PNAS
www.pnas.org

Supplementary Information for

What we talk about when we talk about colors

C.R. Twomey, G. Roberts, D.H. Brainard, and J.B. Plotkin

Colin R. Twomey
E-mail: crtwomey@sas.upenn.edu

Joshua B. Plotkin
E-mail: jplotkin@sas.upenn.edu

This PDF file includes:

Other supplementary materials for this manuscript include the following:

Supporting Information Text

1. Rate-distortion theory

Rate-distortion theory (1, 2) provides a mathematical treatment of the problem of lossy compression, based on information-theoretic quantities. In information theory (1), the entropy of a discrete random variable, X, defined

$$H(X) \triangleq \sum_{x \in \mathcal{X}} p(x) \log \frac{1}{p(x)}, \quad [1]$$

provides a measure of the average length of the shortest description ("amount of information") needed to specify the outcome of random variable X with outcomes in the set $\mathcal{X}$ occurring with probability $p(x)$. The joint entropy of X and a second random variable, Y, $H(X,Y)$, is defined similarly in terms of the joint distribution of X and Y, $p(x,y)$, and measures the average length of the shortest description needed to specify the outcomes of both random variables together. When the outcome of X is related to the outcome of Y in some (possibly nonlinear and stochastic) way, then the shortest description of both X and Y together may be smaller than the shortest descriptions of each of X and Y separately. In general, $H(X,Y) \leq H(X) + H(Y)$, with equality if and only if X and Y are statistically independent. The mutual information between X and Y, defined,

$$I(X;Y) \triangleq H(X) + H(Y) - H(X,Y), \quad [2]$$

then gives a non-negative measure of the average amount of information X and Y contain about each other, which is nonzero if and only if X and Y are not independent.

In the lossy-compression context, for a given source (random variable) X and a description of that source, $\widehat{X}$, the mutual information $I(X;\widehat{X})$ measures the amount of information the description contains about X, and it is this quantity we wish to minimize for compression, subject to a loss function, i.e. a measure of distortion. This can be formalized as

$$R(D) = \min_{p(\hat{x}|x) \ : \ \mathbb{E}d(x,\hat{x}) \leq D} I(X;\widehat{X}), \quad [3]$$

where the loss is measured in terms of an expected distortion, $\mathbb{E}d(x,\hat{x}) = \sum_{x \in \mathcal{X}} \sum_{\hat{x} \in \widehat{\mathcal{X}}} p(\hat{x}|x)p(x)d(x,\hat{x})$, with $p(x)$ a property of the source, and $p(\hat{x}|x)$ the mapping of x to $\hat{x}$ chosen to achieve on average the smallest description size possible, $R(D)$, for a given allowable average distortion, D. Intuitively, the minimum compressed description size, $R(D)$, increases as the allowable average distortion, D, decreases, dependent on the details of the source, X, and loss function, d.

1A. Bregman clustering. The classical formulation of the rate-distortion tradeoff gives an optimal mapping of X to $\widehat{X}$ for fixed $d(x,\hat{x})$. When every x and $\hat{x}$ has coordinates in a vector space, denoted $\mathbf{x}$ and $\hat{\mathbf{x}}$, respectively, then for a large family of distortion measures known as Bregman divergences, optimal coordinates for each $\hat{\mathbf{x}}$ can be found (3) in addition to the optimal mapping between X and $\widehat{X}$. For Bregman divergence $d_\phi(\mathbf{x}\|\hat{\mathbf{x}})$, defined

$$d_\phi(\mathbf{x}\|\hat{\mathbf{x}}) \triangleq \phi(\mathbf{x}) - \phi(\hat{\mathbf{x}}) - \langle \mathbf{x} - \hat{\mathbf{x}}, \nabla_\phi(\hat{\mathbf{x}}) \rangle, \quad [4]$$

with convex function ϕ, gradient $\nabla_\phi(\hat{\mathbf{x}})$ evaluated at $\hat{\mathbf{x}}$, and inner product denoted $\langle \cdot, \cdot \rangle$, the centroid of the mapping from $\widehat{X}$ to X is the minimizer of the average distortion for $\hat{x}$, i.e.

$$\mathbb{E}_{p(x|\hat{x})}\mathbf{x} = \arg\min_{\hat{\mathbf{x}}} \mathbb{E}_{p(x|\hat{x})} d_\phi(\mathbf{x}\|\hat{\mathbf{x}}). \quad [5]$$

Solutions to rate-distortion Bregman clustering (RDBC) problems have the property that each $\hat{\mathbf{x}}$ satisfies Eq. 5.

1B. Compression model of color naming. The first RDBC model of color naming appears in work by Yendrikhovskij (4). Using a perceptual measure of distortion, Yendrikhovskij (4) worked to show that efficient solutions to a tradeoff between average perceptual distortion and vocabulary size account for color categories based on natural image statistics. While the results are likely sensitive to the exact, unreported choices of natural images used to produce the image statistics (5), the conceptual link to a rate-distortion tradeoff has proved significantly productive. Using essentially the same RDBC-based compression model but disregarding scene statistics, instead using the neurophysiological constraints of perceptual discrimination and gamut alone, Regier et al. (6) showed that the compression model of color naming can qualitatively explain many of the typical vocabularies of natural languages in the WCS. Subsequent work by Zaslavsky et al. (7) investigated a "soft" partitioning variant of this same conceptual framework (although with additionally a mixture of Gaussians based measure of distortion derived from a Bayesian listener model of color naming), allowing for uncertainty in the mapping between terms and colors. In all cases, implicitly or explicitly, we can equivalently restate these compression-based accounts of color naming in terms of RDBC.

1C. Focal colors as category centroids. In the World Color Survey, participants were asked to identify among the WCS color chips the "best example" of each basic color term identified in their vocabulary. In the WCS instructions to scientists conducting the field work, this is intended to elicit a response in the participant that identifies a color chip that "... is a good, typical, or ideal..."* example of a given color term. In this work, we hypothesize that focal colors are observations of the centroids defined by Eq. 5. Two objections to this hypothesis immediately arise.

First, past work has shown that empirical measurements of category centroids differ from focal point positions (8–10), which would seem to invalidate our hypothesis. However, the discrepancy can be resolved by understanding how past work measured category centroids. Sturges and Whitfield (9), following earlier work by Boynton and Olson (8), conducted a color naming experiment similar to the WCS but in controlled laboratory conditions (and for English speakers only). Similar to the WCS, participants were asked to name, one by one in randomized sequence, a presented color chip, recording both the response as well as the timing of the response. The chips with shortest response times were considered the focal colors, and despite the difference in method these appear to be in good agreement with the "best example" focal colors recorded by Berlin & Kay for English speakers.

For each participant, the centroid of a category was computed as the average of all the color chips (in a given color space) that the participant named with that category's color term (e.g. "red," "green," etc.). To write this out mathematically, we have a sequence of participant responses, $\hat{x}^{(1)}, \hat{x}^{(2)}, \ldots, \hat{x}^{(n)}$, where each response is a color term, i.e. $\hat{x}^{(i)} \in \widehat{\mathcal{X}}$, elicited by an experimenter presented color chip, $x^{(1)}, x^{(2)}, \ldots, x^{(n)}$, where $x^{(i)} \in \mathcal{X}$. Note that each color chip in $\mathcal{X}$ was presented more than once in the sequence of n presentations. Then the centroid for category $\hat{x}$ was computed as

$$\text{centroid}(\hat{x}) = \frac{1}{\sum_{i=1}^{n} \mathbf{1}(\hat{x}^{(i)} = \hat{x})} \sum_{i=1}^{n} \mathbf{x}^{(i)} \mathbf{1}(\hat{x}^{(i)} = \hat{x}), \quad [6]$$

where $\mathbf{1}(\cdot)$ is the indicator function equal to 1 if its argument is true and 0 otherwise, and $\mathbf{x}^{(i)}$ gives the coordinates of color $x^{(i)}$ in color space. Let $n(\hat{x}|x) = \sum_{i=1}^{n} \mathbf{1}(\hat{x}^{(i)} = \hat{x})\mathbf{1}(x^{(i)} = x)$ count the occurrences of $\hat{x}$ given presentation of color chip x, then we have

$$\text{centroid}(\hat{x}) = \frac{1}{\sum_x n(\hat{x}|x)} \sum_x \mathbf{x} n(\hat{x}|x). \quad [7]$$

Let $n(x) = \sum_{i=1}^{n} \mathbf{1}(x^{(i)} = x)$ count the occurrences of x in the sequence. Then $n(\hat{x}|x) \leq n(x)$, and $p(\hat{x}|x) = n(\hat{x}|x)/n(x)$ gives the fraction of times $\hat{x}$ was used to name x, out of a total of $n(x)$ occurrences. Since each color chip was presented the same number of times, we have further that $n(x) = m$. Then we have equivalently

$$\text{centroid}(\hat{x}) = \frac{1}{m \sum_x p(\hat{x}|x)} \sum_x \mathbf{x} p(\hat{x}|x) m, \quad [8]$$

$$= \frac{1}{\sum_x p(\hat{x}|x)} \sum_x \mathbf{x} p(\hat{x}|x). \quad [9]$$

Lastly, note that $\sum_x p(\hat{x}|x) = p(\hat{x}) n_{\mathcal{X}}$, where $n_{\mathcal{X}}$ is the total number of color chips used, i.e. the cardinality of $\mathcal{X}$, and $p(\hat{x})$ is the fraction of occurrences of $\hat{x}$ in the sequence. Thus we have

$$\text{centroid}(\hat{x}) = \frac{1}{n_{\mathcal{X}}} \frac{1}{p(\hat{x})} \sum_x \mathbf{x} p(\hat{x}|x), \quad [10]$$

which by Bayes rule is equivalent to our definition of centroid with a uniform distribution of communicative need over the color chips, i.e. $p(x) = 1/n_{\mathcal{X}}$. Thus in past work centroids have been shown to differ from focal colors *when a uniform distribution of communicative need over color chips is assumed.* In this paper, by contrast, we show that by inferring and using a non-uniform distribution of communicative need we better predict both empirical color term maps and focal point positions, and that focal point positions coincide with category centroids under this non-uniform distribution of needs.

The second objection stems from work done by Abbott et al. (11) investigating a measure of the "representativeness" of focal colors based on color category extents for the WCS. Representative colors of a given category are not necessarily those with the highest likelihood, i.e. maximizing $p(x|\hat{x})$, but instead are the most likely relative to their likelihood given any other category, weighted by the prior of that category, i.e. maximizing $p(x|\hat{x}) / \sum_{\hat{x}' \neq \hat{x}} p(x|\hat{x}')p(\hat{x}')$. This appears problematic for the hypothesis that category centroids are equivalent to focal points, due to the bijection between Bregman divergences and regular exponential family distributions, and the equivalence between Bregman divergence minimization and maximum likelihood estimation (3, 12). Again, it is crucial to examine definitions to see that the discrepancy is resolved by the assumption placed on the form of $p(x|\hat{x})$. In Abbott et al. (11) $p(x|\hat{x})$ was assumed to be normally distributed. Whereas under the compression hypothesis, the maximum likelihood is taken over the mixture model as a whole, and the form of $p(x|\hat{x}) = p(\hat{x}|x)p(x)/p(\hat{x})$ is not normally distributed in general. The broader message of Abbott et al. (11) is that focal color positions reflect a balance between typicality within a category and distinction from other categories; and this interpretation agrees with our identification of focal colors as category centroids when category centroids "compete" to represent different parts of color space, as in the compression model of color naming.

* *http://www1.icsi.berkeley.edu/wcs/data.html*

2. Inverse inference of source distribution

In this section we address the general problem of inferring an unknown source distribution, $p(x)$, from knowledge of its compressed representation (i.e. a representation $\widehat{X}$ that lies on the rate-distortion curve for some unknown value of the tradeoff parameter, β). Concretely, we wish to find the $q(x)$ that best approximates the unknown distribution $p(x)$ using only what we know about $p(\hat{x})$ and $\hat{\mathbf{x}}$ from its compressed representation, with no other assumptions. For fixed marginal distribution $p(\hat{x})$ over $\widehat{\mathcal{X}}$, this can naturally be expressed as a problem of finding the conditional distributions $q(x|\hat{x})$ that together maximize the entropy of the marginal distribution $q(x) = \sum_{\hat{x}} q(x|\hat{x})p(\hat{x})$ over $\mathcal{X}$, subject to a set of constraints that enforces we recover the known compressed representation, i.e.

$$\max_{q(x|\hat{x}) \,:\, \forall \hat{x} \in \widehat{X},\ d(\tilde{\mathbf{x}}\|\hat{\mathbf{x}}) \,=\, 0} H(X), \tag{11}$$

where $H(X)$ is the Shannon entropy of X and $\tilde{\mathbf{x}} = \sum_x \mathbf{x} q(x|\hat{x})$.

We show that a numerical solution to this problem can be found via an alternating minimization strategy used by Blahut and Arimoto in their solutions to the channel maximization and rate-distortion problems (13, 14) and later generalized by Csiszár & Tusnády (15). To do so, we first note that the objective function can be rewritten as

$$\max_{q(x|\hat{x}) \in Q} I(X;\widehat{X}) + H(X|\widehat{X}), \tag{12}$$

using the fact that $I(X;\widehat{X}) = H(X) - H(X|\widehat{X})$. Here Q is the set of all conditional probability distributions such that $d(\tilde{\mathbf{x}}\|\hat{\mathbf{x}}) = 0$ for all $\hat{x} \in \widehat{\mathcal{X}}$. Since the mutual information term can be written as a maximization over $q(\hat{x}|x)$, (13, 14) i.e.

$$I(X;\widehat{X}) = \max_{q(\hat{x}|x)} \sum_x \sum_{\hat{x}} q(x|\hat{x})p(\hat{x}) \log \frac{q(\hat{x}|x)}{p(\hat{x})}, \tag{13}$$

and $H(X|\widehat{X}) = -\sum_x \sum_{\hat{x}} q(x|\hat{x})p(\hat{x}) \log q(x|\hat{x})$ is constant with respect to varying $q(\hat{x}|x)$, we can rewrite our objective function as a double maximization of the function

$$\begin{aligned} J\left[q(x|\hat{x}), q(\hat{x}|x)\right] &= I(X;\widehat{X}) + H(X|\widehat{X}) && [14]\\ &= \sum_x \sum_{\hat{x}} q(x|\hat{x})p(\hat{x}) \log \frac{q(\hat{x}|x)}{p(\hat{x})} - \sum_x \sum_{\hat{x}} q(x|\hat{x})p(\hat{x}) \log q(x|\hat{x}), && [15]\\ &= \sum_x \sum_{\hat{x}} q(x|\hat{x})p(\hat{x}) \log \frac{q(\hat{x}|x)}{q(x|\hat{x})p(\hat{x})}, && [16] \end{aligned}$$

to change the problem into one of alternating maximizations over $q(x|\hat{x})$ and $q(\hat{x}|x)$, i.e.

$$\max_{q(x|\hat{x}) \in Q} \max_{q(\hat{x}|x)} J\left[q(x|\hat{x}), q(\hat{x}|x)\right]. \tag{17}$$

The inner maximization over $q(\hat{x}|x)$ for constant $q(x|\hat{x})$ is given by $q(\hat{x}|x) = \frac{q(x|\hat{x})p(\hat{x})}{\sum_{\hat{x}} q(x|\hat{x})p(\hat{x})}$, as previously shown by Blahut and Arimoto. The outer maximization over $q(x|\hat{x})$ must respect a set of constraints that ensure we recover $\hat{\mathbf{x}}$ as a minimum distortion representation of $\mathbf{x}$ and that we have a valid probability distribution, i.e.

$$\begin{cases} d(\tilde{\mathbf{x}}\|\hat{\mathbf{x}}) = 0, & [18]\\ \sum_x q(x|\hat{x}) = 1, & [19]\\ q(x|\hat{x}) \geq 0, & [20] \end{cases}$$

where $\tilde{\mathbf{x}} = \sum_{\hat{x}} \mathbf{x} q(x|\hat{x})$. Eq. 18 enforces that there is no difference between the true compressed representation centroids $\hat{\mathbf{x}}$ and those generated by the estimated $q(x|\hat{x})$, while the remaining two constraints ensure that $q(x|\hat{x})$ is a proper probability distribution.

Temporarily setting aside the non-negativity constraint (it will be enforced by the form of the solution), the Lagrangian is then

$$\mathcal{L}\left[q(x|\hat{x})\right] = J\left[q(x|\hat{x}), q(\hat{x}|x)\right] - \sum_{\hat{x}} \lambda(\hat{x}) d(\tilde{\mathbf{x}}\|\hat{\mathbf{x}}) + \sum_{\hat{x}} \gamma(\hat{x}) \sum_x q(x|\hat{x}) \tag{21}$$

for fixed $q(\hat{x}|x)$. Taking the derivative with respect to $q(x|\hat{x})$ and setting equal to zero, we have

$$0 = p(\hat{x}) \left[\log \frac{q(\hat{x}|x)}{q(x|\hat{x})p(\hat{x})} - 1 - \lambda(\hat{x}) \frac{\partial}{\partial q(x|\hat{x})} d\Big(\sum_x \mathbf{x} q(x|\hat{x}) \| \hat{\mathbf{x}}\Big) \right] + \gamma(\hat{x}), \tag{22}$$

where we absorb a $1/p(\hat{x})$ term into each Lagrange multiplier $\lambda(\hat{x})$. If the function d is a Bregman divergence, i.e. it can be written as $d_\phi(\mathbf{u}\|\mathbf{v}) = \phi(\mathbf{u}) - \phi(\mathbf{v}) - \langle \mathbf{u}-\mathbf{v}, \nabla_\phi(\mathbf{v})\rangle$ for some convex function ϕ, then

$$\log \frac{q(x|\hat{x})}{\mu(\hat{x})} = \log \frac{q(\hat{x}|x)}{p(\hat{x})} - \lambda(\hat{x})\langle \mathbf{x}, \nabla_\phi(\tilde{\mathbf{x}}) - \nabla_\phi(\hat{\mathbf{x}})\rangle \quad [23]$$

$$q(x|\hat{x}) = \frac{1}{\mu(\hat{x})}\frac{q(\hat{x}|x)}{p(\hat{x})} e^{-\lambda(\hat{x})\left\langle \mathbf{x}, \nabla_\phi(\tilde{\mathbf{x}}) - \nabla_\phi(\hat{\mathbf{x}})\right\rangle} \quad [24]$$

Where $\log \mu(\hat{x}) = \frac{\gamma(\hat{x})}{p(\hat{x})} - 1$.

For the constraint $\sum_x q(x|\hat{x}) = 1$ to be true, the Lagrange multipliers, $\mu(\hat{x})$, must act as a normalization factor, giving us

$$q(x|\hat{x}) = \frac{q(\hat{x}|x) e^{-\lambda(\hat{x})\left\langle \mathbf{x}, \nabla_\phi(\tilde{\mathbf{x}}) - \nabla_\phi(\hat{\mathbf{x}})\right\rangle}}{\sum_{x'} q(\hat{x}|x') e^{-\lambda(\hat{x})\left\langle \mathbf{x}', \nabla_\phi(\tilde{\mathbf{x}}) - \nabla_\phi(\hat{\mathbf{x}})\right\rangle}}. \quad [25]$$

This also satisfies the non-negativity constraint for each $q(x|\hat{x})$, since $q(\hat{x}|x) \geq 0$, and $e^x \geq 0$ for any $x \in \mathbb{R}$. Finally, we can combine the unknown scalar $-\lambda(\hat{x})$ and vector $\nabla_\phi(\tilde{\mathbf{x}}) - \nabla_\phi(\hat{\mathbf{x}})$ into a single unknown vector $\nu(\hat{x})$, giving

$$q(x|\hat{x}) = \frac{q(\hat{x}|x) e^{\langle \mathbf{x}, \nu(\hat{x})\rangle}}{\sum_{x'} q(\hat{x}|x') e^{\langle \mathbf{x}', \nu(\hat{x})\rangle}}, \quad [26]$$

where $\nu(\hat{x})$ must be chosen such that Eq. 18 is true.

For any Bregman divergence, $d_\phi(\mathbf{u}\|\mathbf{v}) = 0$ iff $\mathbf{u} = \mathbf{v}$ (see Banerjee et al. (3)). Thus to enforce Eq. 18, we need to find $\nu(\hat{x})$ s.t. $\tilde{\mathbf{x}} = \sum_x \mathbf{x} q(x|\hat{x}) = \hat{\mathbf{x}}$. Let $G_{\hat{x}}(\nu) = \log \sum_x q(\hat{x}|x) \exp\langle \mathbf{x}, \nu\rangle$. Then the vector of partial derivatives of $G_{\hat{x}}(\nu)$ with respect to ν are given by

$$\nabla G_{\hat{x}}(\nu) = \sum_x \mathbf{x} \frac{q(\hat{x}|x) e^{\langle \mathbf{x}, \nu\rangle}}{\sum_{x'} q(\hat{x}|x') e^{\langle \mathbf{x}', \nu\rangle}} = \tilde{\mathbf{x}}(\nu). \quad [27]$$

Since $G_{\hat{x}}(\nu)$ is strictly convex, we have by Legendre transform its convex conjugate dual,

$$G^*_{\hat{x}}(\tilde{\mathbf{x}}) = \sup_\nu \langle \tilde{\mathbf{x}}, \nu\rangle - G_{\hat{x}}(\nu). \quad [28]$$

and vector of partial derivatives

$$\nabla G^*_{\hat{x}}(\tilde{\mathbf{x}}) = \arg\sup_\nu \langle \tilde{\mathbf{x}}, \nu\rangle - G_{\hat{x}}(\nu). \quad [29]$$

By the strict convexity of $G_{\hat{x}}$ and the definition of the Legendre transform we have that $\nabla G^*_{\hat{x}}(\tilde{\mathbf{x}}) = (\nabla G_{\hat{x}}(\tilde{\mathbf{x}}))^{-1} = \nu(\tilde{\mathbf{x}})$, i.e. the unique choice of ν for a given value of $\tilde{\mathbf{x}}$. The unique choice of ν to guarantee $\tilde{\mathbf{x}} = \hat{\mathbf{x}}$ is then simply $\nu(\hat{\mathbf{x}}) = \nabla G^*_{\hat{x}}(\hat{\mathbf{x}})$, which can be computed numerically via e.g. BFGS.

