An affirmative answer to John Locke's question of colour scrambling

Donald D. Hoffman[1] & Ann K. Carruthers[2]

[1]*Department of Cognitive Science, University of California, Irvine, California 92697*

[2]*Department of Logic and Philosophy of Science, University of California, Irvine, California 92697*

Is it possible, John Locke pondered in his *Essay* of 1690, that "the idea that a violet produced in one man's mind by his eyes were the same that a marigold produced in another man's, and vice versa."?[1] Could the colours I experience differ from yours, even if experiments reveal no difference between us? Locke's question is raised by inquisitive children, but remains hotly debated by scientists and philosophers because its answer is key to an unsolved mystery of science: What is the relationship between brain activity and conscious experience?[2–8] How, precisely, can brain activity cause, or be, my experience of chartreuse or the fragrance of a rose?[9,10] If colour scrambling of the type Locke envisioned is possible, this would notably constrain the empirical search for relationships between brain activity and consciousness. It would entail that conscious experiences could change without concomitant functional changes in brain states. Locke's question has stirred prolific debate through the ensuing centuries, but no mathematical articulation or proof. Here we prove that the answer is "Yes": colour scrambling is possible without violating scientific laws.

Functionalist theories of conscious experience propose that experience is determined by the functional architecture of a system: its inputs, outputs, internal states, and rules for changing state.[11] Although in humans this architecture is implemented in the nervous system, it could equally well be implemented in other physical systems, such as computers. Proponents of functionalism typically deny that the experiences of one person could be scrambled from those of another with no experimental consequences.[3,4,10,12] For if functional architecture determines conscious experience, then scrambling experiences would require changing functional architecture. Such changes would evoke differences between the two persons in controlled experiments. Some proponents of functionalism assert further that a clear sense to the question of colour scrambling has not yet been articulated.[13]

If, however, colour scrambling is well-defined and possible without violating scientific laws, then functional architecture does not determine conscious experience, and we must look elsewhere, perhaps in quantum processes,[14–16] for physical correlates of consciousness, or countenance fundamentally different approaches to the study of consciousness.[17] Thus it is crucial to articulate a clear sense and definitive answer to Locke's question.

To do this, let us denote possible colour experiences of Jack by *X* and those of Jill by *Y*. We wish to assume as little as possible about these experiences so that our formulation applies not just to colour but to all, conscious, sensory experiences. Such experiences can be more or less certain: thus, at a minimum, we must represent *probabilities* of sensory events, e.g., the probability that Jack sees orange. The probabilistic nature of sensory experiences is a basic assumption universally employed by psychophysicists who use Signal Detection Theory[18,19] and by researchers in

computational vision who model perception as a process of Bayesian inference.[20,21] Therefore, following the standard theory of probability, we assume that X and Y are probability spaces, in which events of X are certain subsets of X.[22] In this setting, we can consider the probabilities, p, of various colour events of X.

We also wish to assume as little as possible about the scrambling function, f, that maps Jack's colour experiences X to Jill's colour experiences Y. At a minimum, we need this scrambling function to respect the structure of colour events for X and Y, so that statements about probabilities for Jill's colour events can be translated into corresponding statements for Jack's colour events. A scrambling function that does this is called a measurable function or, if it is real-valued, a random variable.[23]

Note the generality of our setting. The sensory spaces X and Y need not have a dimension. The scrambling function f need not be linear or continuous. Indeed, for the notions of dimension, linearity or continuity to be defined would require a setting less general than ours.

A shade of red is more readily discriminated from green than from another shade of red. We must model all such empirical similarities and differences between colour experiences. All of them must be preserved by our function that scrambles colour experiences between Jack and Jill, otherwise controlled experiments could detect the scrambling.

One way to describe these similarities and differences is to define a distance metric, d, between colour events such that, the more similar two events are, the less distance lies between them. Again we want to assume as little as possible. We have a probability, p, for Jack's sensory events, and we wish to assume nothing more about them. Therefore we derive our notion of distance solely from this probability.

