Meeting of the Cognitive Science Society*. Austin, TX: Cognitive Science Society; 2015, 896–901.

46. Persaud K, Hemmer P. The influence of knowledge and expectations for color on episodic memory. In: Bello P, Guarini M, McShane M, Scassellati B, eds. *Proceedings of the 36th Annual Meeting of the Cognitive Science Society*. Austin, TX: Cognitive Science Society; 2014, 1162–1167.
47. Hemmer P, Persaud K. Interaction between categorical knowledge and episodic memory across domains. *Front Psychol* 2014, 5:584.
48. Bae G-Y, Olkkonen M, Allred SR, Flombaum JI. Why some colors appear more memorable than others: A model combining categories and particulars in color working memory. *J Exp Psychol Gen* 2015, 144:744–763.
49. Cibelli E, Xu Y, Austerweil JL, Griffiths TL, Regier T. The Sapir-Whorf hypothesis and probabilistic inference: Evidence from the domain of color. *PLoS One* 2016, 11: e0158725. doi:10.1371/journal.pone.0158725.
50. Hanley JR, Roberson D. Categorical perception effects reflect differences in typicality on within-category trials. *Psychon Bull Rev* 2011, 18:355–363.
51. Witzel C, Gegenfurtner KR. Categorical facilitation with equally discriminable colors. *J Vis* 2015, 15:22. doi:10.1167/15.8.22.
52. Witzel C, Gegenfurtner KR. Categorical perception for red and brown. *J Exp Psychol Hum Percept Perform* 2016, 42:540–570.
53. Pica P, Lemer C, Izard V, Dehaene S. Exact and approximate arithmetic in an Amazonian indigene group. *Science* 2004, 306:499–503.
54. Gordon P. Numerical cognition without words: Evidence from Amazonia. *Science* 2004, 306:496–499.
55. Frank MC, Fedorenko E, Lai P, Saxe R, Gibson E. Verbal interference suppresses exact numerical representation. *Cogn Psychol* 2012, 64:74–92.

  