The alternating maximization algorithm is then to iterate

$$\begin{cases} q_t(\hat{x}|x) = \dfrac{q_t(x|\hat{x})p(\hat{x})}{\sum_{\hat{x}} q_t(x|\hat{x})p(\hat{x})} & [30] \\ q_{t+1}(x|\hat{x}) = \dfrac{q_t(\hat{x}|x) e^{\langle \mathbf{x}, \nu_t(\hat{x})\rangle}}{\sum_{x'} q_t(\hat{x}|x') e^{\langle \mathbf{x}', \nu_t(\hat{x})\rangle}}, & [31] \end{cases}$$

with $\nu_t(\hat{x}) = \nabla G^*_{\hat{x},t}(\hat{\mathbf{x}})$, and starting from any initial $q_0(x|\hat{x})$. By construction, the choice of $q_t(\hat{x}|x)$ maximizes J for fixed $q_t(x|\hat{x})$, and $q_{t+1}(x|\hat{x})$ maximizes J for fixed $q_t(\hat{x}|x)$, subject to their respective constraints. We thus have a sequence indexed by t of non-decreasing values for J, which converges whenever the maximum entropy is finite. The solution for the marginal distribution of X is then given by $q(x) = \sum_{\hat{x}} q_\infty(x|\hat{x})p(\hat{x})$.

2A. Convergence to the global optimum. In this section we will show that the alternating minimization algorithm defined by Eq. 30 and Eq. 31 converges to the global maximum of $J\left[q(x|\hat{x}), q(\hat{x}|x)\right]$ for any initial choice of $q_0(\hat{x}|x)$. We will do this using a geometric approach developed by Csiszár & Tusnády (15),[†] which for example can be used to prove convergence to the global optimum for the alternating minimization algorithm proposed by Blahut (14) to find numerical solutions to the rate-distortion problem. First, note that maximizing $J\left[q(x|\hat{x}), q(\hat{x}|x)\right]$ is equivalent to minimizing $D\left[q(x|\hat{x}), q(\hat{x}|x)\right] = -J\left[q(x|\hat{x}), q(\hat{x}|x)\right]$. Then by Theorems 1 and 2 of Csiszár & Tusnády (15), to show convergence to the global minimum via alternating minimizations of D it is sufficient to show that the "three points property" and "four points property" both hold for D and a choice of functional, δ.

[†] See also Byrne (16) as a helpful reference.

Definition 1. (From Csiszár & Tusnády (15)) Let $\delta[p, p']$ be a non-negative valued function on $P \times P$ such that $\delta[p, p] = 0$ for each $p \in P$. Given D and δ, for a $p \in P$ the three points property holds if

$$\delta[p, p_{t+1}] + D[p_{t+1}, q_t] \leq D[p, q_t], \tag{32}$$

whenever $p_{t+1} = \arg\min_p D[p, q_t]$. The four points property holds for a $p \in P$ if for every $q \in Q$

$$D[p, q_t] \leq \delta[p, p_t] + D[p, q], \tag{33}$$

whenever $q_t = \arg\min_q D[p_t, q]$.

We will show that the three and four point properties hold for D and the following choice of δ,

$$\delta\left[q(x|\hat{x}), q'(x|\hat{x})\right] = \sum_{\hat{x}} p(\hat{x}) \sum_x q(x|\hat{x}) \log \frac{q(x|\hat{x})}{q'(x|\hat{x})}. \tag{34}$$

Non-negativity of Eq. 34 follows directly from the non-negativity of the KL-divergence and $p(\hat{x})$, as does equality holding iff $q(x|\hat{x}) = q'(x|\hat{x})$.

We will also make use of the fact that we can rewrite both the definition of δ given by Eq. 34 and D in terms of the following Bregman divergence,

$$d_\psi(\mathbf{x}\|\mathbf{y}) = \sum_i w_i \sum_j x_{ij} \log \frac{x_{ij}}{y_{ij}} - \sum_i w_i \sum_j (x_{ij} - y_{ij}), \tag{35}$$

where w_i are constant non-negative weights that sum to one, and $x_{ij}, y_{ij} \geq 0$, not necessarily summing to one. In this case ψ is the strictly convex function $\psi(\mathbf{x}) = \sum_i w_i \sum_j x_{ij} \log x_{ij}$. Then with $w_i = p(\hat{x})$, i indexing elements of $\widehat{X}$, and j indexing elements of X, we have that

$$\delta\left[q(x|\hat{x}), q'(x|\hat{x})\right] = d_\psi(q(x|\hat{x})\|q'(x|\hat{x})), \tag{36}$$

and

$$\begin{aligned}
D[q(x|\hat{x}), q(\hat{x}|x)] &= -J[q(x|\hat{x}), q(\hat{x}|x)], && [37]\\
&= \sum_x \sum_{\hat{x}} q(x|\hat{x}) p(\hat{x}) \log \frac{q(x|\hat{x}) p(\hat{x})}{q(\hat{x}|x)}, && [38]\\
&= \sum_{\hat{x}} p(\hat{x}) \sum_x q(x|\hat{x}) \log \frac{q(x|\hat{x})}{q(\hat{x}|x)} + \sum_{\hat{x}} p(\hat{x}) \log p(\hat{x}), && [39]\\
&= \sum_{\hat{x}} p(\hat{x}) \sum_x q(x|\hat{x}) \log \frac{q(x|\hat{x})}{q(\hat{x}|x)} - H(\widehat{X}) \\
&\quad - \left[\sum_{\hat{x}} p(\hat{x}) \sum_x (q(x|\hat{x}) - q(\hat{x}|x))\right] + \left[\sum_{\hat{x}} p(\hat{x}) \sum_x (q(x|\hat{x}) - q(\hat{x}|x))\right], && [40]\\
&= d_\psi(q(x|\hat{x})\|q(\hat{x}|x)) + 1 - \sum_{\hat{x}} p(\hat{x}) \sum_x q(\hat{x}|x) - H(\widehat{X}). && [41]
\end{aligned}$$

Lemma 1. The three points property, $\delta[q(x|\hat{x}), q_{t+1}(x|\hat{x})] + D[q_{t+1}(x|\hat{x}), q_t(\hat{x}|x)] \leq D[q(x|\hat{x}), q_t(\hat{x}|x)]$, where $q_{t+1}(x|\hat{x}) = \arg\min_{q(x|\hat{x})} D[q(x|\hat{x}), q_t(\hat{x}|x)]$, holds.

Proof. Rewriting using Eq. 36 and Eq. 41, we must show that

$$\begin{aligned}
&d_\psi(q(x|\hat{x})\|q_{t+1}(x|\hat{x})) + d_\psi(q_{t+1}(x|\hat{x})\|q_t(\hat{x}|x)) + 1 - \sum_{\hat{x}} p(\hat{x}) \sum_x q_t(\hat{x}|x) - H(\widehat{X}) && [42]\\
&\quad \leq d_\psi(q(x|\hat{x})\|q_t(\hat{x}|x)) + 1 - \sum_{\hat{x}} p(\hat{x}) \sum_x q_t(\hat{x}|x) - H(\widehat{X}). && [43]
\end{aligned}$$

Cancelling, we need to show

$$d_\psi(q(x|\hat{x})\|q_{t+1}(x|\hat{x})) + d_\psi(q_{t+1}(x|\hat{x})\|q_t(\hat{x}|x)) \leq d_\psi(q(x|\hat{x})\|q_t(\hat{x}|x)), \tag{44}$$

which follows immediately from the Generalized Pythagoras Theorem (3) and the fact that by construction solutions of Eq. 31 maximize J for fixed $q_t(\hat{x}|x)$, so that

$$q_{t+1}(x|\hat{x}) = \arg\min_{q(x|\hat{x})} D\left[q(x|\hat{x}), q_t(\hat{x}|x)\right], \tag{45}$$

$$= \arg\min_{q(x|\hat{x})} d_\psi(q(x|\hat{x})\|q_t(\hat{x}|x)) + \underbrace{1 - \sum_{\hat{x}} p(\hat{x}) \sum_{x} q_t(\hat{x}|x) - H(\widehat{X})}_{\text{constant}}, \tag{46}$$

$$= \arg\min_{q(x|\hat{x})} d_\psi(q(x|\hat{x})\|q_t(\hat{x}|x)). \tag{47}$$

■

Lemma 2. The four points property, $D\left[q(x|\hat{x}), q_t(\hat{x}|x)\right] \leq \delta\left[q(x|\hat{x}), q_t(x|\hat{x})\right] + D\left[q(x|\hat{x}), q(\hat{x}|x)\right]$, where

$$q_t(\hat{x}|x) = \arg\min_{q(\hat{x}|x)} D\left[q_t(x|\hat{x}), q(\hat{x}|x)\right],$$

holds.

Proof. From the definitions of D and δ, we must show that

$$\sum_{\hat{x}} p(\hat{x}) \sum_{x} q(x|\hat{x}) \log \frac{q(x|\hat{x})p(\hat{x})}{q_t(\hat{x}|x)} \leq \sum_{\hat{x}} p(\hat{x}) q(x|\hat{x}) \log \frac{q(x|\hat{x})}{q_t(x|\hat{x})} + \sum_{\hat{x}} p(\hat{x}) \sum_{x} q(x|\hat{x}) \log \frac{q(x|\hat{x})p(\hat{x})}{q(\hat{x}|x)}. \tag{48}$$

By subtraction, equivalently we must show that

$$0 \leq \sum_{\hat{x}} p(\hat{x}) q(x|\hat{x}) \log \frac{q(x|\hat{x})}{q_t(x|\hat{x})} + \sum_{\hat{x}} p(\hat{x}) \sum_{x} q(x|\hat{x}) \log \frac{q_t(\hat{x}|x)}{q(\hat{x}|x)}. \tag{49}$$

Denoting $q_t(x) = \sum_{\hat{x}} q_t(x|\hat{x})p(\hat{x})$, from Eq. 30 we have that $q_t(\hat{x}|x) = q_t(x|\hat{x})p(\hat{x})/q_t(x)$. Then by substitution we have

$$0 \leq \sum_{\hat{x}} p(\hat{x}) q(x|\hat{x}) \log \frac{q(x|\hat{x})}{q_t(x|\hat{x})} + \sum_{\hat{x}} p(\hat{x}) \sum_{x} q(x|\hat{x}) \log \frac{q_t(x|\hat{x})p(\hat{x})}{q(\hat{x}|x)q_t(x)}, \tag{50}$$

$$= \sum_{\hat{x}} p(\hat{x}) q(x|\hat{x}) \log \frac{q(x|\hat{x})p(\hat{x})}{q(\hat{x}|x)q_t(x)}, \tag{51}$$

$$= \sum_{\hat{x}} p(\hat{x}) q(x|\hat{x}) \log \frac{q(x)}{q_t(x)}, \tag{52}$$

$$= \sum_{x} q(x) \log \frac{q(x)}{q_t(x)}, \tag{53}$$

where Eq. 52 follows from the fact that $q(x) = q(x|\hat{x})p(\hat{x})/q(\hat{x}|x)$, and Eq. 53 from the fact that $q(x) = \sum_{\hat{x}} q(x|\hat{x})p(\hat{x})$. Then this is equivalent to the statement that $0 \leq D_{\mathrm{KL}}\left[q(x)\|q_t(x)\right]$, which is true by non-negativity of the KL-divergence. ■

Theorem 1. The sequence of alternating maximizations defined by Eq. 30 and Eq. 31 converges to the global maximum of $J\left[q(x|\hat{x}), q(\hat{x}|x)\right]$ for any initial choice of $q_0(\hat{x}|x)$.

Proof. Proof of Theorem 2A follows from satisfying the five point property of Csiszár & Tusnády (15), which is implied by satisfying the three and four points properties from Lemma 1 and Lemma 2, respectively. ■

2B. Uniqueness. In the previous section we showed that the solution found by the alternating maximization algorithm is globally optimal. Here we show that the optimal $q(x)$ distribution is also unique.

Theorem 2. The distribution $q^*(x) = \sum_{\hat{x}} q^*(x|\hat{x})p(\hat{x})$ for the $q^*(x|\hat{x})$ achieving the maximum of $J\left[q(x|\hat{x}), q(\hat{x}|x)\right]$ is unique.

Proof. Assume $q^*(x)$ is not unique, and there exists a distinct solution $q'(x)$ that also achieves the maximum of $J\left[q(x|\hat{x}), q(\hat{x}|x)\right]$ with $q'(x|\hat{x})$. Then two things must be true.

First, since $q^*(x)$ and $q'(x)$ are distinct, then $0 < D_{\mathrm{KL}}\left[q^*(x)\|q'(x)\right]$. From the definition of the KL-divergence and using the fact that $q(x) = \sum_{\hat{x}} q(x|\hat{x})p(\hat{x})$, we have that

$$0 < \sum_{x} \sum_{\hat{x}} q^*(x|\hat{x})p(\hat{x}) \log \frac{q^*(x)}{q'(x)}. \tag{54}$$

Since $q(x) = q(x|\hat{x})p(\hat{x})/q(\hat{x}|x)$ (for any choice of $\hat{x}$), the definition $q(x|\hat{x})$ from Eq. 31, and the equivalence of $\nu^*(\hat{x}) = \nu'(\hat{x}) = \nu(\hat{x})$, we have

$$0 < \sum_x \sum_{\hat{x}} q^*(x|\hat{x})p(\hat{x}) \log \frac{\frac{e^{\langle \mathbf{x}, \nu(\hat{x})\rangle}}{\sum_{x'} q^*(\hat{x}|x)e^{\langle \mathbf{x}', \nu(\hat{x})\rangle}}}{\frac{e^{\langle \mathbf{x}, \nu(\hat{x})\rangle}}{\sum_{x'} q'(\hat{x}|x)e^{\langle \mathbf{x}', \nu(\hat{x})\rangle}}}, \quad [55]$$

$$= \sum_{\hat{x}} p(\hat{x}) \log \frac{\sum_{x'} q'(\hat{x}|x)e^{\langle \mathbf{x}', \nu(\hat{x})\rangle}}{\sum_{x'} q^*(\hat{x}|x)e^{\langle \mathbf{x}', \nu(\hat{x})\rangle}}, \quad [56]$$

since after cancellation none of the terms depend on x except $q^*(x|\hat{x})$, and $\sum_x q^*(x|\hat{x}) = 1$.

Second, since both $q^*(x)$ and $q'(x)$ achieve the global optimum, we must have that $J\left[q^*(x|\hat{x}), q^*(\hat{x}|x)\right] = J\left[q'(x|\hat{x}), q'(\hat{x}|x)\right]$. Then after cancelling we have

$$\sum_{\hat{x}} p(\hat{x}) \sum_x q^*(x|\hat{x}) \log \frac{q^*(\hat{x}|x)}{q^*(x|\hat{x})} = \sum_{\hat{x}} p(\hat{x}) \sum_x q'(x|\hat{x}) \log \frac{q'(\hat{x}|x)}{q'(x|\hat{x})}. \quad [57]$$

From the definition of $q(x|\hat{x})$ in Eq. 31 and the equivalence of $\nu^*(\hat{x}) = \nu'(\hat{x}) = \nu(\hat{x})$,

$$\sum_{\hat{x}} p(\hat{x}) \sum_x q^*(x|\hat{x}) \log \frac{e^{\langle \mathbf{x}, \nu(\hat{x})\rangle}}{\sum_{x'} q^*(\hat{x}|x)e^{\langle \mathbf{x}', \nu(\hat{x})\rangle}} = \sum_{\hat{x}} p(\hat{x}) \sum_x q'(x|\hat{x}) \log \frac{e^{\langle \mathbf{x}, \nu(\hat{x})\rangle}}{\sum_{x'} q'(\hat{x}|x)e^{\langle \mathbf{x}', \nu(\hat{x})\rangle}}. \quad [58]$$

Then, since $\sum_x \mathbf{x} q^*(x|\hat{x}) = \sum_x \mathbf{x} q'(x|\hat{x}) = \hat{\mathbf{x}}$, we can cancel the $\sum_{\hat{x}} p(\hat{x})\langle \hat{\mathbf{x}}, \nu(\hat{x})\rangle$ term from both sides, and using the fact that $\sum_x q^*(x|\hat{x}) = \sum_x q'(x|\hat{x}) = 1$, we have

$$0 = \sum_{\hat{x}} p(\hat{x}) \log \frac{\sum_{x'} q'(\hat{x}|x)e^{\langle \mathbf{x}', \nu(\hat{x})\rangle}}{\sum_{x'} q^*(\hat{x}|x)e^{\langle \mathbf{x}', \nu(\hat{x})\rangle}}. \quad [59]$$

But this contradicts the inequality established by Eq. 56. Thus $q^*(x)$ must be unique. ■

2C. Example inference and comparison to prior work. As an illustrative example, we present the results of the inverse inference method above for a known distribution of needs, $p(x)$. This toy example allows us to study the properties of the inverse inference when the ground truth, $p(x)$, is known. We also use this example to illustrate the difference between our inference method and inferences based on two prior methods in the literature. Rather than solving for the maximum entropy distribution consistent with a rate-distortion optimal vocabulary, the "capacity achieving prior" (CAP) method (7) assumes instead that the true $p(x)$ will be one such that, given a vocabulary of term mappings $p(\hat{x}|x)$, we only ever need to communicate the x's that are maximally unambiguous to specify with that vocabulary. The CAP distribution is the one that achieves the maximum channel capacity for the given term map $p(\hat{x}|x)$ (the specification of the channel from X to $\widehat{X}$), i.e. satisfying

$$p_{\mathrm{CAP}}(x) = \arg\max_{q(x)} \sum_{x,\hat{x}} p(\hat{x}|x)q(x) \log \frac{p(\hat{x}|x)}{q(\hat{x})}, \quad [60]$$

where $q(\hat{x}) = \sum_{x'} p(\hat{x}|x')q(x')$. This is a strong assumption in general, and when it is violated, as we will see in this section, the CAP provides a poor approximation of the true distribution $p(x)$.

We also compare our approach to the word-frequency method (here abbreviated WF) proposed in (17), which asks for the maximum entropy distribution $p_{\mathrm{WF}}(x)$ that satisfies the linear constraints $p(\hat{x}) = \sum_x p(\hat{x}|x)p_{\mathrm{WF}}(x)$. In other words, the WF method solves for the distribution of communicative needs consistent with a given term mapping, $p(\hat{x}|x)$, and known word frequencies, $p(\hat{x})$. To understand what this does in practice, it is instructive to consider the case of "hard" clusters where $p(\hat{x}|x)$ equals either 1 or 0. In this case, we derive an analytical solution for the WF inference (see SI Sec. 4A), which is given by $p_{\mathrm{WF}}(x) = p(\hat{x}(x))/\sum_{x'} p(\hat{x}(x)|x')$, where $\hat{x}(x)$ is the unique nonzero $\hat{x}$ in $p(\hat{x}|x)$; i.e., the $\hat{x}$ chosen for a given x. In effect, then, the WF method approximates communicative need by dividing the frequency of a given word, $p(\hat{x})$, uniformly across its mapped domain. In the "soft" case, solutions behave similarly but with an additional factor accounting for "fuzzy" boundaries between $\hat{x}$ domains. While this gives considerably more reasonable estimates of needs than the CAP approach, it depends on the availability of word frequencies, $p(\hat{x})$, and it again requires knowledge of the term maps $p(\hat{x}|x)$; the former is unknown for almost all WCS languages, and the latter introduces circularity when aiming to predict term maps based on language-specific communicative needs.

In our toy example $x \in X$ covers the unit grid ($n = 100 \times 100$) with an arbitrary but specified distribution $p(x)$, as shown in Fig. S1A (ground truth). The figure also shows the RDBC solutions for either 4 or 8 terms (Fig. S1A and S1B, respectively; this example uses squared Euclidean distance as the distortion measure). The ground truth distribution $p(x)$ was chosen to be nonuniform, with a broad probability gradient from $(0, 0)$ to $(1, 1)$, and a smaller-scale low to high to low to high oscillation in probability along the x-axis. The RDBC centroids and Voronoi (nearest-centroid) regions show a non-uniform division of X into clusters (or "terms," to link this to the terminology of the color naming problem), as a result of using a non-uniform $p(x)$.

Based on only the positions of the focal terms $\hat{\mathbf{x}}$ and the term frequencies $p(\hat{x})$, our inverse method produces an estimate of $p(x)$ that recapitulates the broad-scale features of the ground truth. The inverse inference performs well even with as few as 4 terms (Fig. S1A), with some additional, fine-scale details captured when inferring from 8 terms (Fig. S1B). By contrast, the distributions inferred by the CAP method, which are not based on $\hat{\mathbf{x}}$ and $p(\hat{x})$ but instead require knowing the full term map $p(\hat{x}|x)$, deviate significantly from the ground truth (Fig. S1A and S1B; note different scale).

In Fig. S1C, the entropy of inferred and CAP solutions are shown for a broader range of vocabulary sizes (from 2 to 10). The figure also quantifies the dissimilarity between the ground truth and the estimated distributions, based on their KL-divergence. Successive iterations of the inverse inference algorithm show monotonic convergence to a maximal entropy value that lies between the ground-truth entropy and the unconstrained maximum entropy distribution (uniform over X). Note there are only small differences between the maximum entropy values achieved when varying the vocabulary size used (the equivalent of the number of color terms). While not directly constrained by the inverse inference method, since the ground truth distribution is assumed unknown, the inverse method converges to distributions that are very close to the true distribution. Solutions become closer to ground truth as the vocabulary size increases, but even small vocabularies provide inferences that closely approximate the ground truth. By comparison, CAP solutions have entropies that are substantially lower than the maximum or even the ground truth entropy, and they are sensitive to vocabulary size. CAP solutions are orders of magnitude more divergent from ground truth, compared to the results of the inverse inference method we have developed.