This can be done simply. As illustrated in Figure 2, let event *A* represent one colour experience and *B* a second colour experience. The symmetric difference of *A* and *B*, denoted $A \Delta B$, is the blue-shaded region in Figure 2. Intuitively, it is the part of *A* that is outside of *B,* plus the part of *B* that is outside of *A* . Then the distance, *d(A, B)*, between colour experience *A* and colour experience *B* is just the probability *p* of this symmetric difference.[24,25] That is, $d(A,B)=p(A \Delta B)$. This probability-of-symmetric-difference (PSD) metric specifies the distances between all pairs of colour experiences, not just a particular pair *A* and *B*, and therefore captures all empirically measurable similarities and differences among these experiences. According to functionalism the distances, *d*, between colour experiences, but not the colour experiences themselves, enter into functional architecture.

Given the probability *p* of Jack's colour experiences *X* and given our scrambling function *f*, we can canonically transport *p* to Jill's colour experiences *Y*. This transport, which is a probability measure *q* on *Y*, is simply the distribution of *f*. Recall that, if *D* is any event for Jill's colour experiences *Y*, and if *C* is its corresponding unscrambled event for Jack's colour experiences *X*, then the probability *q(D)* is just *p(C)*.

We have canonically transported Jack's probabilities *p* of colour experiences *X* to Jill's probabilities *q* of colour experiences *Y*. Now we can derive from *q* a new distance metric, *d'*, on Jill's colour experiences *Y*, just as we derived from *p* the distance metric, *d*, on Jack's colour experiences *X*. And here is the striking result:

For *every* choice of probability *p* and *every* choice of scrambling *f*,

Jack's distance metric *d* and Jill's distance metric *d' always* agree.

A proof of this "Scrambling Theorem" is in the Methods section.

The Scrambling Theorem entails that, if A' and B' are any two sensory experiences for Jill, and if A and B are the corresponding unscrambled sensory experiences for Jack, then the distance between A' and B' for Jill is identical to the distance between A and B for Jack. Thus all perceptual experiments involving Jill evince the same results as for Jack. A simple illustration of this is shown in Figure 3a, where the colour experiences of Jill are scrambled relative to those of Jack, but all distances between corresponding experiences remain unchanged. This example uses no inputs from the world and no outputs from Jack or Jill; scrambling is possible nonetheless.

Figure 3b extends this example by adding inputs and outputs, and illustrates that Jack and Jill will apply the same colour names to the same objects, so that one cannot tell by talking with them that their experiences differ. This example uses language, but the same result applies to nonlinguistic experiments. Thus the colour experiences of Jack and Jill could be differently connected to the external world without any experimental divergences to betray that phenomenal difference. If Jack makes finer discriminations among external stimuli in the region of colour space called “green” than in the region called “blue”, so also will Jill, even if Jill's colour experience upon viewing grass is the same as Jack's upon viewing the sky. The nonuniformity of colour space,[26] or of any other phenomenal space, is no obstacle to application of the Scrambling Theorem. But if we suppose instead that Jack is red-green colour blind and Jill is not, then of course experiments would reveal differences between them, with no attendant violation of the Scrambling Theorem.

The Scrambling Theorem models subjective similarities between conscious experiences with the PSD metric, a metric that is derived solely from probabilities governing those experiences and, as described in the Methods section, that generalizes

standard measures of perceptual discrimination from Signal Detection Theory.[18,19] This metric, however, is not required for the Scrambling Theorem. The Theorem holds if, instead of probabilities of symmetric differences, one uses probabilities of *any* measurable function of the relevant events, or of their unions, intersections and differences.

One can obviate the Scrambling Theorem by denying that conscious experiences can be represented by any well-defined space. This deflationary tactic exacts a high price. It precludes a functionalist account of experience, since such accounts require well-defined states.[11] It also precludes using probabilities of experiences, a tool that perceptual psychologists and researchers in computational vision have found essential.[18–21]

Some functionalists assert that colour experiences are *identical* with certain functional brain states, and that therefore to ask why certain experiences are correlated with certain brain states is mistaken, for such correlations can be sought only for entities that are distinct.[27,28] The Scrambling Theorem puts the burden of proof squarely on the shoulders of such identity theorists, for it shows that any colour experience can, in principle, play the functional role of any other colour experience. Why then should we be tempted to assert that any one of these colour experiences is identical to a particular functional role? And whose colour experience–Jack's or Jill's–should we assert is identical to a particular functional role, or to a particular band of the electromagnetic spectrum?

We have shown that Locke's suggestion that colour experiences might vary from person to person, with no measurable differences, is indeed conceivable and could occur without violating scientific laws. Theories of conscious experience that assume or entail

otherwise must therefore be reconsidered. Perhaps more importantly, we now know what to say the next time a child asks if their colours might differ from ours.

Methods

Scrambling Theorem. Let (X,ξ) and (Y, ζ) be measurable spaces, $f: X \rightarrow Y$ a measurable function, p a probability measure on (X,ξ) and d the metric on ξ induced by p defined, for all $A, B \in \xi$, by $d(A,B)=p(A \Delta B)$, where Δ denotes symmetric difference. Then the probability measure q on (Y, ζ) defined, for all $E \in \zeta$, by $q(E)=p(f^{-1}(E))$, induces a metric d' on ζ satisfying, for all $E, F \in \zeta$, $d'(E,F) = d(f^{-1}(E), f^{-1}(F))$.

Proof. $d'(E,F) = q(E \Delta F) = p(f^{-1}(E \Delta F)) = p(f^{-1}(E) \Delta f^{-1}(F)) = d(f^{-1}(E), f^{-1}(F))$. ♦

If f is bimeasurable, then the Scrambling Theorem also holds in the other direction, i.e., for all $G, H \in \xi$, $d(G,H) = d'(f(G), f(H))$. The Theorem also holds if, instead of the PSD metric, one uses the probability of *any* measurable function of the relevant sets. For instance, it holds for the quasi-pseudometric[25] $d^*(A,B)=p(B\backslash A)$, where $p(B\backslash A) + p(A\backslash B) = p(A \Delta B)$; d^*, like some psychological judgments of similarity,[30] is not symmetric. The Theorem assumes that ξ and ζ are σ-algebras, but holds more generally. If, for instance, ξ and ζ are closed under *disjoint* union then the metric, where defined, is preserved under scrambling. This more general formulation permits incomparable pairs of subjective experiences, for which there is no distance: What is the distance between the smell of garlic and the sound of a flute?

Some functionalists object to letting points $x \in X$ represent conscious experiences because, on their theory, it is not x itself but instead the collection of its

distance relations $d_x = \{d(x,z): z \in X\}$ that corresponds to experience. This objection does not obviate the Scrambling Theorem. We let $D = \{d_x : x \in X\}$ and note that the map $g: X \rightarrow D$ given by $x \rightarrow d_x$ is bijective. Any measure p on X can be transported by g to D, and we can then apply the Scrambling Theorem to all measurable functions $f: D \rightarrow D$, because points of D represent conscious experiences according to functionalism. Others deny that points $x \in X$ represent conscious experiences because such points could, instead, be taken to represent internal codes or states of, say, unconscious robots. But this possibility does not preclude applying the Scrambling Theorem to conscious experiences any more than using the integers to count apples precludes using them to count oranges.