3. Application to color categories

We use the inverse inference method of SI Sec. 2 to find the distributions of communicative need for empirical color vocabularies via the following correspondence (outlined in Fig. 1C). In this application, the source, X, denotes the visible colors that need to be communicated, which are the WCS stimuli set. Each WCS stimulus color, x, in the set of WCS stimuli, $\mathcal{X}$, has a position $\mathbf{x}$ in CIE Lab , a perceptually uniform color space. The unknown distribution of communicative need we wish to infer is $p(x)$. Our estimate of $p(x)$ will be the one that best matches the known position, $\hat{\mathbf{x}}$, of each "best-example," or focal, color for each term, $\hat{x}$, in the language's color vocabulary, $\widehat{\mathcal{X}}$, and otherwise maximizes the entropy of the inferred distribution.

Intuitively, in the inverse inference procedure (SI Sec. 2), the vectors $\nu(\hat{x})$ can be thought of as "pulling" on the inferred distribution such that the inferred centroids match the position of the true centroids. In the example shown in SI Sec. 2C, the positions of the true centroids lie in the interior of the boundary of all the x positions. To match prior work and the WCS itself, we use the WCS color chips (Fig. 1A) as the support set for the inverse inference. Since WCS participants selected focal colors from this same set, the average focal color position across participants could lie on or near the boundary of the support set if there is high agreement among participants. To match these positions with the given support set, the inverse method would be forced to "pull" with overly large magnitudes towards these remote points, when this is just an artifact of the constraints on participants and the choice of support.

To check if this was the case, and to mitigate any impact it may have, we constrained the maximum magnitude that any $\nu(\hat{x})$ could have, and varied this value as a parameter, λ. At $\lambda = 0$ the inverse method makes no attempt to match the language centroids, and we recover only the uniform distribution over the WCS color chips. At $\lambda = \infty$, we recover the unconstrained inverse inference method. At intermediate values of λ, pathologically large magnitudes have limited impact on the inference. If indeed there are pathologically large magnitudes at play, then there should be a large difference between the entropy at $\lambda = \infty$, where the inferred distribution becomes overly concentrated at the problematic focal point, and at intermediate values of λ for which the nearly the same RMSE between inferred and true focal points is achieved. Fig. S2A shows that this is exactly the case, and suggests that $\lambda \leq 0.25$ is sufficient to achieve RMSE's close to the unconstrained solutions, while maintaining substantially higher entropies.

Note that RMSE is measured using the empirical focal points and the position of the focal points for the optimal rate-distortion fit using the inferred distribution at a given value of λ. Rate-distortion solutions were found using the standard alternating minimization algorithm (see Banerjee (3)),

$$\begin{cases} \hat{\mathbf{x}}_t = \sum_x \mathbf{x} q_t(x|\hat{x}), & [61] \\ q_t(\hat{x}) = \sum_x q_t(\hat{x}|x) p(x), & [62] \\ q_{t+1}(\hat{x}|x) = \frac{q_t(\hat{x}) e^{-\beta d(\mathbf{x}\|\hat{\mathbf{x}}_t)}}{\sum_{\hat{x}'} q_t(\hat{x}') e^{-\beta d(\mathbf{x}\|\hat{\mathbf{x}}'_t)}}, & [63] \end{cases}$$

where $q_t(x|\hat{x}) = q_t(\hat{x}|x)p(x)/\sum_{x'} q_t(\hat{x}|x')p(x')$, and β is a parameter that acts as an "inverse temperature," controlling the "softness" (low values of β) or "hardness" (high values) of the boundaries between terms given by $p(\hat{x}|x)$. Since RDBC solutions are not unique, we run the algorithm starting from many different initial conditions (initial $\hat{\mathbf{x}}$ positions drawn uniformly at random from the set of WCS color chips) until convergence (change in $\hat{\mathbf{x}}$ positions between iterations is $< 1 \times 10^{-5}$ or the maximum number of iterations is reached; max iterations 1×10^4 used in searches for the optimal value of β; max iterations 5×10^4 for calculation of RDBC solution using the optimal value β), and keep the solution with lowest mean squared error. We used a standard derivative-free nonlinear optimization method (bound optimization by quadratic approximation (18), via the `nloptr` (v1.2.1) package for R v3.6.3) to search for lowest mean squared error values of β.

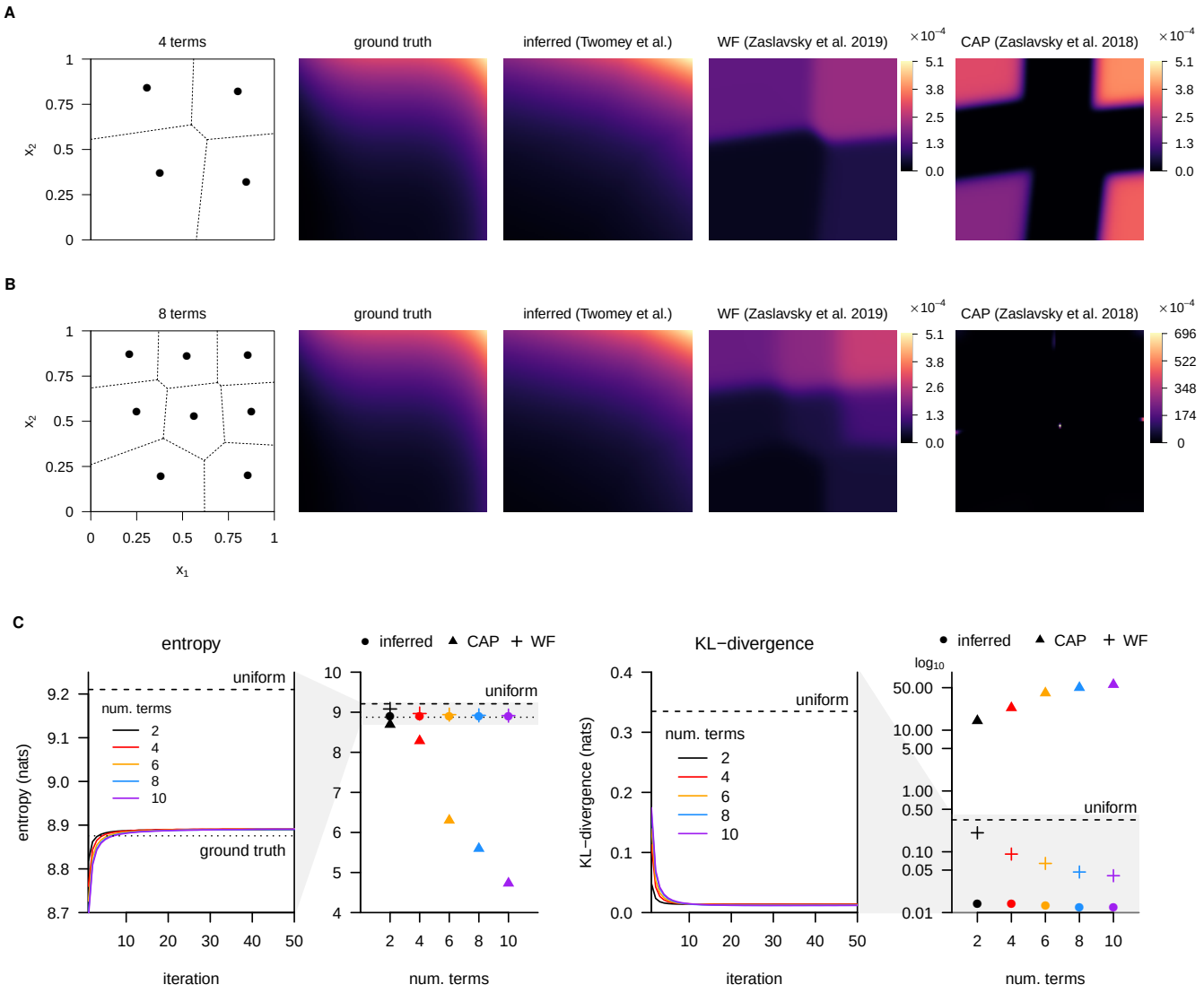

Fig. S1. Example inference of the distribution, $p(x)$, underlying a given rate-distortion solution. **(A)** An example rate-distortion vocabulary (*left*) with 4 terms (black points show position of term centroids, $\hat{\mathbf{x}}$) giving a compressed representation of a dense square grid (100×100) of elements X with distribution $p(x)$ (*middle-left* "ground truth"). Using the positions, $\hat{\mathbf{x}}$, and probabilities, $p(\hat{x})$, of the vocabulary terms alone, the inverse algorithm infers distribution $p_{\mathrm{inf}}(x)$ (*middle-center*), which approximates the main features of the true distribution $p(x)$ (same color scale). For comparison, we also show the distributions inferred under the word-frequency (WF) method (*middle-right*) and the capacity achieving distribution (CAP, *right*). **(B)** The same example as in (A) but for an 8-term rate-distortion vocabulary (grid and ground truth distributions are identical). **(C)** Evolution of the key quantities in the inference algorithm's iterative solutions to the example used in (A) & (B) , for vocabulary sizes 2, 4, 8, and 16, with comparisons to WF and CAP used in prior work. (*Left*) The entropy of the inferred distribution monotonically increases with each iteration of the inverse inference algorithm, and converges to values between the true entropy (dotted line labeled "ground truth") and the maximum entropy (dashed line labeled "uniform") for this example as expected. The entropy is reduced and approaches the true entropy as the number of terms is increased, but only to a small degree compared to CAP. The adjacent figure with expanded y-axis shows converged values for inferred, WF, and CAP distributions (circles, crosses, and triangle plotting symbols, respectively). CAP solutions have lower entropy than the true distribution, are more sensitive to the number of vocabulary terms, and become increasingly different as the number of terms increases. (*Right*) The KL-divergence between the inferred distribution and true distribution tends to decrease and converges to small values. This is a consequence of matching the term centroids since the true distribution is not known to the inverse inference algorithm. The adjacent figure compares the inferred, WF, and CAP distributions KL-divergence to the true distribution at convergence (log scale). The distributions inferred by our method are close to the true distribution, and become even closer with increasing vocabulary size; while the CAP distributions are far from the ground truth and become increasingly farther, even more so than uniform. The WF solutions are sensitive to the number of terms available, and at 10 terms give solutions that are nearly a factor 3 further from ground truth than our inferred solutions for 2 terms (0.04 vs. 0.014).

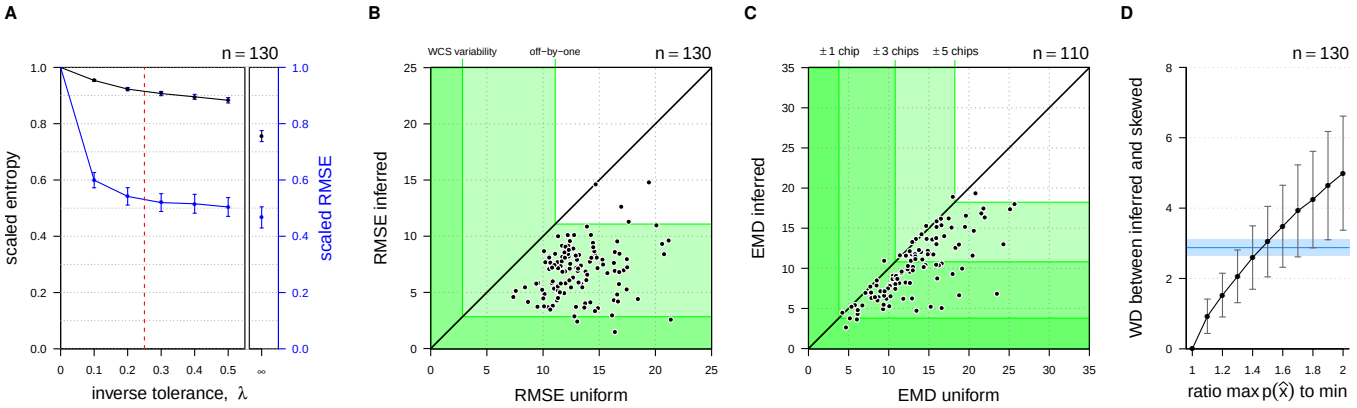

Fig. S2. Application of inverse inference algorithm to WCS. **(A)** In the constrained-maximum version of the inverse inference method, small values of the inverse tolerance parameter, λ, (x-axis) can achieve values of RMSE (mean and 95% confidence intervals shown scaled relative to uniform) comparable to the non-relaxed inverse inference ($\lambda = \infty$), while maintaining a much higher entropy (shown scaled relative to uniform). **(B)** RMSE between rate-distortion optimal vocabularies under inferred distribution, $p_{\text{inf}}(x)$, (y-axis) and empirical ground truth, compared to RMSE under uniform distirbution, $p_{\text{unif}}(x)$, (x-axis). All points lie below 1–1 line (black), showing that inferred strictly improves over uniform for matching focal point posisitions. Regions bounded by reference lines for median RMSE from within-language variability in focal point position ("WCS variability") and median all focal point positions off-by-one WCS chip ("off-by-one") are shown overlapping in green. **(C)** Average Earth mover's distance (EMD) between rate-distortion predicted and empirically observed term maps for each WCS language vocabulary under a uniform distribution of communicative need (x-axis) or the language inferred distribution (y-axis). Reference lines at ± 1, ± 3, and ± 5 chips show the median EMD across languages comparing empirically observed languages to themselves $\pm$ a rotation in hue (rotation of WCS columns 1:40). **(D)** Sensitivity of inferred language-specific communicative needs to the assumption of uniform term frequencies, $p(\hat{x})$. Mean and standard deviation Wasserstein distance is shown between inferred distributions under a uniform $p(\hat{x})$ and an asymmetric ("skewed") distribution constructed with varying ratios of max $p(\hat{x})$ to min $p(\hat{x})$ (x-axis). Reference line (blue) shows median Wasserstein distance and 95% CI between inferred distributions derived from language mean focal color position and focal colors resampled from language speaker responses (based on WCS languages).

For each B&K+WCS language, the minimal RMSE for inverse inference with $\lambda \leq 0.25$ is shown in Fig. S2B (y-axis), and compared with the minimal RMSE (same optimization procedure for non-unique RDBC solutions and choice of β) for uniform (x-axis). In all cases, use of the inferred distribution reduces RMSE compared to uniform (all points below 1–1 line). As useful references, we quantified the RMSE for within-language variability in focal point positions among participants (via bootstrap resampling of participant responses and measuring their RMSE with respect to the mean focal point positions for that language), as well as the RMSE when all terms are off by one WCS color chip. Most inferred distribution RMSE's are between the median values of these two reference quantities, which is not the case for uniform.

Similarly, in Fig. S2C we show the absolute improvement in term map predictions for the WCS languages shown in Fig. 3B, comparing the Earth mover's distance (EMD) between predicted and empirical term maps based on inferred (y-axis) and uniform (x-axis) distributions. WCS languages were used for term map comparisons both for the ability to resample from among speaker responses (the B&K data surveyed only one speaker per language) to assess confidence intervals on improvement in Fig. 3B, and because the B&K study design substantially differed methodologically from the WCS in the color naming task.[‡] In the WCS color naming was assayed for each color chip, whereas in B&K participants selected chips out of the full set of stimuli (19). While the B&K term maps are related to $p(\hat{x}|x)$, they are not straightforward estimates of $p(\hat{x}|x)$ as in the WCS, and behave qualitatively differently. As useful reference points, we computed the EMD between empirical vocabularies and rotations thereof, approximated by cycling WCS columns 2:41. This transform preserves the structure of each vocabulary while increasing the displacement (in hue) between the true and rotated terms, and has been used in prior work on color naming (6). Here it provides a more meaningful distance scale for the EMD measurements than e.g. chip-wise randomization.

3A. RMSE reference points. We provide three points of comparison for the RMSE distributions shown in Fig. 2B. First, the "WCS variability" reference line was computed by resampling participant focal point choices by language, recomputing the mean focal point across resampled participants, and measuring the RMSE between the recomputed focal points and the actual language focal points. We used the median computed RMSE as a useful reference point approximating a lower bound on how well predicted focal points might be expected to perform. Second, the "off-by-one" reference line was computed by repeatedly offsetting each focal point by one WCS chip sampled uniformly at random from the neighborhood of WCS color chips in Fig. 1A and measuring the RMSE between the set of perturbed focal points for a language and the actual focal points. The median computed RMSE in this case gives an intermediate point of comparison for predicted focal point RMSE distributions. Third, the "random" reference line was computed by resampling each language's focal points from the WCS color chips uniformly at random without replacement, then assigning each resampled focal point to the nearest true focal point, and measuring the RMSE of the two sets of focal points under this assignment. This gives an approximate upper bound on how poorly a predicted set of focal points might perform, using the same procedure for assigning predicted focal points under the rate-distortion model to actual language focal points.

[‡] The focal color assays of B&K and the WCS were essentially the same, however. Hence the inclusion of both data sets in other analyses where only focal color estimates are necessary.

3B. Sensitivity of inferred distributions to term frequencies, $p(\hat{x})$. The inverse inference algorithm uses the frequency of vocabulary terms, $p(\hat{x})$, as part of the inference process for determining $p(x)$. This information is not available for color vocabularies in the B&K and WCS datasets, and further field work would be required to estimate these quantities directly (with the additional caveat that vocabularies may have changed since having been originally surveyed). And so the present work uses the simplifying assumption of a uniform $p(\hat{x})$. This is a reasonable approximation, given the WCS selection criteria for basic color terms and evidence in English that basic color terms are elicited with approximately equal frequency under a free naming task (10). Nevertheless, to investigate the sensitivity of inferred distributions to this choice, we compared inferred distributions under increasingly asymmetric ("skewed") distributions for $p(\hat{x})$, sampled from a linearly increasing set of probabilities between the minimum and maximum $p(\hat{x})$. Fig. S2D shows the Wasserstein distance between the inferred distributions under uniform $p(\hat{x})$ and the skewed distribution as a function of the ratio between the maximum and minimum $p(\hat{x})$. As a useful reference point, we computed the median Wasserstein distance between inferred distributions under the uniform $p(\hat{x})$ assumption re-sampled from the WCS language speaker populations. Ratios in usage greater than approximately 1.5 would be needed before non-uniformity would begin to have a more significant impact on inferred distributions than the among-speaker variability inherent in the data. While this suggests the choice of a uniform $p(\hat{x})$ is reasonable to a first approximation, the extent to which this assumption of uniformity may be violated in some languages remains an open, but potentially tractable, question for future field work.

3C. Sensitivity of inferred distributions to vocabulary size, $|\widehat{\mathcal{X}}|$. Under the rate distortion hypothesis, color vocabularies optimize the information a listener can infer about the color being referenced, based on the color term chosen by a speaker. Because there are far fewer terms than perceivable colors, there is by necessity some loss of information caused by the compression of colors into terms. As a result, the size of a vocabulary (number of terms) should have some impact on our ability to infer the underlying distribution of communicative needs, $p(x)$: larger vocabularies should provide more resolution and more detail in the inferred distribution. For the B&K+WCS languages, we can expect that fewer terms will result in the recovery of only broad-scale features of a language's communicative needs, while more terms allow for additional detail.

This effect is demonstrated in Figure S3A. Here we generated rate-distortion efficient vocabularies for simulated languages, each generated from the same underlying distribution of underlying communicative needs and differing only in the number of color terms. We then inferred the distribution of needs from the simulated focal color positions, using our inverse inference method. As expected, we find that having more color terms allows for more detail to be recovered in the inferred distribution of needs, although the results are qualitatively similar across a range of vocabulary sizes.

We also investigated the relationship between vocabulary size and inferred needs in more systematic detail. To do so, we again generated rate-distortion efficient vocabularies for pairs of languages sharing the same underlying communicative needs and differing only in vocabulary size. We used the B&K+WCS average inferred distribution as a "ground truth," and the number of terms in each simulated vocabulary was restricted to the range of terms found in the B&K+WCS data. We then inferred the communicative needs for each simulated vocabulary in the pair, and we measured their Wasserstein distance. Figure S3B shows a small but statistically significant impact of differences in vocabulary size on the measured distance between inferred distributions of need – which arises because vocabulary size has an impact on the resolution of the inference. For comparison to these simulations, in which the underlying needs are kept constant, we also plotted the distances between inferred needs measured in the empirical data, for all pairs of B&K+WCS languages. The empirical distances between inferred needs are much larger than can be explained by the simulated data. These results imply that differences in vocabulary size alone cannot explain the large differences observed among B&K+WCS inferred communicative needs. Moreover, the relationship between differences in vocabulary size and differences in inferred needs has substantially greater magnitude in the empirical data than in the simulated languages. This suggests that there may be typical ways in which communicative needs evolve as the vocabularies of languages change in size – which is an interesting hypothesis for future study.

3D. Capacity achieving distributions for individual WCS languages. The capacity achieving distributions, which are referred to as priors (CAP) in the literature, should not in general be expected to approximate the true distribution of communicative need, as shown in SI Sec. 2C. Here we reproduce the average CAP across the WCS languages reported in Zaslavsky et al. (7, 17). The average CAP differs by several orders of magnitude from the average distribution $p(x)$ inferred in this paper (Fig. S4A). The language-specific CAP's for Waorani and Martu-Wangka are shown in Fig. S4B: they each differ radically from the communicative needs we estimate by our inference method. The CAP distributions feature implausible variation in communicative need across nearby colors.