The PSD metric used in the Scrambling Theorem generalizes the function A_z, the area under a receiver operating characteristic curve, used in Signal Detection Theory.[18,19] A_z describes an observer's ability to discriminate between two signals with Gaussian distributions G_1 and G_2 on the measurable real line (R,ξ). Using the PSD metric, one obtains this special case by letting $p=(G_1 + G_2)/2$, and choosing two events, A and $B \in \xi$, satisfying: (1) event A is radially symmetric about the mean of G_1, and B about the mean of G_2, and (2) A and B together maximize $p(A \Delta B)$. An example is shown in Figure 2b. This special case of the PSD metric varies monotonically with A_z as the means and variances of the gaussians vary, as shown in Figure 4a,b. If an observation results in an event C, and one must decide between G_1 and G_2 given C, then an unbiased decision criterion is the value 1 for the ratio of the distances from C to G_1 and G_2, viz., the ratio $G_1(R \Delta C)/G_2(R \Delta C)$, as shown in Figure 4c.

The PSD metric is preserved, up to a global scale factor, by the entailment relation of the Lebesgue logic on probability measures, in which p entails q iff p is a

normalized restriction of q.[29] Perceptual inferences of observers are morphisms of the Lebesgue logic,[21,29] and therefore respect the PSD metric.

1. Locke, J. *An Essay Concerning Human Understanding* (1690).

2. Crick, F. & Koch, C. *Cereb. Cort.* **8**, 97–107 (1998).

3. Tye, M. *Consciousness, Color, and Content* (MIT Press, Cambridge, MA, 2000).

4. Chalmers, D. *The Conscious Mind* (Oxford University Press, 1996).

5. Braddon-Mitchel, D. & Jackson, F. *Philosophy of Mind and Cognition* (Blackwell, Oxford, 1996).

6. Gregory, R. *Brit. Med. J.* **317**, 1693–1695 (1998).

7. Metzinger, T. (Ed) *Neural Correlates of Consciousness: Empirical and Conceptual Questions* (MIT Press, Cambridge, MA, 2000).

8. Chalmers, D. (Ed) *Philosophy of Mind: Classical and Contemporary Readings* (Oxford University Press, 2002).

9. Bickle, J. *Philosophy and Neuroscience: A Ruthlessly Reductive Account* (Kluwer Academic Publishers, Dordrecht 2003).

10. Churchland, P.S. *Brain-Wise: Studies in Neurophilosophy* (MIT Press, Cambridge, MA, 2002).

11. Block, N. & Fodor, J. *Philos. Rev.* **81**, 159–181 (1972).

12. Dennett, D. in *Brainchildren* 141– 152 (MIT Press, Cambridge, MA, 1998).

13. Clark, A. *South. J. Philos.* **22**, 431–443 (1983).

14. Penrose, R. *Shadows of the Mind* (Oxford University Press, 1994).

15. Albert, D. *Quantum Mechanics and Experience* (Harvard University Press, 1994).

16. Barrett, J. *The Quantum Mechanics of Minds and Worlds* (Oxford University Press, 1999).

17. Searle, J. *The Rediscovery of the Mind* (MIT Press, Cambridge, MA, 1994).

18. Green, D. & Swets, J. *Signal Detection Theory and Psychophysics* (Peninsula Publishing, Los Altos, CA, 1988).

19. Wickens, T. *Elementary Signal Detection Theory* (Oxford University Press, 2002).

20. Knill, D. & Richards, W. *Perception as Bayesian Inference* (Cambridge University Press, 1996).

21. Bennett, B., Hoffman, D., & Prakash, C. *Observer Mechanics: A Formal Theory of Perception* (Academic Press, San Diego, 1989).

22. Fine, T. *Theories of Probability: An Examination of Foundations* (Academic Press, London, 1973).

23. Ash, R. & Doleans-Dade, C. *Probability and Measure Theory* (Academic Press, San Diego, 2000).

24. Blumenthal, L. & Menger, K. *Studies in Geometry* (Freeman, San Francisco, 1970).

25. Ferrer, J. *App. Gen. Topol.* **4**, 243–253 (2003).

26. Mausfeld, R. & Heyer, D. (Eds) *Colour Perception: Mind and the Physical World* (Oxford University Press, 2003).

27. Chalmers, D. in *Conscious Experience* (ed Metzinger, T.) 309–328 (Imprint Academic, Exeter, UK, 1995).

28. Churchland, P.M. *J. Philos.* **93**, 211–228 (1996).

29. Bennett, B., Hoffman, D., & Prakash, C. *J. Math. Psychol.* **37** 63–103 (1993).

30. Tversky A. *Psych. Rev.* **84** 327–352 (1977).

Acknowledgements We thank Jeff Barrett, Mike Braunstein, Dave Chalmers, Charlie Chubb, Paul Churchland, Stuart Cracraft, Dan Dennett, Jason Ford, Ron Frostig, Richard Gregory, Duncan Luce, Rainer Mausfeld, Louis Narens, Alan Nelson, Barron Ramos, Whitman Richards, Bob Schwartz, Terry Sejnowski and Manish Singh for helpful comments on earlier drafts. We thank Bill Batchelder, Geoff Iverson and Louis Narens for helpful discussions. This work was supported by a grant from the US National Science Foundation.

Competing Interests statement. The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to D.H.. (e-mail: ddhoff@uci.edu).

Figure 1. Violets and marigolds. Could subjective experiences of colour differ from person to person as radically as these two images, with no experimental consequences?

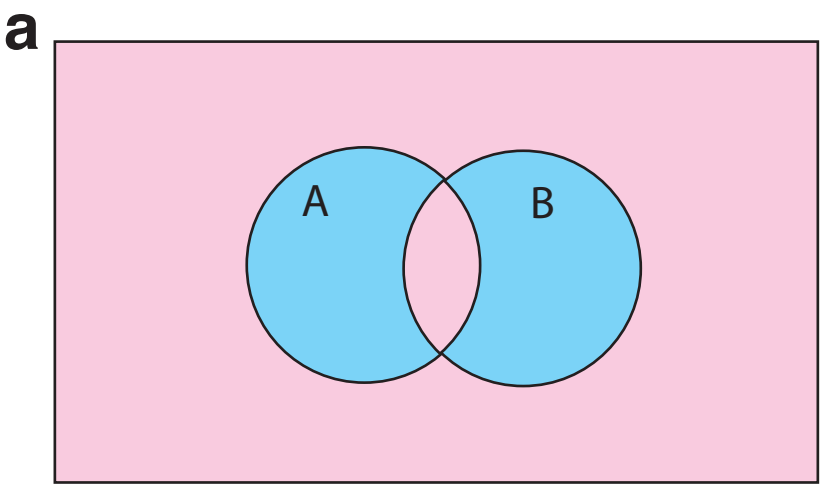
Figure 2. The probability of symmetric difference (PSD) metric. **a,** The symmetric difference, shaded blue, of events *A* and *B*. The symmetric difference of *A* and *B* is their union minus their intersection. **b,** The PSD distance between events *A* and *B*. This distance is the green area above *A* plus the red area above *B*. The probability measure in this example is $(G_1 + G_2)/2$, where G_1 is a Gaussian with mean 0 and standard deviation 1, and G_2 is a Gaussian with mean 4 and standard deviation 2. **c,** The PSD distance between events *A* and *B* that overlap. The region of overlap is not included in the symmetric difference, and so the PSD distance is less than in Figure 2b.