3E. Field work variability. Variability in how the field work for the WCS was conducted for different languages does not appear to explain the instances of non-improvement in Fig. 3B term map predictions. For the WCS, native speakers were asked to use only the basic color terms of their language, as previously identified according to a set of specific linguistic criteria. However in some cases it seems that native speakers apparently were not so constrained, either by experimenter or participant choice. Based on the identification of these two modes in the WCS by Gibson et al. (20) in their supplementary materials, there was no apparent relationship between the choice of methodology and a language showing improvement or no improvement under the inferred distribution vs. uniform.

3F. Potential correlates of communicative needs. Here we consider possible correlates of the communicative needs we infer, based on proxy measures and potential confounds suggested by prior work.

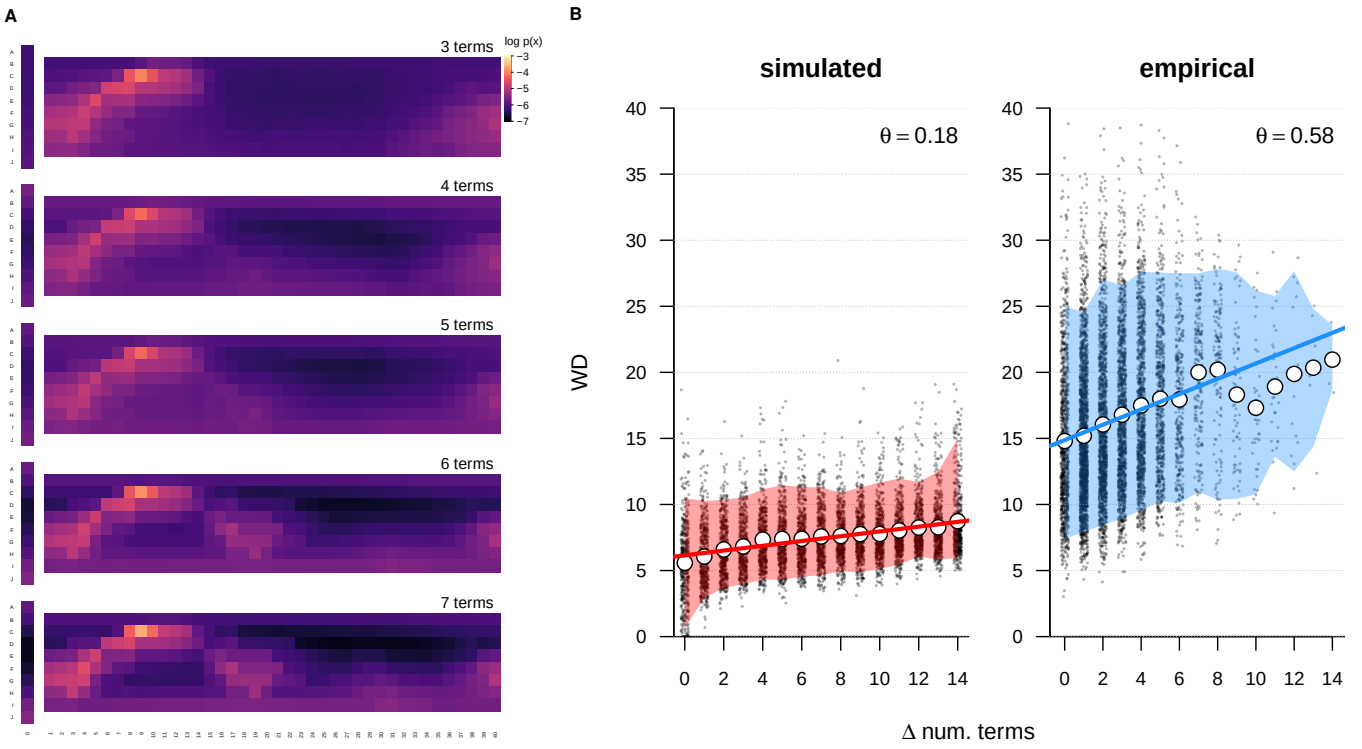

Fig. S3. Sensitivity to number of color terms. **(A)** Average inferred distributions of communicative need for rate-distortion optimal vocabularies simulated with different numbers of terms but the same underlying communicative need (the B&K+WCS average inferred distribution). A larger number of terms provides more resolution and detail in the inferred distribution of needs, but the inferred distributions are nonetheless qualitatively the same. **(B)** Simulated (*left*) and empirical (*right*) Wasserstein distances (WD) between inferred distributions of need for pairs of languages, shown as a function of the difference in the number of their terms (Δ num. terms, white points show mean WD for each Δ num. terms). Differences in inferred communicative needs (WD) are substantially smaller in the simulations, which isolate the effects attributable to vocabulary size alone, compared to the differences observed among empirical languages (red and blue bands show 90% extent of simulated and empirical distances, respectively). Also, the relationship between vocabulary size and differences in inferred needs (WD) is substantially smaller (slope of linear regression $\theta = 0.18$, red line) in the simulated data with a single shared distribution of needs, compared to the relationship observed in the empirical data (slope $\theta = 0.58$, blue line).

3F.1. Communicative efficiency (surprisal). Gibson et al. (20) sought to understand communicative needs by estimating communicative efficiency, or surprisal, defined as

$$S(x) = -\sum_{\hat{x}} p(\hat{x}|x) \log p(x|\hat{x}), \qquad [64]$$

(Eq. 1 of Gibson et al. 2017), where x are color stimuli (i.e. the WCS color chips), $\hat{x}$ are color terms, and $p(x|\hat{x}) \propto p(\hat{x}|x)$ by Bayes rule under the assumption of a uniform $p(x)$. This is distinct from our work, where by "communicative needs" we in fact mean the quantity $p(x)$ – namely, the chance that a speaker needs to reference color x – which we infer directly for individual languages by maximum entropy (SI Sec. 2). Nevertheless, we can ask whether surprisal $S(x)$ is predictive of the language-specific communicative needs we infer. While we do find a moderate positive correlation (Pearson's $\rho = 0.41$; $n = 36{,}300$; see SI Fig. S8A), the vast majority of variance in needs $p(x)$ remains unexplained ($1 - R^2 = 0.83$). The substantial differences between these two quantities explains why the needs that we infer have a comparatively weak correlation with the warm-cool trend in colors of salient objects, whereas the warm-cool trend is stronger for the surprisal measure.

3F.2. Color saturation (chroma). The WCS color chip stimuli were chosen to cover a range of Munsell lightness values and hues at approximately maximal "saturation" (Munsell chroma). Maximal color saturation, or chroma, thus varies as a function of Munsell lightness value and hue for these stimuli. Our inference of communicative needs depends on the positions of color chips used in the WCS. One concern, then, might be that the variation in color saturation, or chroma, in some way directly determines our inferred communicative needs. If this were the case, then we would expect to find a systematic relationship between inferred language-specific communicative needs and Munsell chroma. But we do not find any systematic relationship (SI Fig. S8B), even though the highest values of chroma tend to correspond to high values of $p(x)$. This relationship for high chroma values may reflect the observation that participants choice of focal colors can be biased towards highly saturated colors (26). Or it could alternatively, or additionally, reflect a common cause in the determinants of communicative needs and perceptual discrimination.

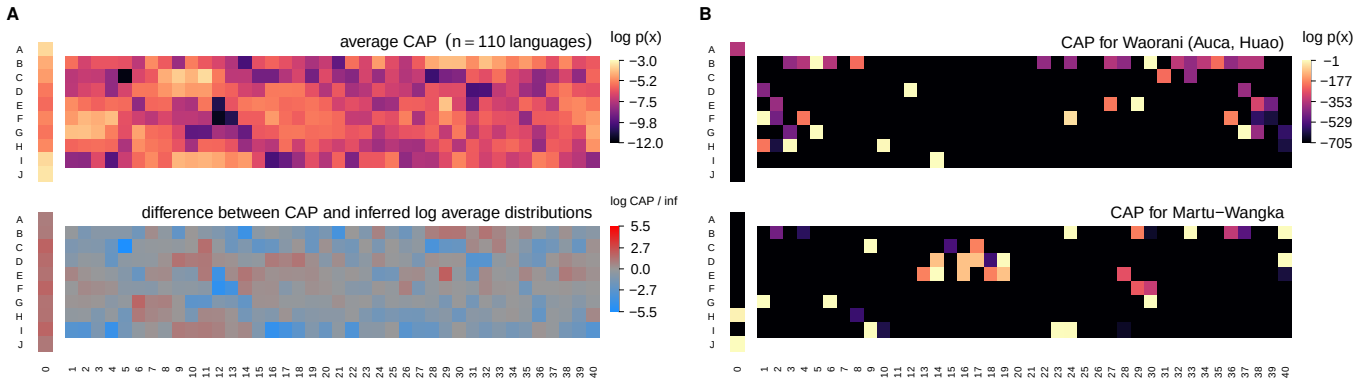

Fig. S4. Comparison to the capacity maximizing distributions ("capacity achieving priors") for WCS languages. **(A)** (Top) Approximate replication of Fig. 3a from Zaslavsky et al. (17) showing capacity achieving prior (CAP) averaged across WCS languages (here we include all WCS languages; some were excluded in Zaslavsky et al. (17)). (Bottom) The difference between the average CAP and average prior we infer (see Fig. 2A) ranges over several orders of magnitude (log scale). **(B)** CAP distributions for two languages used as examples in Fig. 5A (note different scale). Under the CAP inference, two neighboring Munsell color chips may exhibit a 10^{300}-fold difference in communicative need.

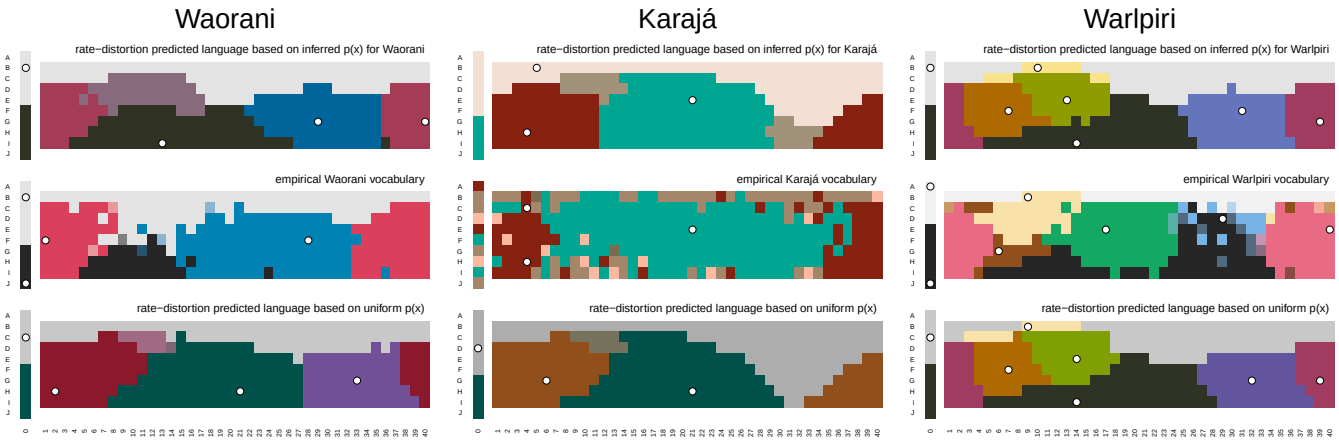

Fig. S5. Results for WCS languages previously identified in the literature as possible outliers. RDBC results shown for the languages labeled in Fig. 3B (Pirahá shown in Fig. 2C). Prior work has hypothesized that Pirahá (21), Warlpiri (22), Waorani (23), and Karajá (23), may be exceptions in some way to the broad trends identified in the WCS. All but Warlpiri appear to be substantially improved when we account for language specific communicative needs.

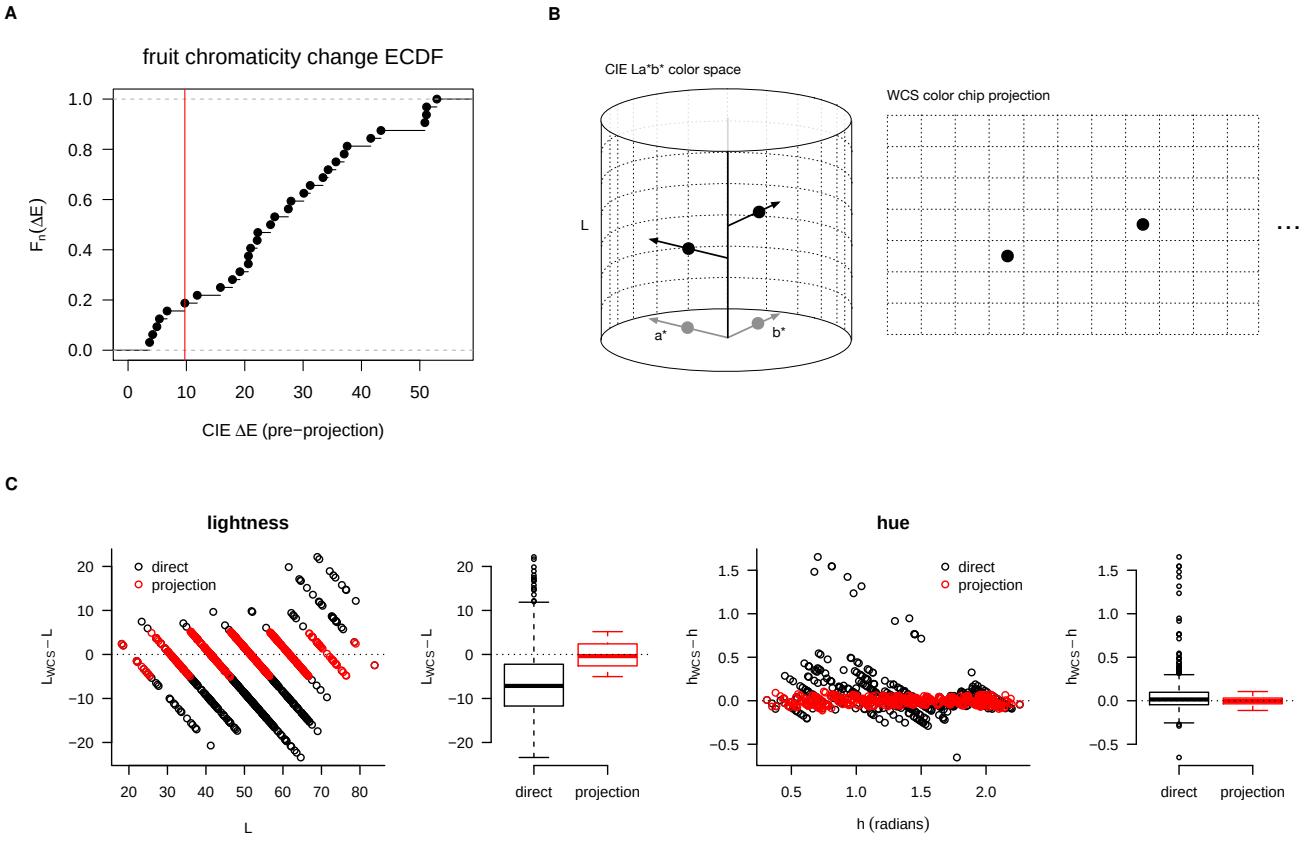

Fig. S6. Treatment of Sumner & Mollon fruit chromaticity measurements. **(A)** Empirical cumulative distribution function (ECDF) of the change in fruit chromaticity between unripe and ripe states. Not all fruits signal ripeness by a change in chromaticity (24, 25); other indicators may include size or smell. For each species collected by Sumner & Mollon having at least one measurement in each of the 'unripe' and 'ripe' classifications, the species' chromaticity measurements is included in our analysis (Fig. 4C & D) if the CIE Lab difference (ΔE^*) between the mean unripe and ripe measurements is greater than a threshold (red vertical line). This threshold is determined by the minimum ΔE of a subset of the species measurements for which we could establish a significant change in mean CIE Lab coordinates at the $p < 0.01$ level based on a Hotelling T^2 test. **(B)** After conversion from spectral measurements to CIE Lab coordinates, the final step is to find the nearest WCS color chip in CIE Lab space. The WCS color chips form a high-saturation outer shell of the Munsell color array, privileging lightness (L) and hue angle over saturation. We adopt this same choice by selecting nearest neighbors based on L and hue angle (i.e. normalizing the (a^*, b^*) position sub-vector), ignoring saturation. **(C)** The choice of matching by projection rather than directly by ΔE^* better constrains the difference in lightness (L) and hue (h) between the matched WCS color chips and the true CIE Lab coordinates, with the tradeoff of a small increase in the overal mean ΔE^* (35.5 vs. 44.7). However this tradeoff appears to be necessary to make meaningful comparisons between fruit ripeness categories; without projection there is substantial variation in the residuals as a function of L and h. Box-plots show median and first and third quartiles; whiskers extend to the minimum (maximum) up to 1.5 times the interquartile range, with outliers shown as individual points.

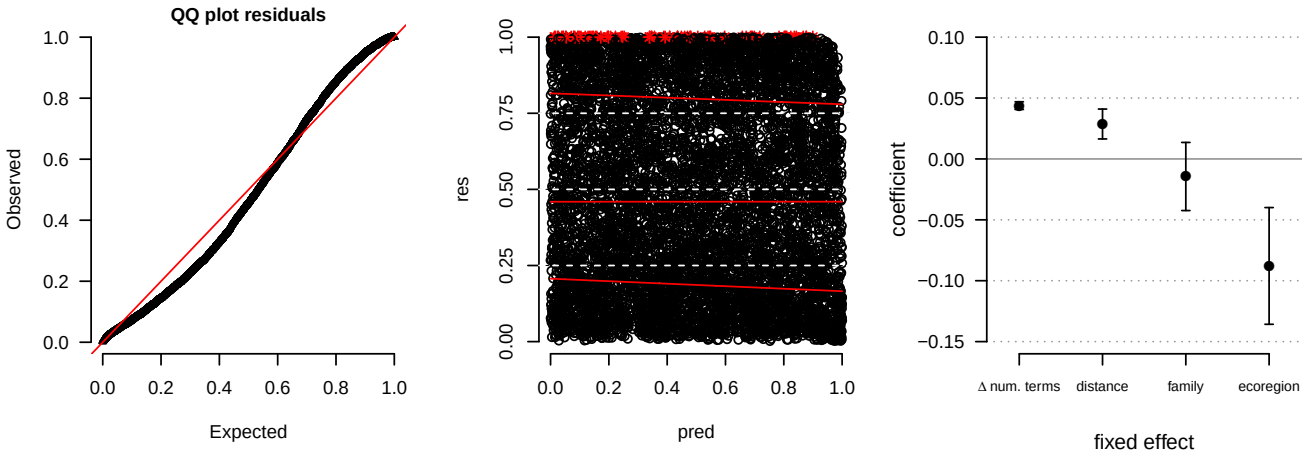

Fig. S7. Diagnostics for GLMM of differences in communicative need between languages. (*Left*) Uniform quantile–quantile (QQ) plot of expected vs. observed GLMM model residuals. (*Middle*) Rank transformed model predicted values (pred) vs. residuals (res), with quantile regressions (red lines) compared to theoretical quantiles (dashed white lines at 0.25, 0.50, and 0.75); simulation outliers shown as red stars. (*Right*) Fixed effect coefficients and 95% confidence intervals for geodesic distance (Haversine method; standardized units), shared linguistic family (TRUE=1, FALSE=0), and shared ecoregion (TRUE=1, FALSE=0). Positive coefficients indicate an increase in dissimilarity (increase in Wasserstein distance), while negative coefficients indicate a decrease. Out of $n = 125^2$ language pairs, 73% and 60% shared the same linguistic family or ecoregion, respectively. Variance inflation factors (VIFs) for distance, family, and ecoregion were 1.291, 1.219, 1.149, respectively. All VIFs are less than 5, showing low multicollinearity.

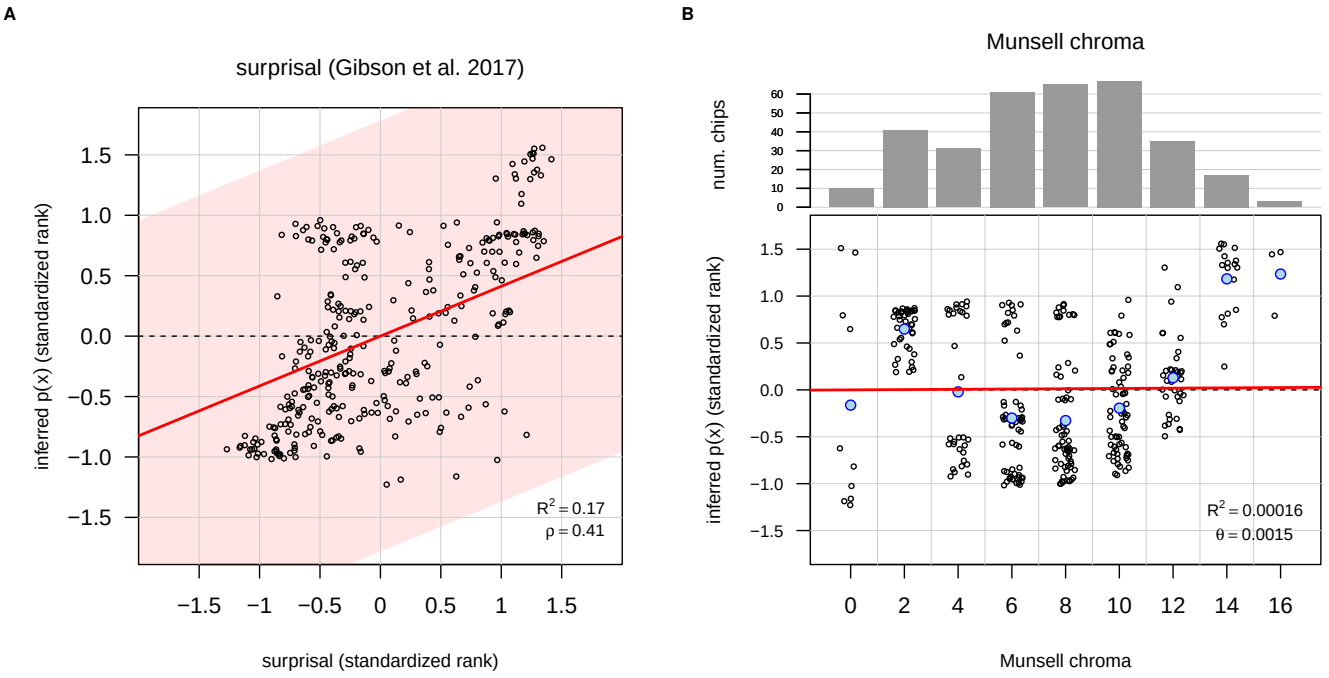

Fig. S8. Relationship of inferred language-specific communicative needs to surprisal (Gibson et al. 2017) and color saturation of WCS stimuli (Munsell chroma). **(A)** Surprisal is moderately correlated, but not strongly predictive, of inferred communicative needs. Red line and legend show linear relationship between standardized rank values; light-red area indicates 95% prediction interval. Points show average (over languages) standardized rank inferred communicative needs (y-axis) compared with average (over languages) standardized rank surprisal (x-axis) for each WCS color chip. **(B)** Munsell chroma (color saturation) of WCS stimuli are not predictive of language-specific inferred communicative needs. Red line and legend describe linear relationship based on 110 languages and 330 color chips per language; $n = 36{,}300$ total comparisons. Points show the Munsell chroma ($+$ jitter; x-axis) and average (over languages) standardized rank inferred communicative needs for each WCS color chip (y-axis). Standardized ranks are computed per language. The total number of WCS color chips of each Munsell chroma used in the WCS is shown at top. Blue points show expected values of inferred communicative needs by Munsell chroma. Expected values are poor summaries for intermediate chroma values due to multi-modality.

4. Comparison to alternative approaches

In this section we provide additional discussion and comparison of our approach to inferring communicative needs with prior approaches. While past work proposed methods in the context of approximating a single, global distribution of need (7, 17, 27), it is reasonable to consider whether or not those approaches could also be used to estimate language-specific needs. First, for intuition, we provide a brief derivation of an analytical solution to Zaslavsky et al.'s word-frequency (WF) approach (17) for the special case of "hard" category boundaries, i.e. $p(\hat{x}|x)$ equal to either 1 or 0 only. Second, we give illustrative examples comparing our approach with that of WF (17) and CAP (7) for inference of communicative needs and prediction of color naming maps. Finally, we provide a systematic comparison of estimation methods for communicative needs when word frequencies are unknown (which is the case for almost all languages in the WCS).

4A. Analytical solution for a special case of the WF method. The WF method (17) approximates communicative needs by finding the maximum entropy distribution over colors, x, such that the marginalization of the joint distribution formed by measured term maps, $p(\hat{x}|x)$, and estimated $p_{\text{WF}}(x)$, matches the measured word frequencies, $p(\hat{x})$; i.e., such that $p(\hat{x}) = \sum_x p(\hat{x}|x)p_{\text{WF}}(x)$. The method assumes that word frequencies $p(\hat{x})$ are known. By Lagrange multipliers, one can show that solutions (for "hard" or "soft" conditions), must have the form $p_{\text{WF}}(x) \propto \exp\left[\sum_{\hat{x}} \nu(\hat{x})p(\hat{x}|x)\right]$, where the constant of proportionality normalizes $p_{\text{WF}}(x)$, and $\nu(\hat{x})$ are Lagrange multipliers that need to be chosen to enforce the constraint that $p(\hat{x}) = \sum_x p(\hat{x}|x)p_{\text{WF}}(x)$.

Let μ be the constant of proportionality; in the case of hard clustering, let $\hat{x}(x)$ denote the one nonzero $\hat{x}$ for a given x's $p(\hat{x}|x)$ distribution; and with a slight abuse of notation, let $x \in \hat{x}$ indicate the set of all x s.t. $p(\hat{x}|x) = 1$. Then for hard clustering we can decompose the constraint into two parts,

$$p(\hat{x}) = \sum_{x\in\hat{x}} p(\hat{x}|x)\mu^{-1}e^{\nu(\hat{x})p(\hat{x}|x)} + \sum_{x\notin\hat{x}} p(\hat{x}|x)\mu^{-1}e^{\nu(\hat{x}(x))p(\hat{x}(x)|x)}, \quad [65]$$

$$= \mu^{-1}e^{\nu(\hat{x})} \sum_{x\in\hat{x}} p(\hat{x}|x), \quad [66]$$

$$\nu(\hat{x}) = \log \frac{p(\hat{x})\mu}{\sum_{x\in\hat{x}} p(\hat{x}|x)}, \quad [67]$$

where the second step follows because (for hard clustering) $p(\hat{x}|x) = 1$ for any $x \in \hat{x}$ and $= 0$ for any $x \notin \hat{x}$. Then after plugging this into the form of the solution and cancelling out μ, we have

$$p_{\text{WF}}(x) = \frac{p(\hat{x}(x))}{\sum_{x'\in\hat{x}(x)} p(\hat{x}(x)|x')}, \quad [68]$$

where the denominator simplifies to $\sum'_x p(\hat{x}(x)|x') = |\{x \in \hat{x}\}|$. In words, the maximum entropy solution for hard clustering under the WF constraints simply apportions the frequency of a word, $p(\hat{x})$, evenly across its domain as given by $p(\hat{x}|x)$. This analytical solution for the hard clustering case gives a helpful intuition for understanding the performance of WF more generally, as we describe in the subsequent two sections.

4B. Illustrative comparison of approaches for language-specific needs. Here, we consider two illustrative examples of our inference method and model predictions in comparison to past work, where the ground truth is known. For each example, we generated a simple, arbitrary ground truth distribution of communicative needs over a unit square domain. (We consider a comparison across a more diverse set of systematically generated examples in the subsequent section, SI Sec. 4C). In the first example, shown in SI Fig. S9A, the true distribution of communicative needs has a maximum near the top right, and a minimum near the bottom left. The ground truth RDBC for this example provides just three categories (centroids and largest-likelihood $p(\hat{x}|x)$ shown with unique colors). Without using knowledge of the term map, $p(\hat{x}|x)$, our inference method is able to recover the coarse features of the true distribution of needs. The WF method approximately evenly divides the frequency of each term, $p(\hat{x})$, across its mapped domain, $p(\hat{x}|x)$, similar to the "hard" partitioning case solved analytically in SI Sec. 4A. The CAP method exponentially concentrates probability mass away from the boundaries between terms (all distributions in SI Fig. S9 are shown on a log scale).

In these examples, the ground truth rate-distortion Bregman clustering (RDBC) is based on a squared-error measure of distortion. It then seems surprising that predictions of term maps using the CAP distribution, which does not well approximate the ground truth communicative need distribution, nonetheless closely resemble the true term mapping, $p(\hat{x}|x)$, when using with the information bottleneck (IB) model proposed in Zaslavsky et al. (7). Under the IB model, categories are not characterized by a centroid at a single point in space, but by a distribution over all points in space, which gives a large degree of additional flexibility. When coupled with language-specific inferences based on e.g. CAP, this evidently allows for recovery of the ground truth term maps despite the difference between the true generative process (based on centroids at single points in space with distortion measured by squared-error between points) in this example and the process specified by the IB model (based on mixtures of Gaussians over all points in space, with distortion measured by the KL-divergence between Gaussian mixtures). More critically, the requirement of empirical term maps, $p(\hat{x}|x)$, as inputs for both the CAP and WF approaches necessarily leads to circularity if these methods are then used for predicting language term maps based on language-specific inferences of communicative need.

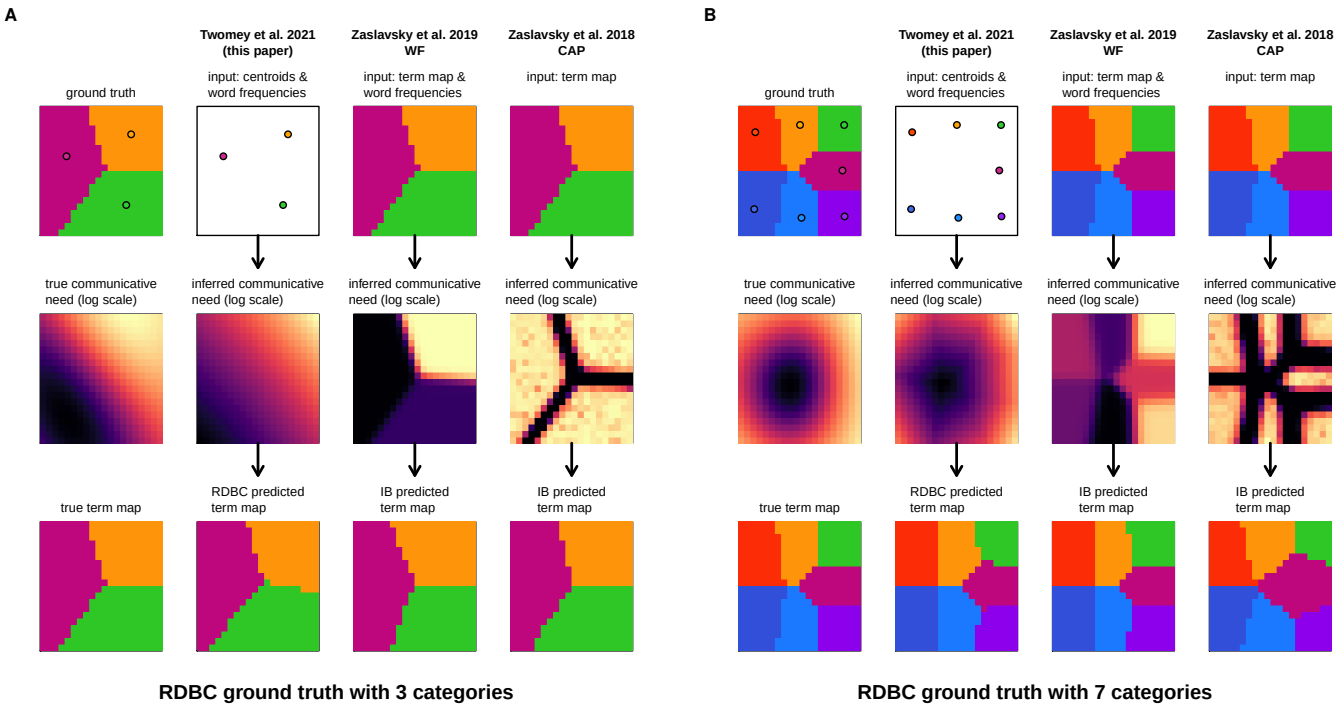

Fig. S9. Comparison of methods for inference of communicative needs and prediction of term maps, when the ground truth is known. **(A)** A rate-distortion optimal "vocabulary" with three terms for the unit square is shown at top-left, with category centroids (points) and best-choice term maps (colored regions). The true distribution of communicative needs, $p(x)$, is shown in the middle row (ground truth). Our inference method (second column) takes as input the category centroids and, if available, their frequencies, $p(\hat{x})$, and produces an estimate of communicative need (middle row, same column). Our model of color naming then predicts term maps (bottom row, same column) using the inferred communicative need. Inferences based on the word-frequency based approach of Zaslavsky et al. (17) (third column), and the CAP approach of Zaslavsky et al. (7) (fourth column) do not accurately reconstruct the true distribution of needs. These two methods each use the IB model for prediction of term maps. Note that for language-specific communicative needs, term maps would necessarily be inputs for both the WF and CAP methods, leading to circularity in prediction of term maps. **(B)** A seven-term vocabulary example.

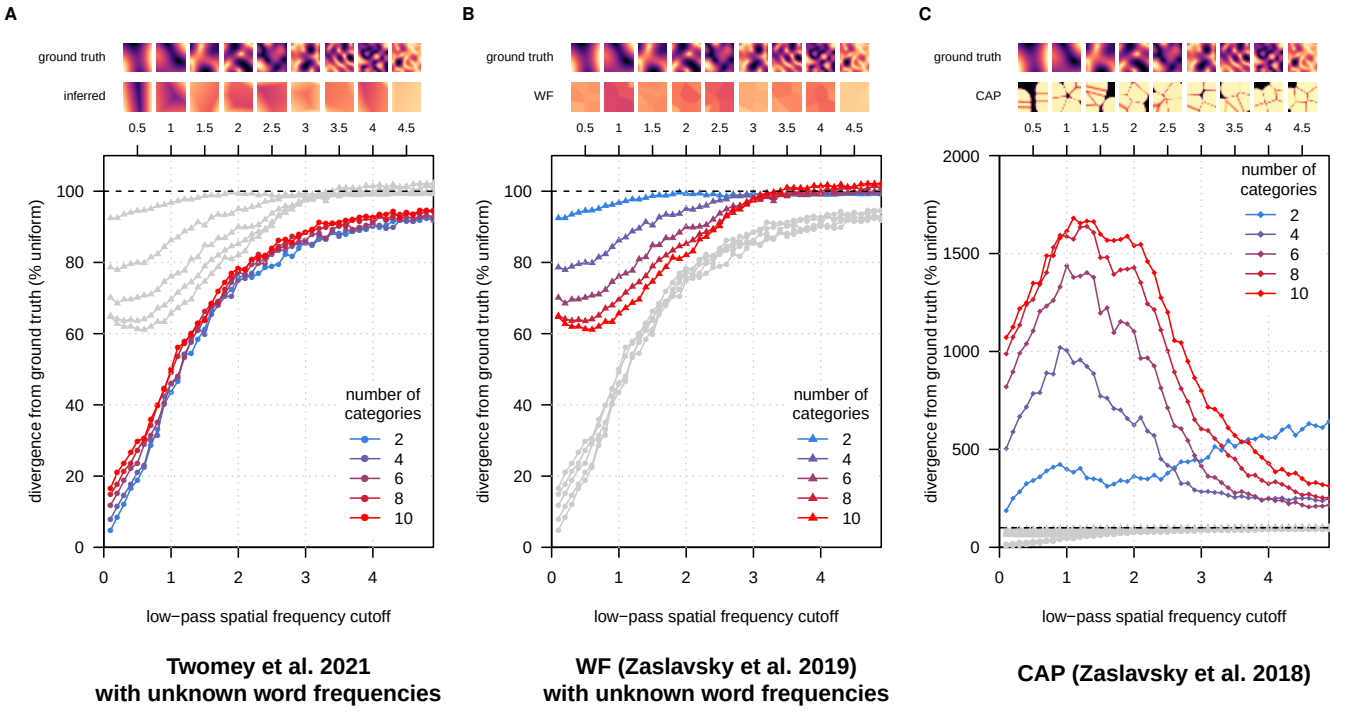

Fig. S10. Systematic comparison of inference methods for language-specific communicative needs when word frequencies are unknown (e.g. in the WCS). Ground truth distributions of communicative need were generated with spatial variation up to a given cutoff spatial frequency (x-axis), and scaled such that the entropy was held constant across all generated examples. Accuracy of inference (y-axis; lower is better) is measured as the KL-divergence between inferred versus ground truth communicative needs, for each example, expressed as a percentage of the KL-divergence between uniform and ground truth (dashed line at 100%). Curves are shown for inferences based on rate-distortion vocabularies for 2, 4, 6, 8, and 10 "words" (number of categories), when word-frequencies are unknown. **(A)** Inferences using the method proposed in this paper (Twomey et al); **(B)** based on the word-frequency (WF) method of Zaslavsky et al. (17); and **(C)** based on the CAP approach of Zaslavsky et al. (7). For language-specific inferences, our method achieves high accuracy for recovering low-frequency information about the true communicative need, and it does so in a manner that is invariant to the number of available "words" (categories).

4C. Systematic comparison of approaches for language-specific needs. The examples shown in SI Fig. S9 and earlier in SI Fig. S1 illustrate the relative performance of our method, the WF approach, and CAP, under a few ideal test conditions. Next, we conduct a systematic investigation of these three methods controlling for the level of spatial detail in the ground truth distributions of communicative need, and withholding information about $p(\hat{x})$, which is unknown for virtually all WCS languages, from all inference methods. Ground truth distributions were generated on a log scale from sinusoids with random phase and amplitude below a given spatial frequency. Distributions were scaled such that entropy was held constant across all generated examples for all spatial frequency cutoffs. This scaling fixes the KL-divergence between each generated ground truth distribution and uniformity at a constant, providing a consistent scale across examples.

SI Fig. S10 shows the KL-divergence between the distributions of communicative need recovered by each method and the generated ground truth distribution, as a percentage of the divergence between uniformity and ground truth, averaged over 500 examples at each spatial frequency cutoff. Our method (SI Fig. S10A) recovers the low-frequency features of the ground truth distribution even when $p(\hat{x})$ is unknown (assumed uniform) and in a manner that is relatively insensitive to the number of categories (number of terms) of the RDBC on which it is based. High-frequency information is lost, as expected, though recovery of low-frequency information across all spatial frequency cutoffs always improves estimates over uniform.

By contrast, the WF method (SI Fig. S10B) is highly sensitive to the number of categories available to it through $p(\hat{x}|x)$, and it requires a large number of categories to improve by even 40% over uniform when $p(\hat{x})$ is unknown (assumed uniform). High spatial frequency information is lost, and when word frequencies are unknown this apparently inhibits recovery of low-frequency information as well. Performance at intermediate scales can approach our method for large enough vocabularies (number of categories), but even at these scales our method provides consistently better performance across vocabulary sizes. Inferences by CAP (SI Fig. S10C) never improve over uniform and they are highly sensitive to vocabulary size (number of categories).

5. Language-specific inferences of communicative needs

In the following figures (SI Fig. S11–S27), we show each of the 130 language-specific distributions of communicative need we inferred using our method, for all languages recorded in the WCS and B&K survey data. We quantified the uncertainty in our inferences of communicative need by resampling with replacement from WCS participant focal color choices, recalculating mean focal color positions, and re-running our inference algorithm, 100 times per language (13,000 times in total). These bootstrapped distributions of communicative need for each language were used to highlight regions of color space that deviate significantly from the average communicative need across languages.

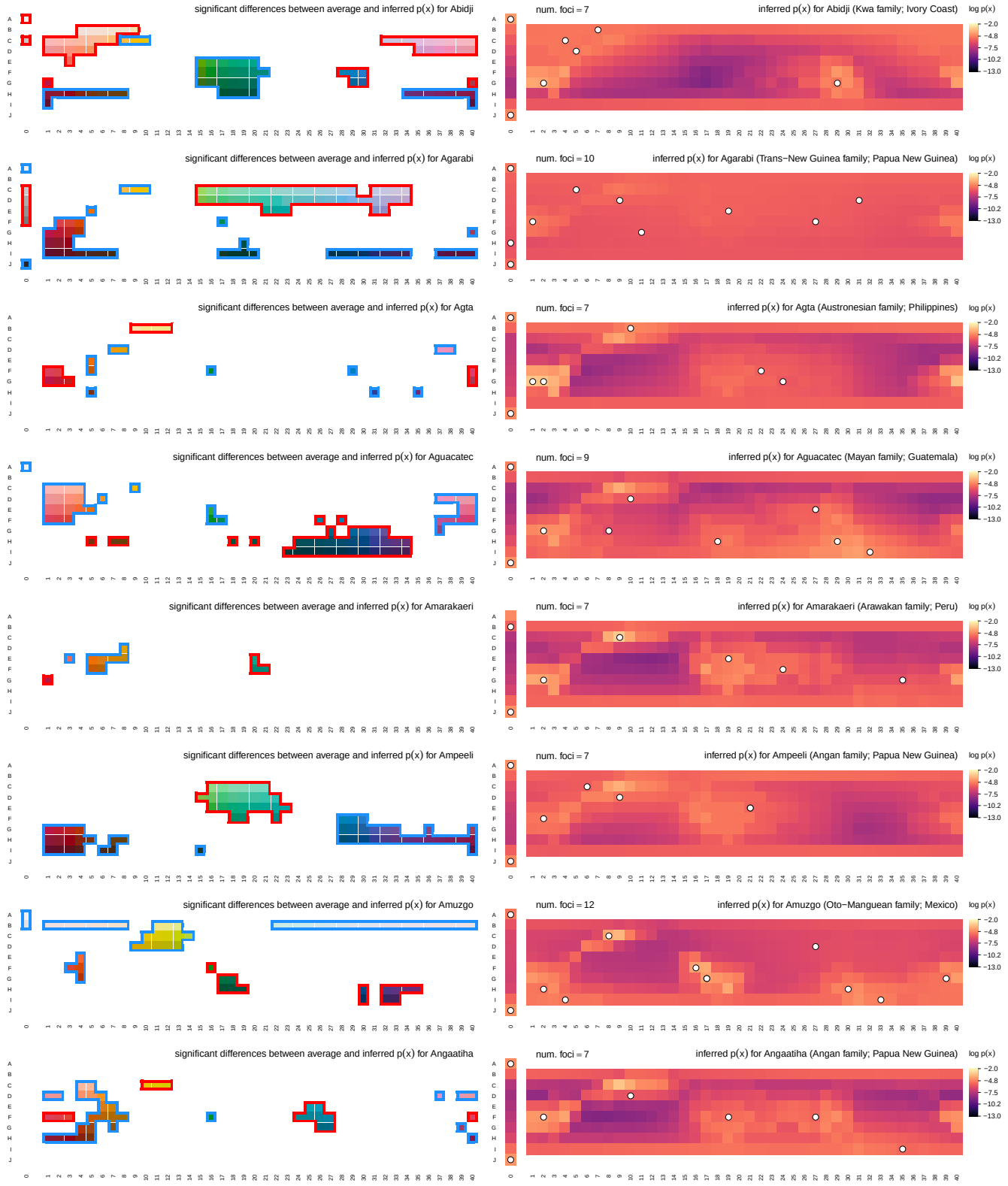

Fig. S11. Inferred communicative needs for 130 languages on a common scale. Each row corresponds to a language in the combined WCS+B&K survey data. (*Left column*) Significant differences between language-specific and across-language average communicative needs, shown as in Fig. 5. Deviations that exceed $\sigma/2$ with 95% confidence are highlighted in red (elevated) or blue (suppressed). (*Right column*) Language-specific communicative needs (log scale) shown with language focal color positions projected on to the WCS color chips (white points). Focal points may overlap on the same color chip.