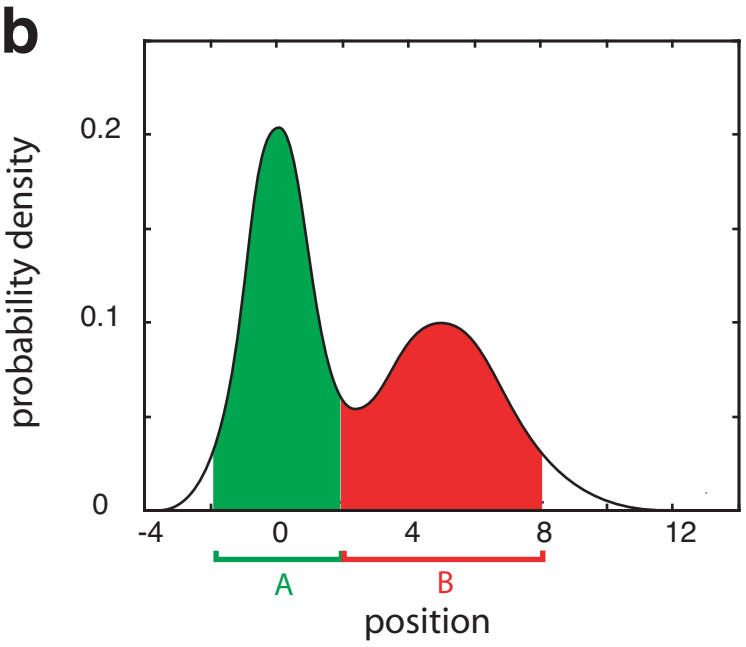
Figure 3. Simple examples of the Scrambling Theorem. **a,** A scrambling of experiences that preserves all distances. The arrows define the scrambling, and transport the probability *p* to its distribution *q*. Since the four basic colour events

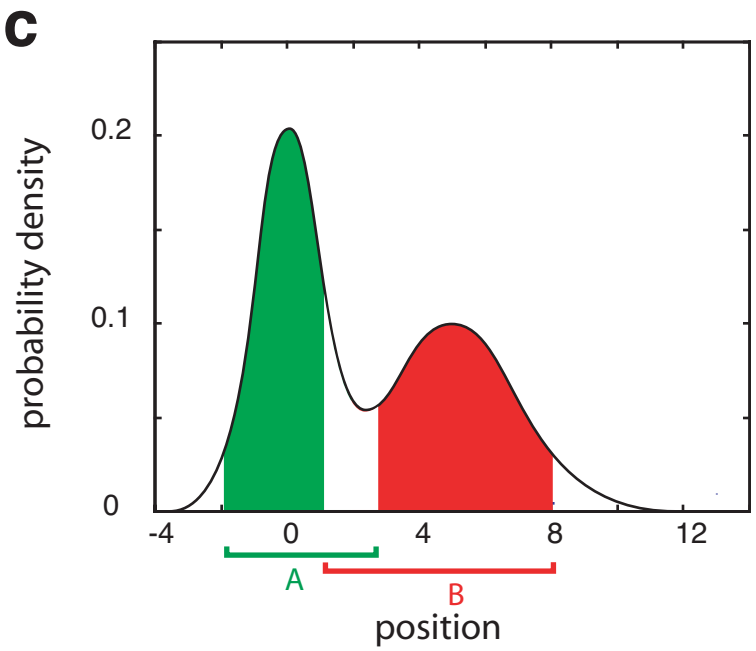
are disjoint, the measure of the symmetric difference, and therefore of the distance, between any two of these colour events is just the sum of their measures. One easily verifies that the distances *d* and *d′* between corresponding colour events remain unchanged. **b,** Objects, subjective experience, and language. Although Jack and Jill differ in their colour experiences, they still use the same colour names to describe objects. An apple leads to one colour experience for Jack, as indicated by its arrow on the left, and to another colour experience for Jill, as indicated by its arrow on the right. But Jack and Jill both call their colour experience "red", as indicated by their arrows to the colour words. Mathematically, one says that the diagram commutes.