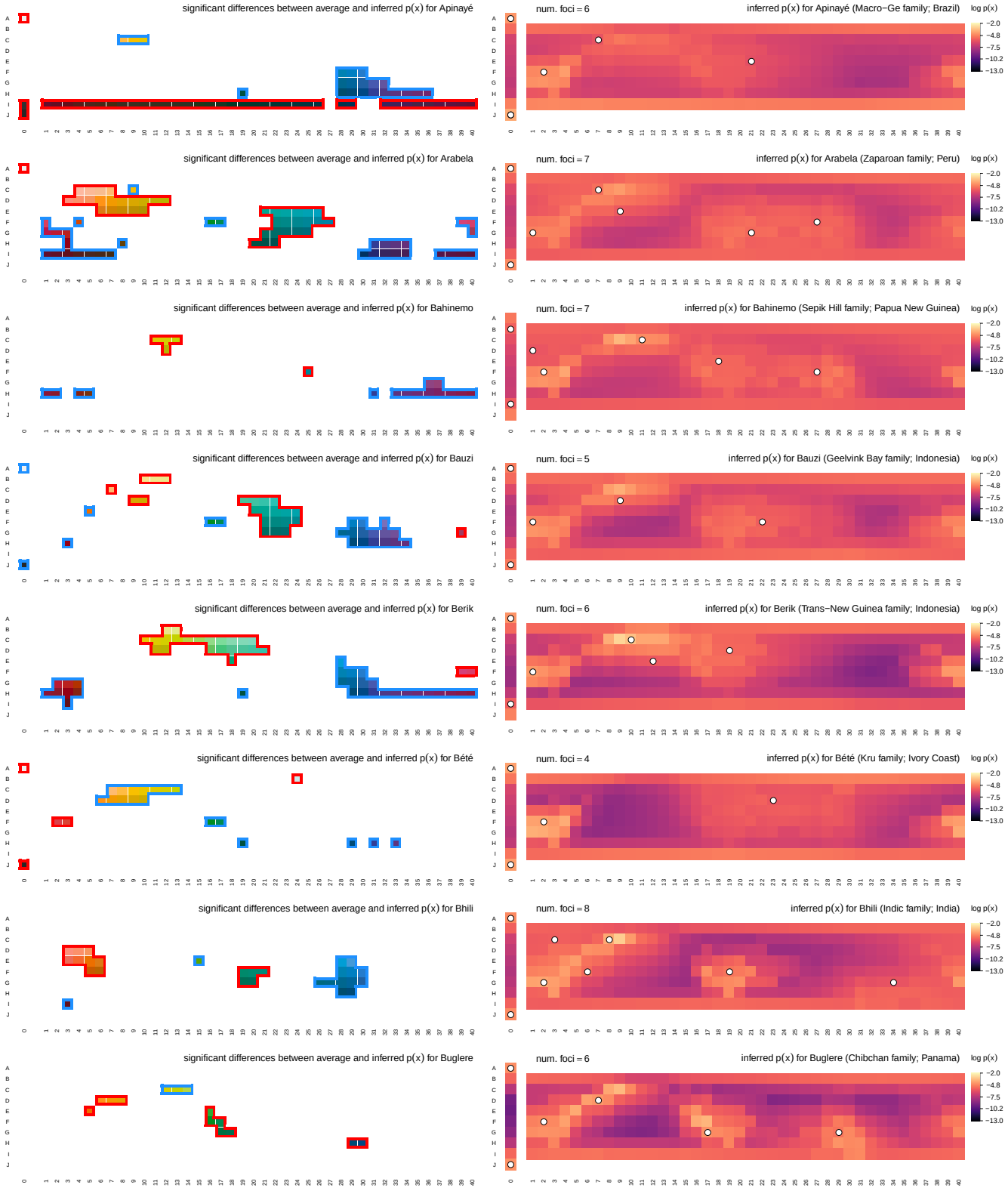

Fig. S12. Inferred communicative needs for 130 languages on a common scale (continued).

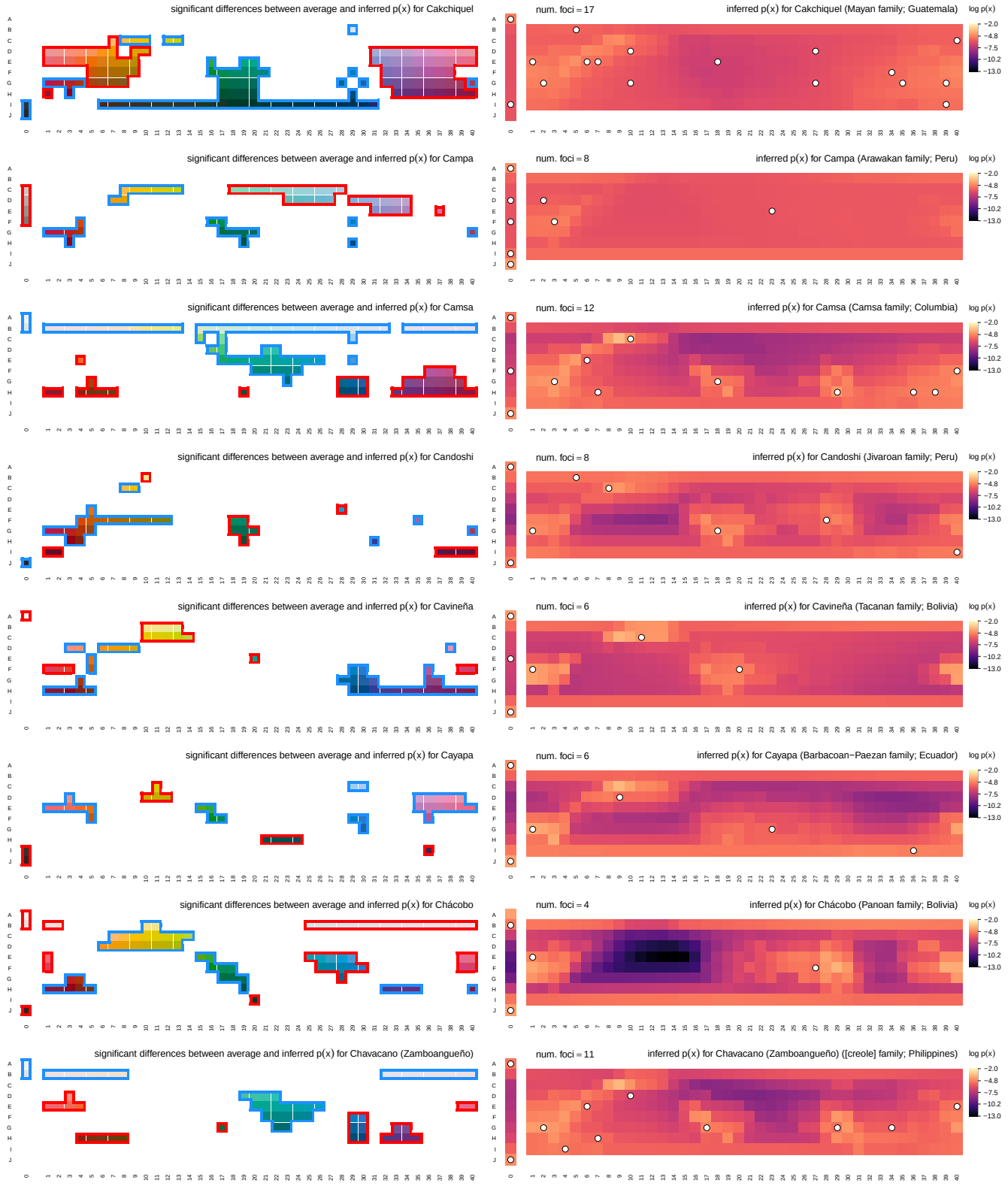

Fig. S13. Inferred communicative needs for 130 languages on a common scale (continued).

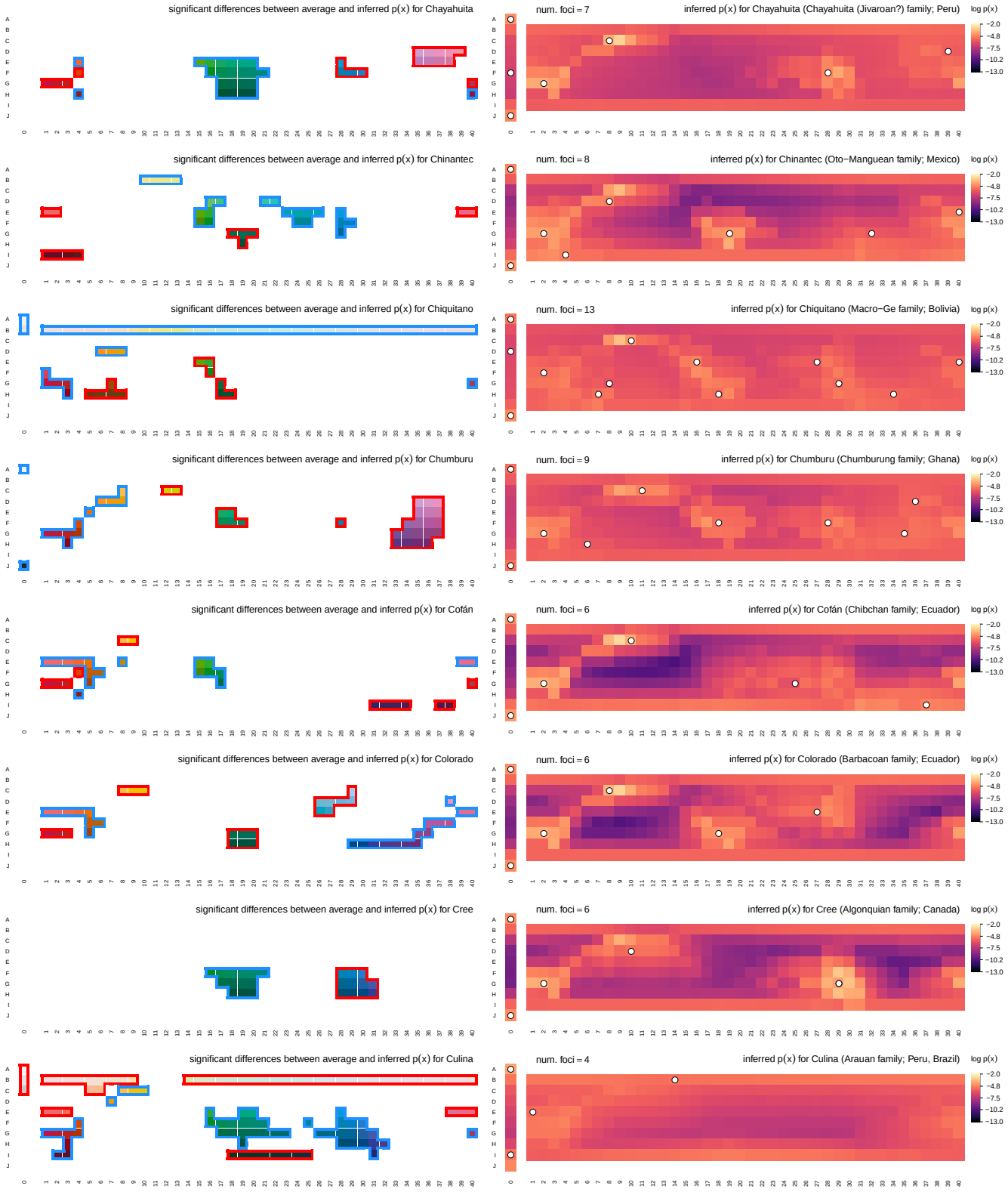

Fig. S14. Inferred communicative needs for 130 languages on a common scale (continued).

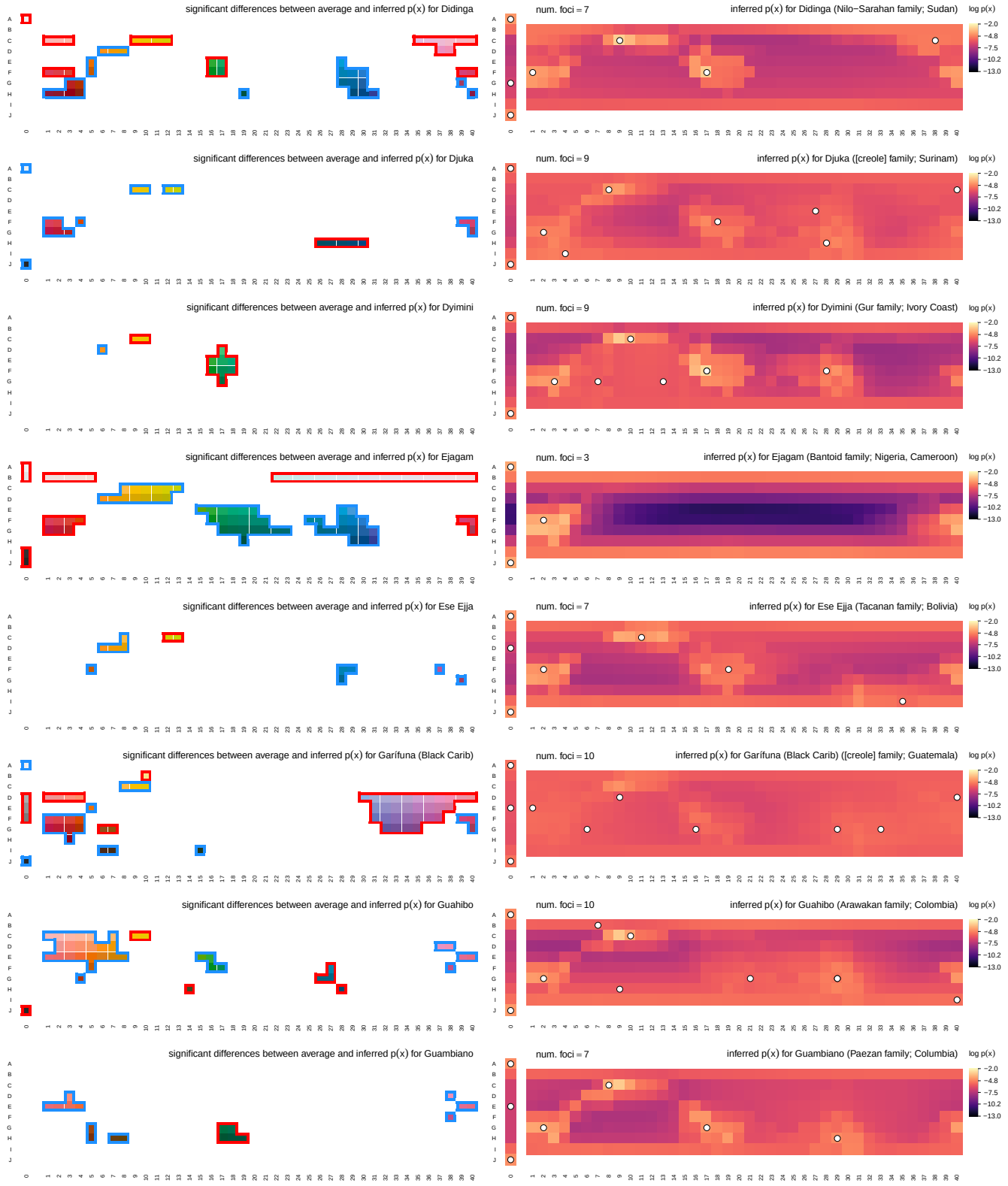

Fig. S15. Inferred communicative needs for 130 languages on a common scale (continued).

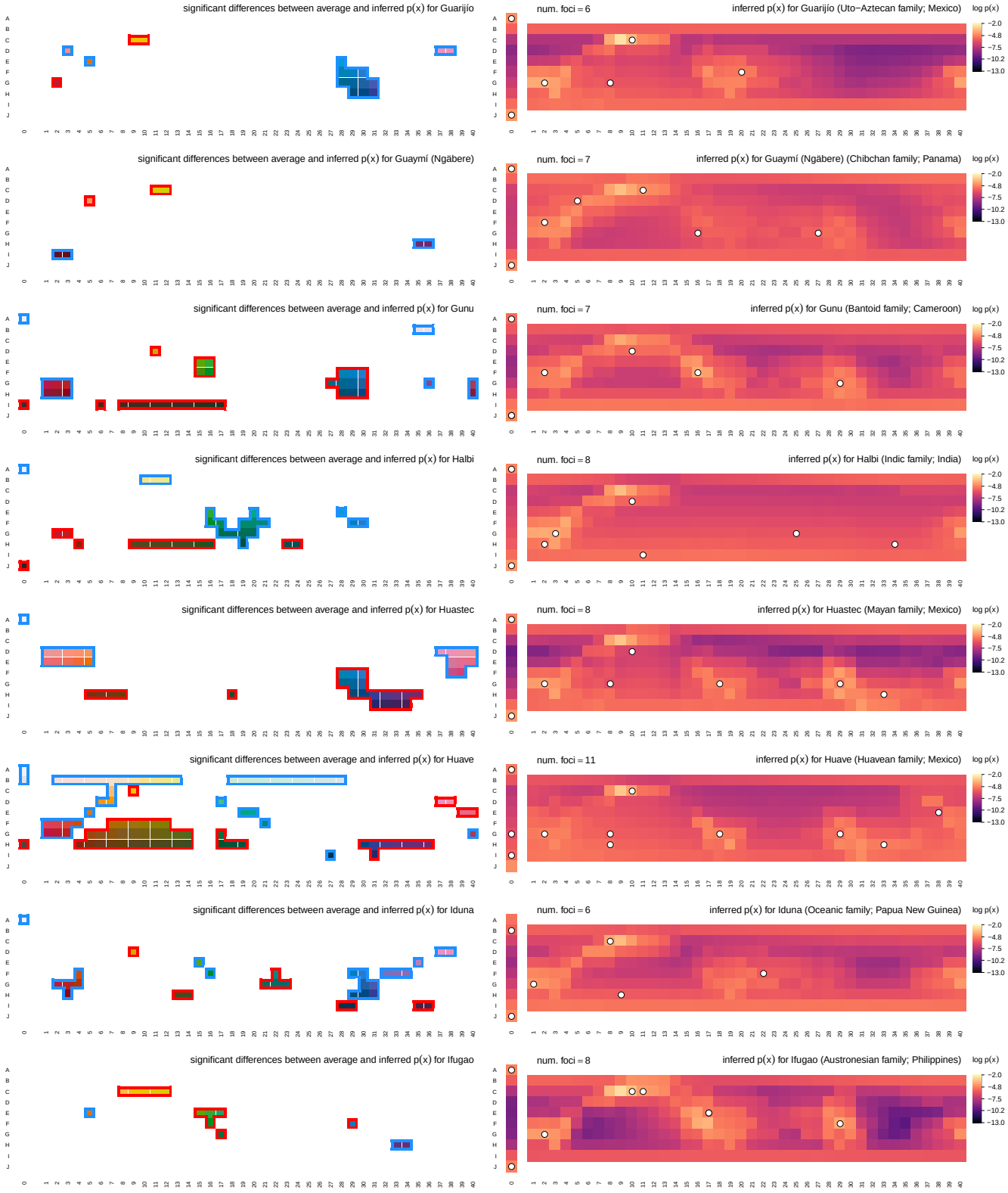

Fig. S16. Inferred communicative needs for 130 languages on a common scale (continued).

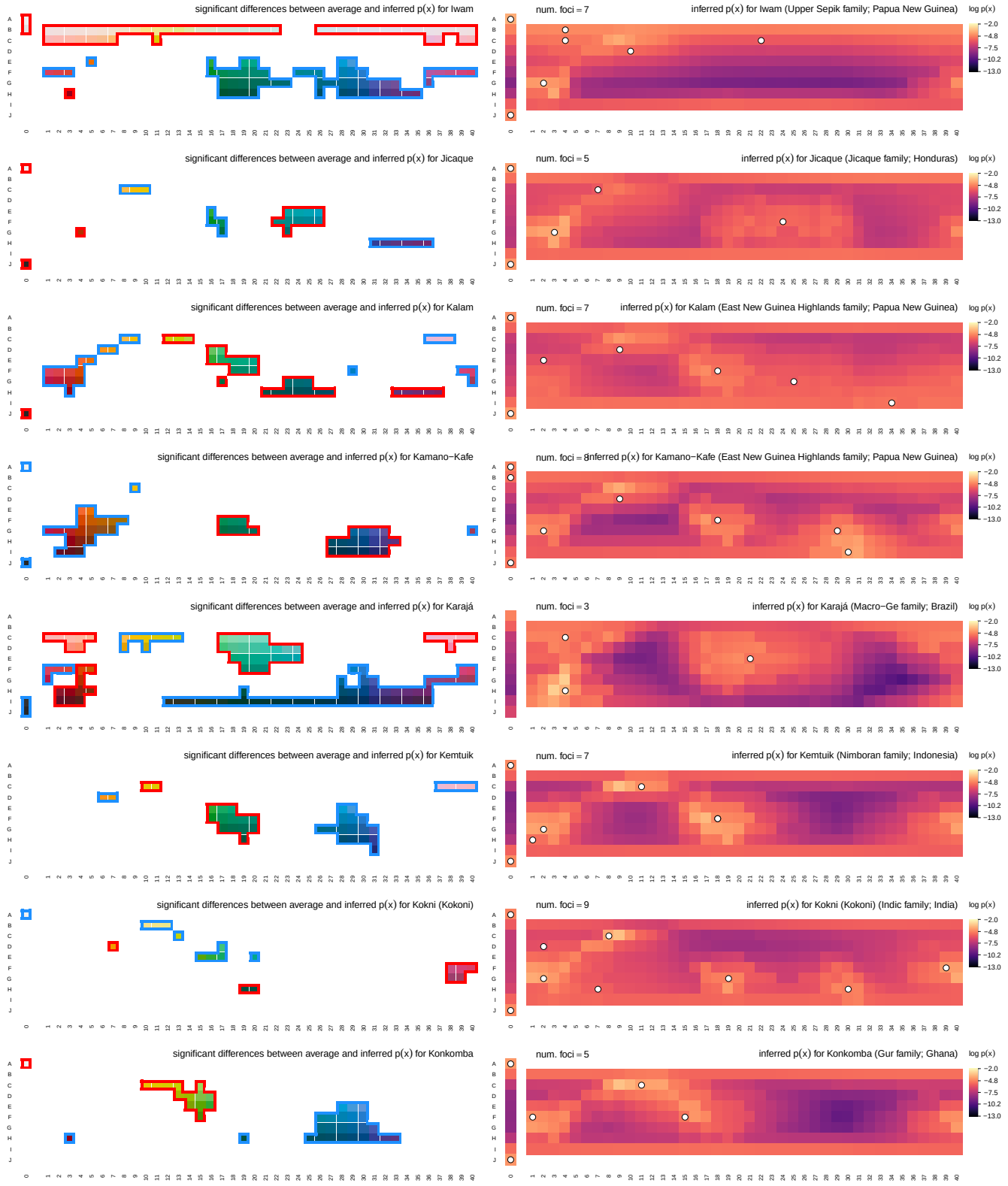

Fig. S17. Inferred communicative needs for 130 languages on a common scale (continued).

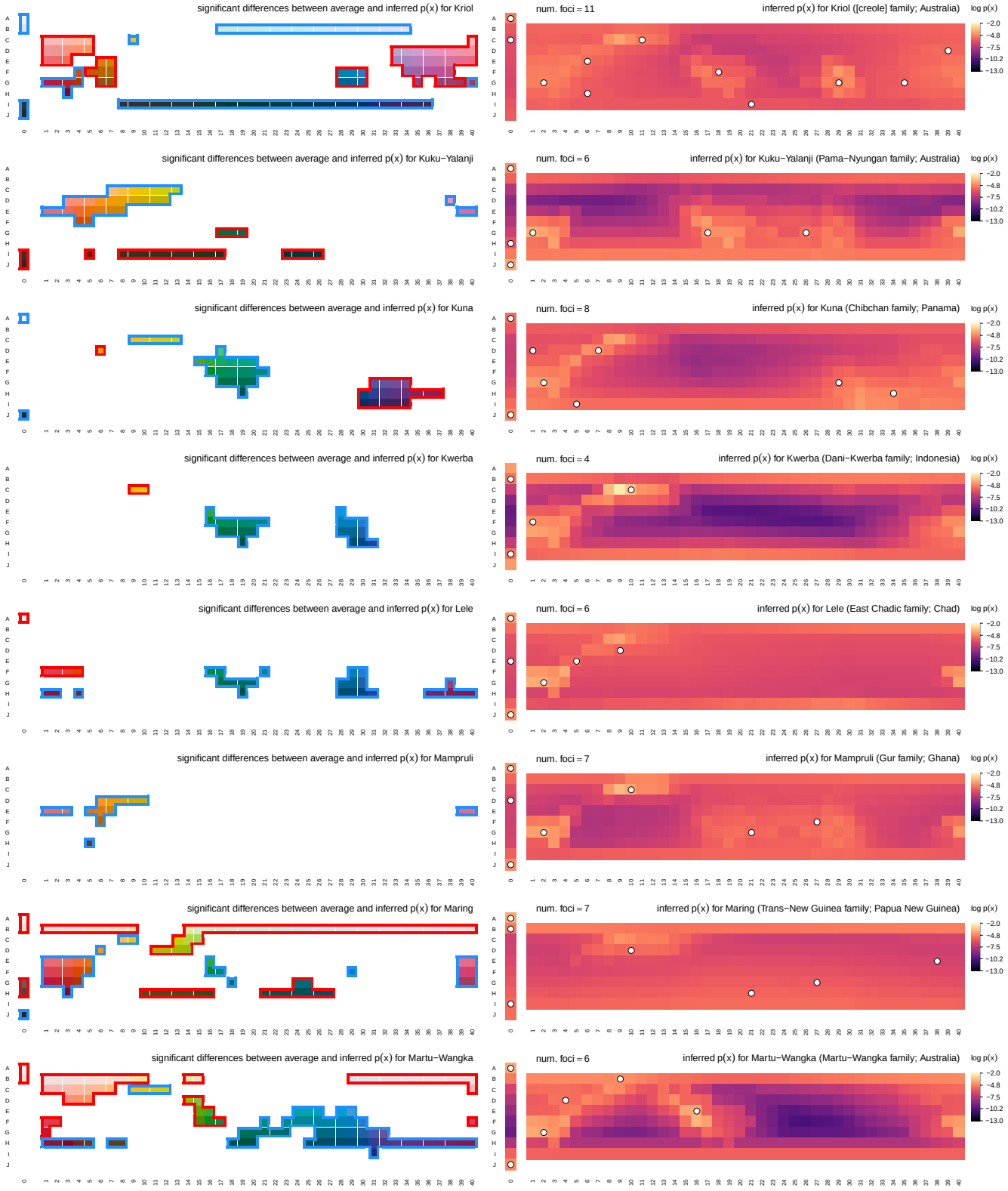

Fig. S18. Inferred communicative needs for 130 languages on a common scale (continued).

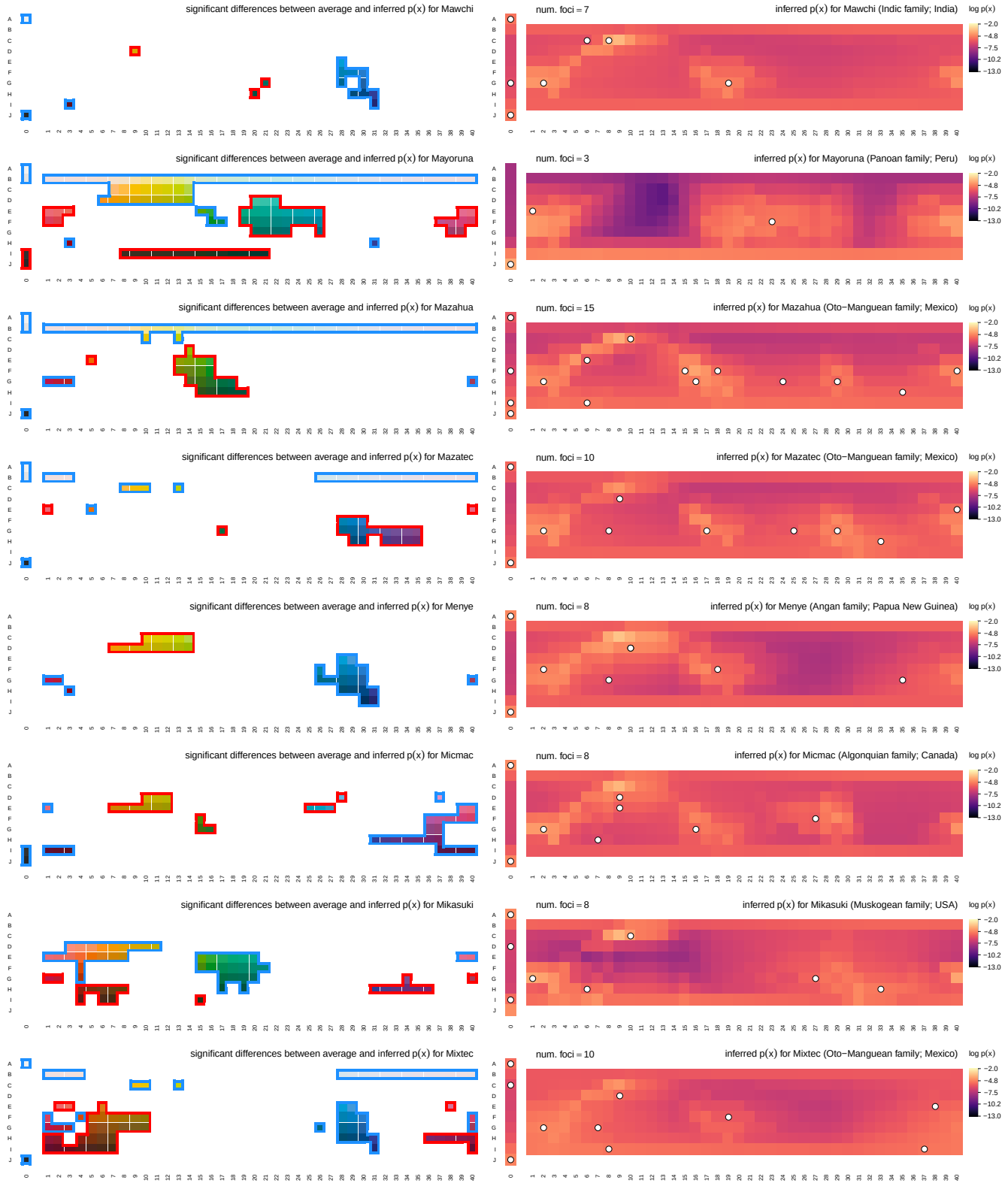

Fig. S19. Inferred communicative needs for 130 languages on a common scale (continued).

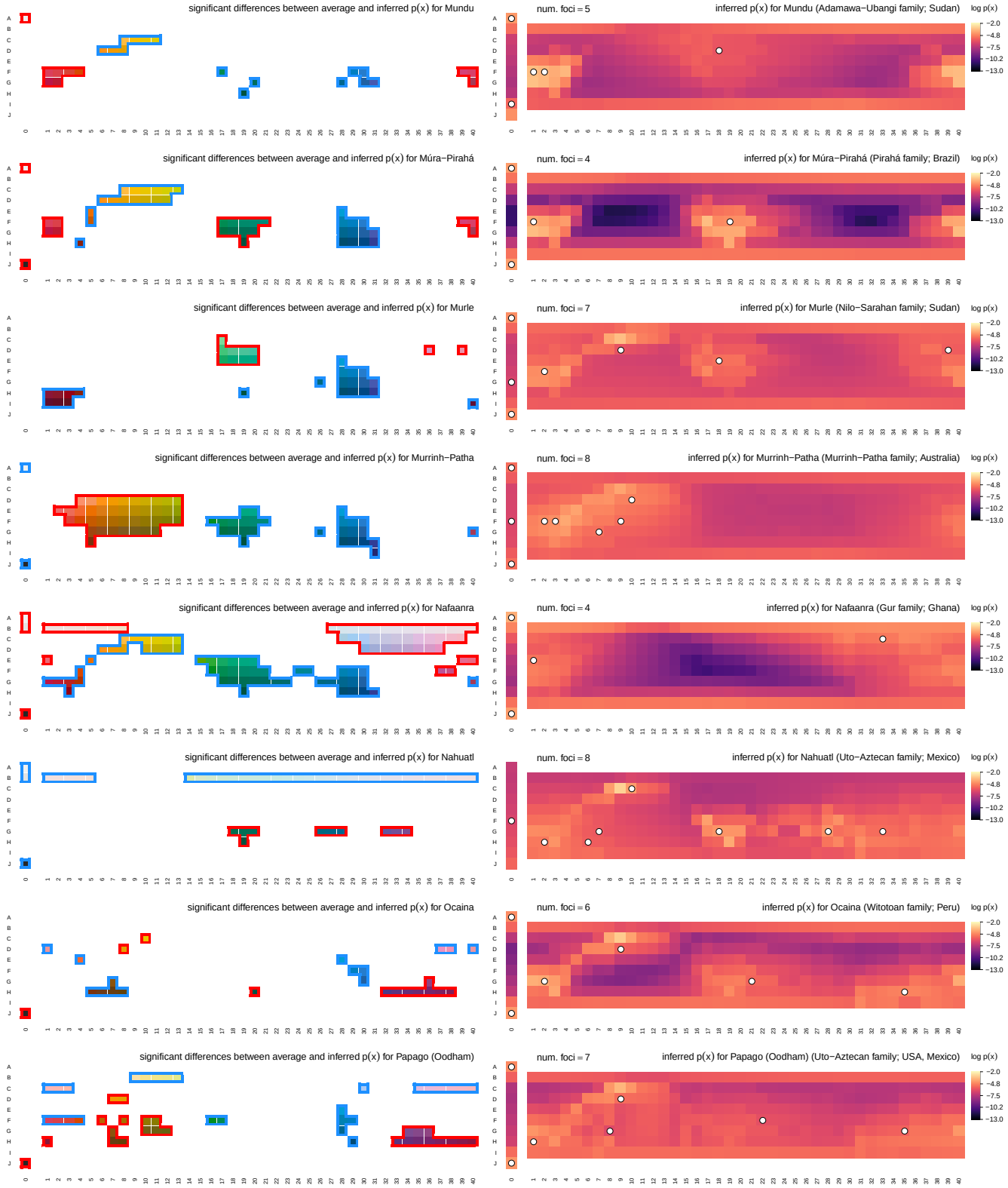

Fig. S20. Inferred communicative needs for 130 languages on a common scale (continued).

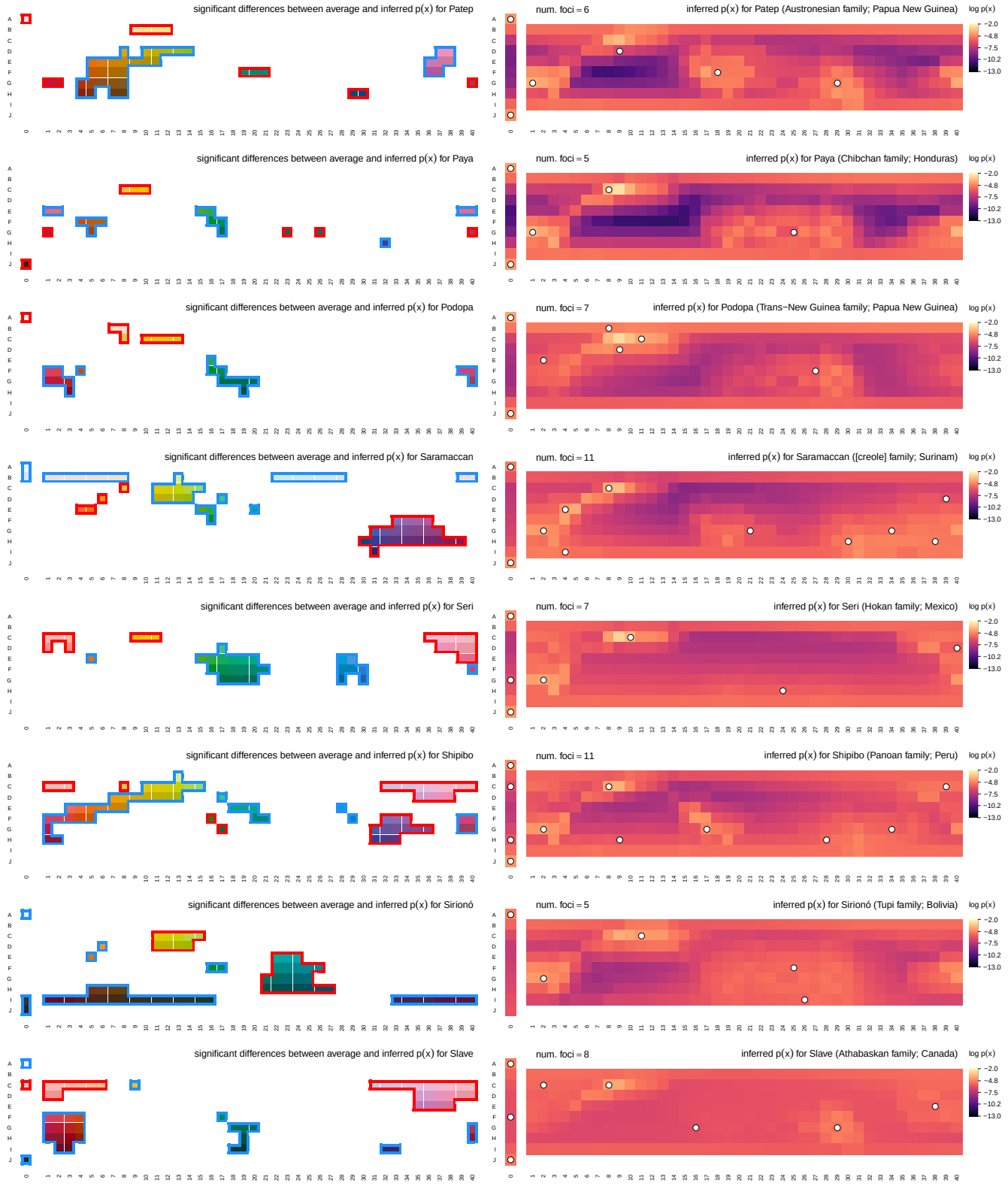

Fig. S21. Inferred communicative needs for 130 languages on a common scale (continued).

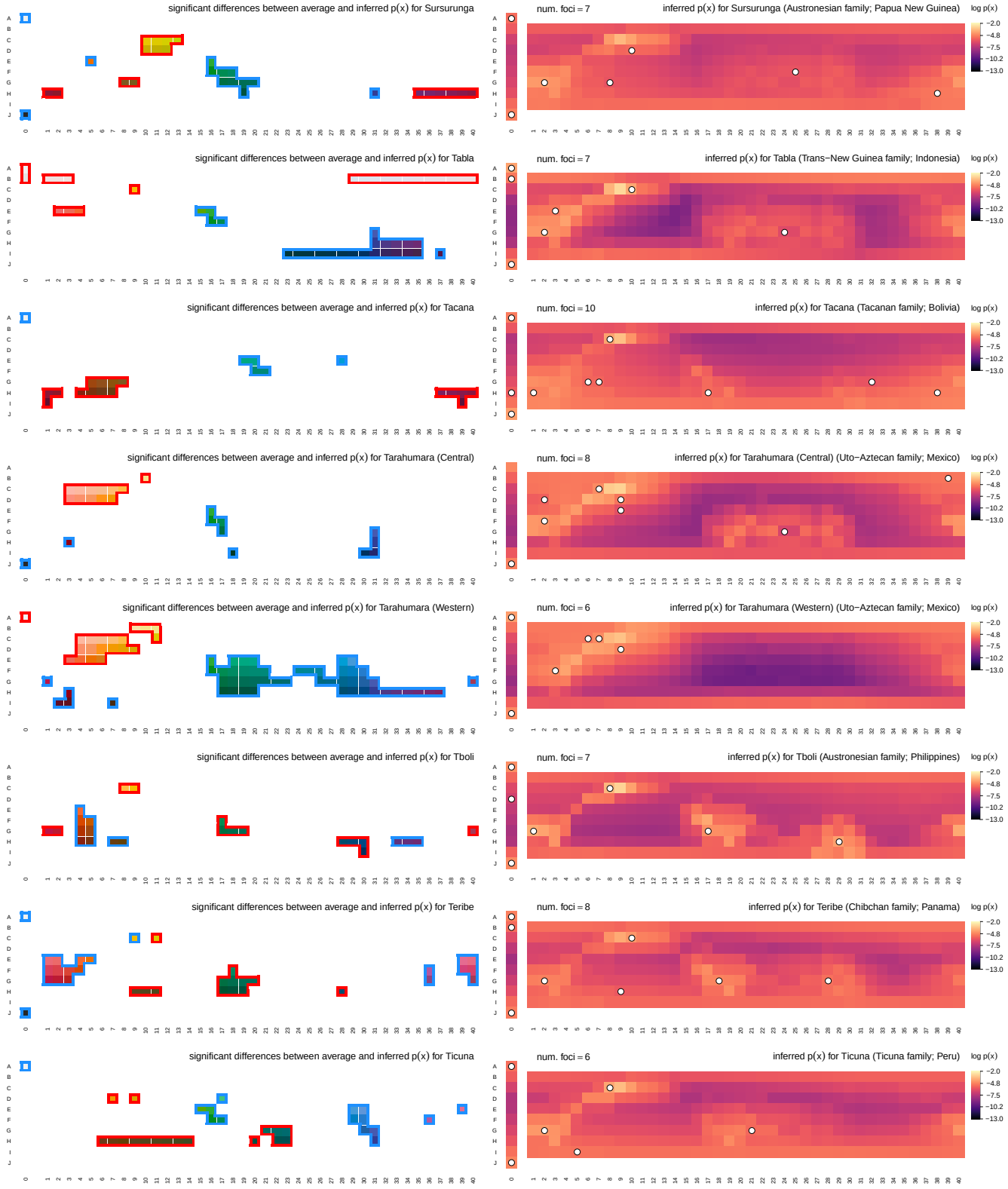

Fig. S22. Inferred communicative needs for 130 languages on a common scale (continued).

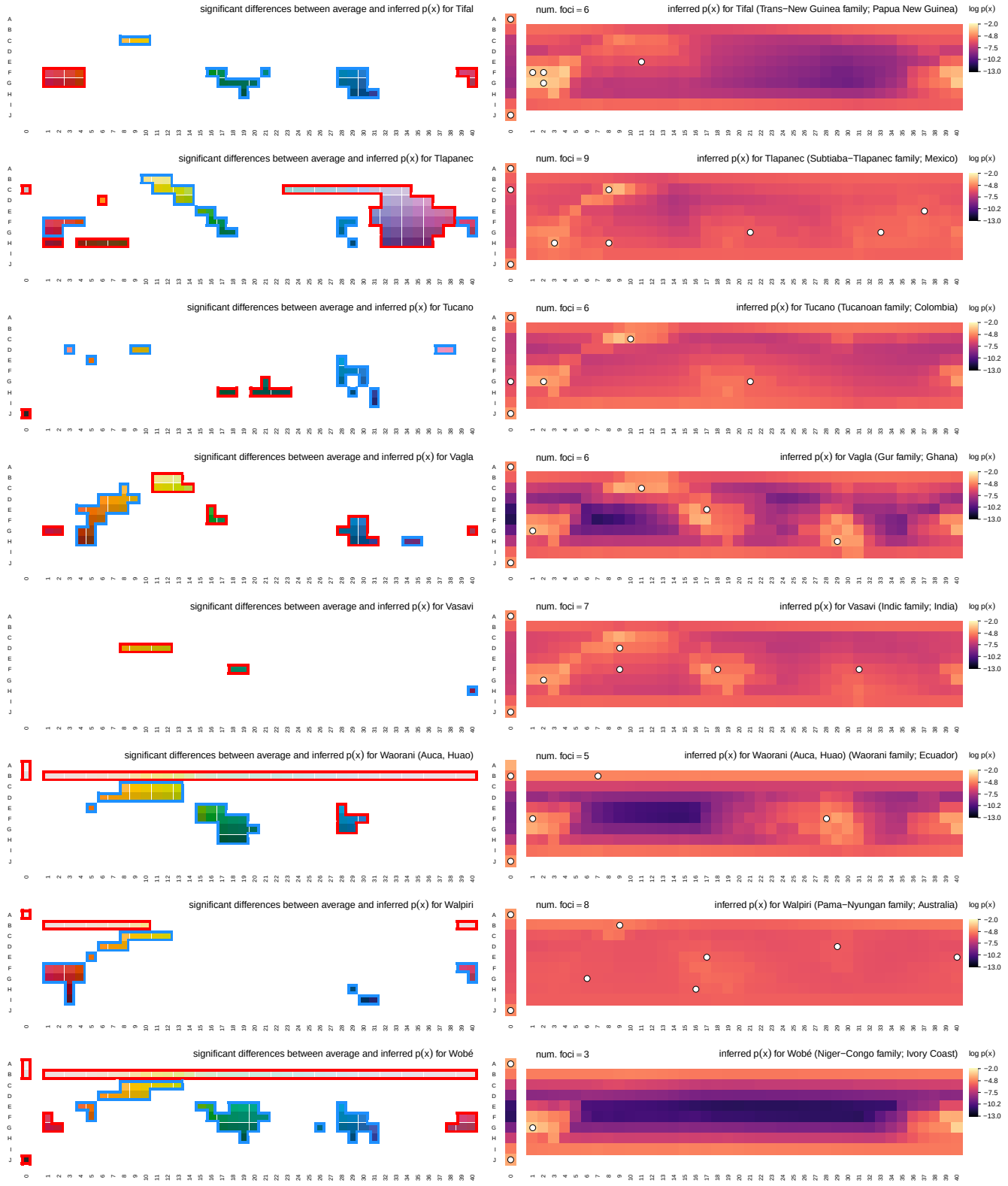

Fig. S23. Inferred communicative needs for 130 languages on a common scale (continued).

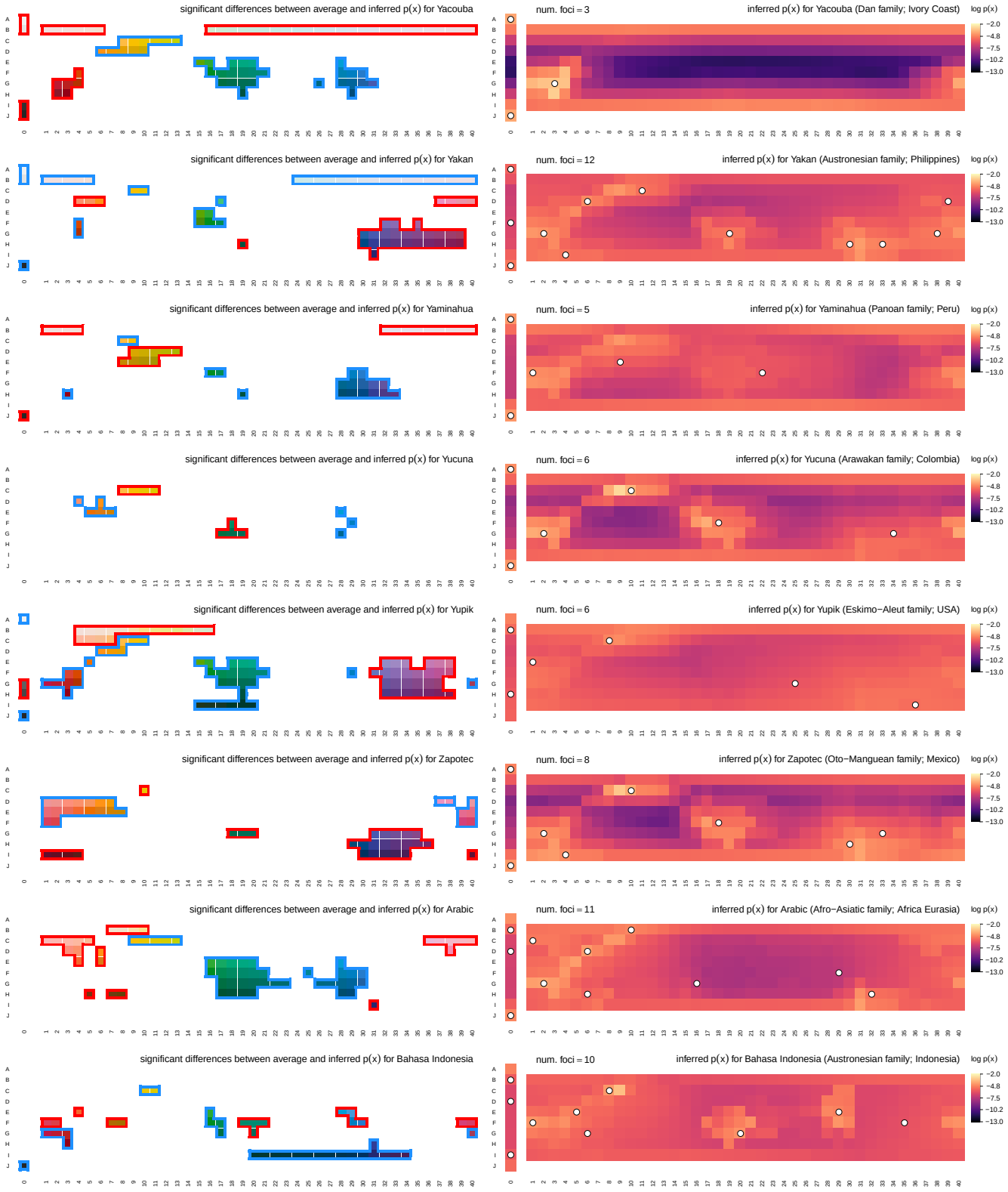

Fig. S24. Inferred communicative needs for 130 languages on a common scale (continued).

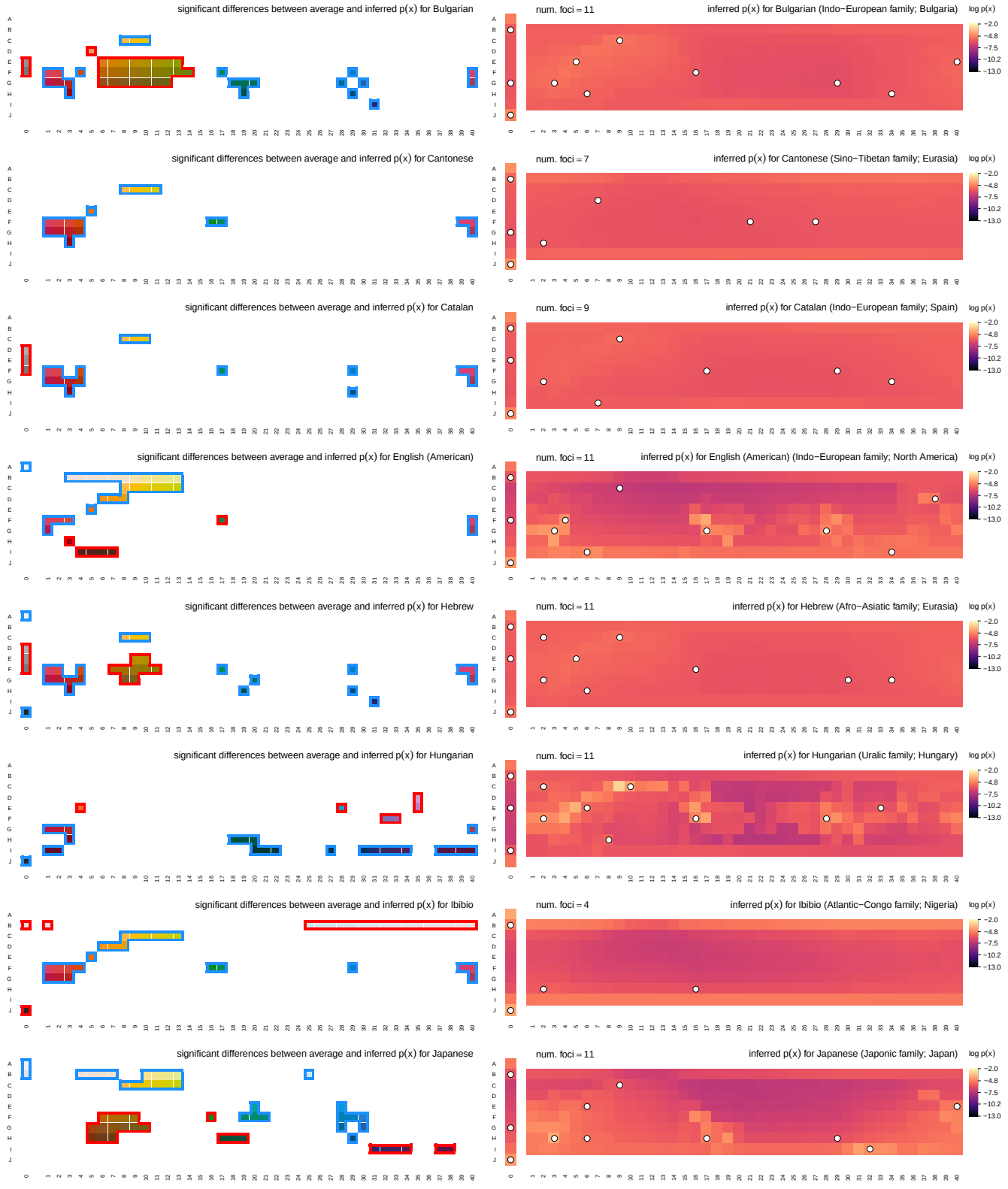

Fig. S25. Inferred communicative needs for 130 languages on a common scale (continued).

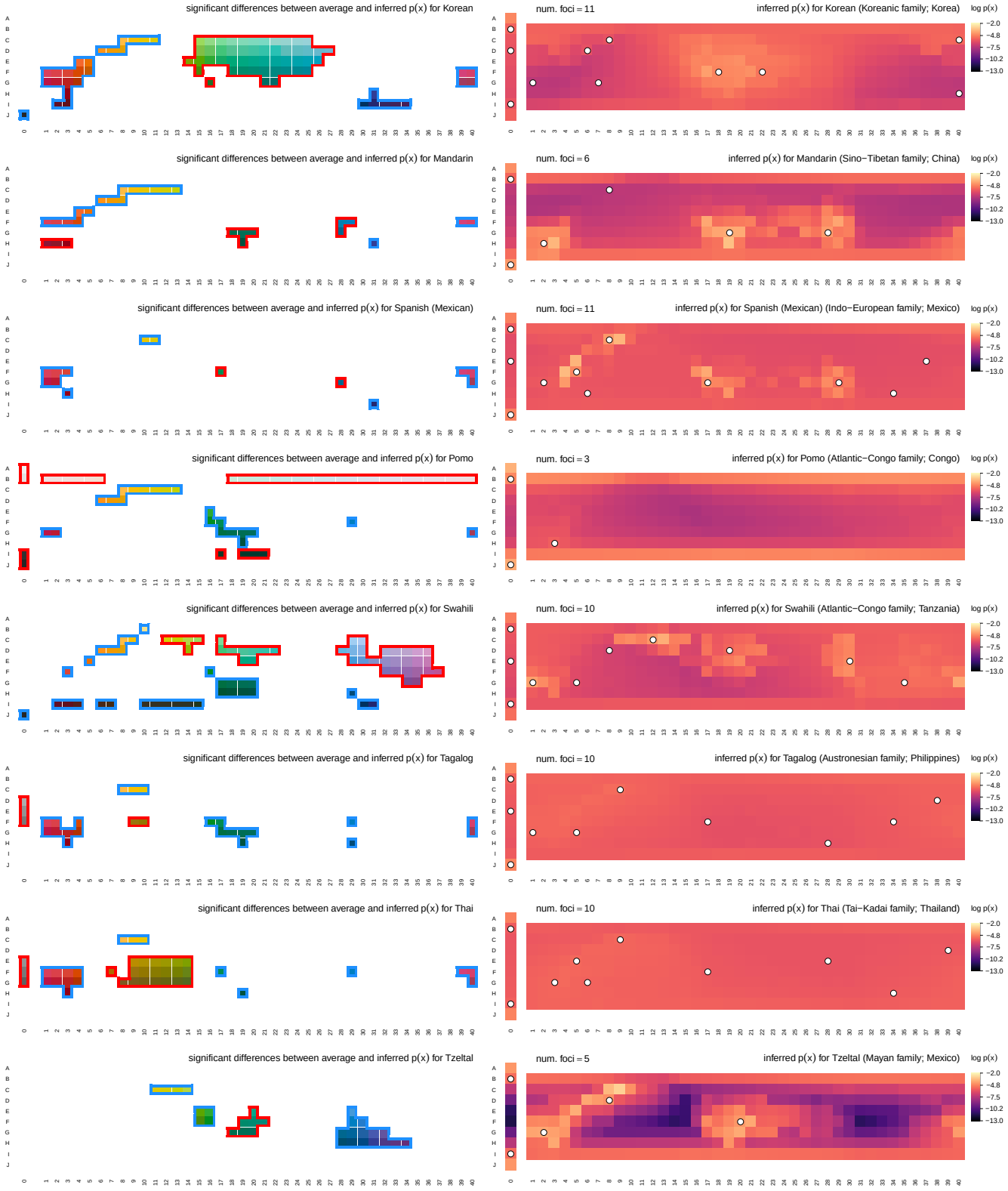

Fig. S26. Inferred communicative needs for 130 languages on a common scale (continued).

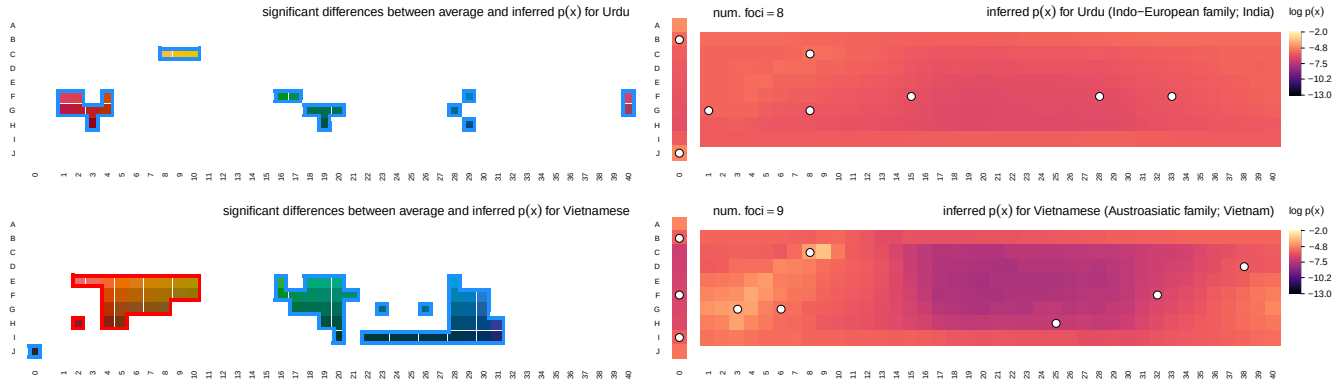

Fig. S27. Inferred communicative needs for 130 languages on a common scale (continued).

SI Dataset S1 (communicative_needs.csv)
Language-specific inferred distributions, $p(x)$, for each language in the combined WCS+B&K, with one row for each language. The `lnum` field for the first 1–110 rows correspond to the `lnum` fields for the languages in the WCS data archived (http://www1.icsi.berkeley.edu/wcs/data.html). The remaining 20 rows, 111–130, correspond to the 20 languages in the B&K dataset (available in the same archive as the WCS data above). Subtracting 110 from the `lnum`s for these languages gives the corresponding `lnum`s for the B&K dataset. The name of each language is provided as the second column in the CSV file. The remaining columns, numbered 1–330 corresponding to their `cnum` in the WCS, contain the language specific inferred values $p(x)$ for each language and color chip.

SI Dataset S2 (average_communicative_needs.csv)
The average inferred distribution across all 130 languages. Each row corresponds to a single color chip in the WCS data, indexed by the `cnum` field. The value of the average inferred distribution for each chip is provided in the `average` column. The values of `cnum` correspond to the chip numbering used in `cnum-vhcm-lab-new.txt` from the World Color Survey archive (http://www1.icsi.berkeley.edu/wcs/data/cnum-maps/cnum-vhcm-lab-new.txt).

SI Dataset S3 (bootstrapped_communicative_needs.csv)
Distributions of language-specific communicative needs inferred by resampling language focal colors with replacement. The structure of the data is the same as for SI Dataset S1, except with 100 consecutive rows per language.

References

1. CE Shannon, A mathematical theory of communication. *Bell Syst. Tech. J.* **27**, 379–423 (1948).
2. CE Shannon, Coding theorems for a discrete source with a fidelity criterion. *IRE Natl. Conv. Rec.* **7**, 142–163 (1959).
3. A Banerjee, S Merugu, IS Dhillon, J Ghosh, Clustering with bregman divergences. *J. Mach. Learn. Res.* **6**, 1705–1749 (2005).
4. SN Yendrikhovskij, Computing color categories from statistics of natural images. *J. Imaging Sci. Technol.* **45**, 409–417 (2001).
5. L Steels, T Belpaeme, Coordinating perceptually grounded categories through language: a case study for colour. *Behav. Brain Sci.* **28**, 469–489 (2005).
6. T Regier, P Kay, N Khetarpal, Color naming reflects optimal partitions of color space. *PNAS* **104**, 1436–1441 (2007).
7. N Zaslavsky, C Kemp, T Regier, N Tishby, Efficient compression in color naming and its evolution. *PNAS* **115**, 7937–7942 (2018).
8. RM Boynton, CX Olson, Locating basic colors in the osa space. *Color. Res. Appl.* **12**, 94–105 (1987).
9. J Sturges, TWA Whitfield, Locating basic colours in the munsell space. *Color. Res. Appl.* **20**, 364–376 (1995).
10. DT Lindsey, AM Brown, The color lexicon of american english. *J. Vis.* **14**, 17, 1–25 (2014).
11. JT Abbott, TL Griffiths, T Regier, Focal colors across languages are representative members of color categories. *PNAS* **113**, 11178–11183 (2016).
12. A Agarwal, H Daumé III, A geometric view of conjugate priors. *Mach Learn.* **81**, 99–113 (2010).
13. S Arimoto, An algorithm for computing the capacity of arbitrary discrete memoryless channels. *IEEE Trans. Inf. Theory* **18**, 14–20 (1972).
14. R Blahut, Computation of channel capacity and rate-distortion function. *IEEE Trans. Inf. Theory* **18**, 460–473 (1972).
15. I Csiszár, G Tusnády, Information geometry and alternating minimization procedures. *Stat. Decis.* **Supplement Issue 1**, 205–237 (1984).
16. CL Byrne, *Iterative optimization in inverse problems.* (CRC Press, Boca Raton, FL), (2014).
17. N Zaslavsky, C Kemp, N Tishby, T Regier, Communicative need in colour naming. *Cogn. Neuropsychol.* **37**, 312–324 (2019).
18. MJD Powell, The BOBYQA algorithm for bound constrained optimization without derivatives, (Department of Applied Mathematics and Theoretical Physics, Cambridge University, UK), Technical report (2009).
19. P Kay, B Berlin, L Maffi, W Merrifield, Color naming across languages in *Color categories in thought and language*, eds. CL Hardin, L Maffi. (Cambridge University Press, Cambridge), pp. 21–56 (1997).
20. E Gibson, et al., Color naming across languages reflects color use. *PNAS* **114**, 10785–10790 (2017).
21. DL Everett, Cultural constraints on grammar and cognition in pirahã: another look at the design features of human language. *Curr. Anthropol.* **46**, 621–646 (2005).
22. A Wierzbicka, Why there are no 'colour universals' in language and thought. *J. R. Anthropol. Inst.* **14**, 407–425 (2008).
23. T Regier, P Kay, N Khetarpal, Color naming and the shape of color space. *Language* **85**, 884–892 (2009).
24. P Sumner, JD Mollon, Catarrhine photopigments are optimized for detecting targets against a foliage background. *J. Exp. Biol.* **203**, 1963–1986 (2000).
25. P Sumner, JD Mollon, Chromaticity as a signal of ripeness in fruits taken by primates. *J. Exp. Biol.* **203**, 1987–2000 (2000).
26. C Witzel, Variation of saturation across hue affects unique and typical hue choices. *i-Perception* **10** (2019).
27. N Zaslavsky, C Kemp, N Tishby, T Regier, Color naming reflects both perceptual structure and communicative need. *Cog. Sci.* **11**, 207–219 (2018).