Figure 4. Plots of the PSD metric and the function A_z of Signal Detection Theory. **a,** PSD and A_z for two Gaussians with standard deviations 1 and 5, as the difference in their means varies from 0 to 20. PSD and A_z are related monotonically, but PSD has greater sensitivity than A_z where sensitivity is needed most, viz., for small differences between the means. **b,** PSD and A_z for two Gaussians with means 0 and 5 as the difference in their standard deviations varies from 0 to 20. Again, PSD and A_z are related monotonically, but PSD has greater sensitivity than A_z for small differences between the standard deviations. **c,** Ratio of PSD distances. For each position of an event, *C*, of width 0.2 we plot the ratio $G_1(R \Delta C)/G_2(R \Delta C)$., where *R* denotes the real line. G_1 is always a Gaussian with mean 0 and standard deviation 1. G_2 is a Gaussian with mean 4 and standard deviation 1 for the blue curve, standard deviation 2 for the green curve, and standard deviation 4 for the red curve. On the far left of the plot, the red curve lies above the value 1, indicating that in the PSD metric the event *C* in this region is closer to the Gaussian with mean 4 than to the Gaussian with mean 0.

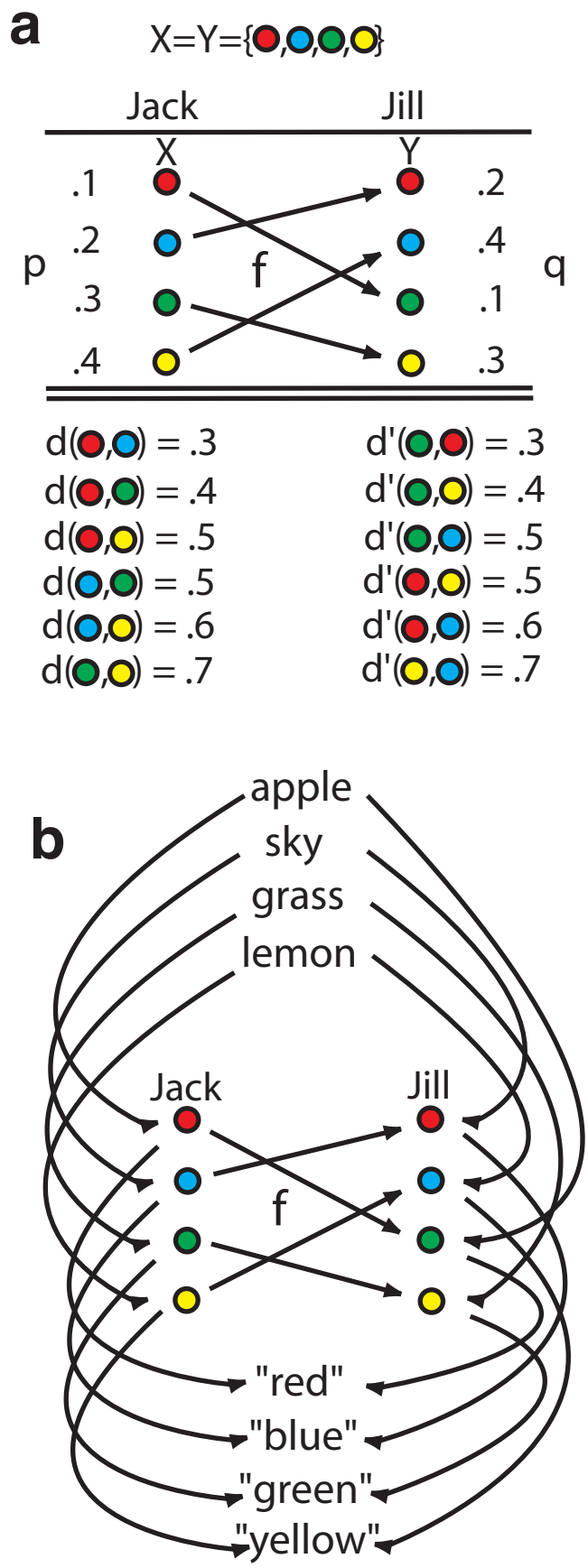
hoffman_fig1



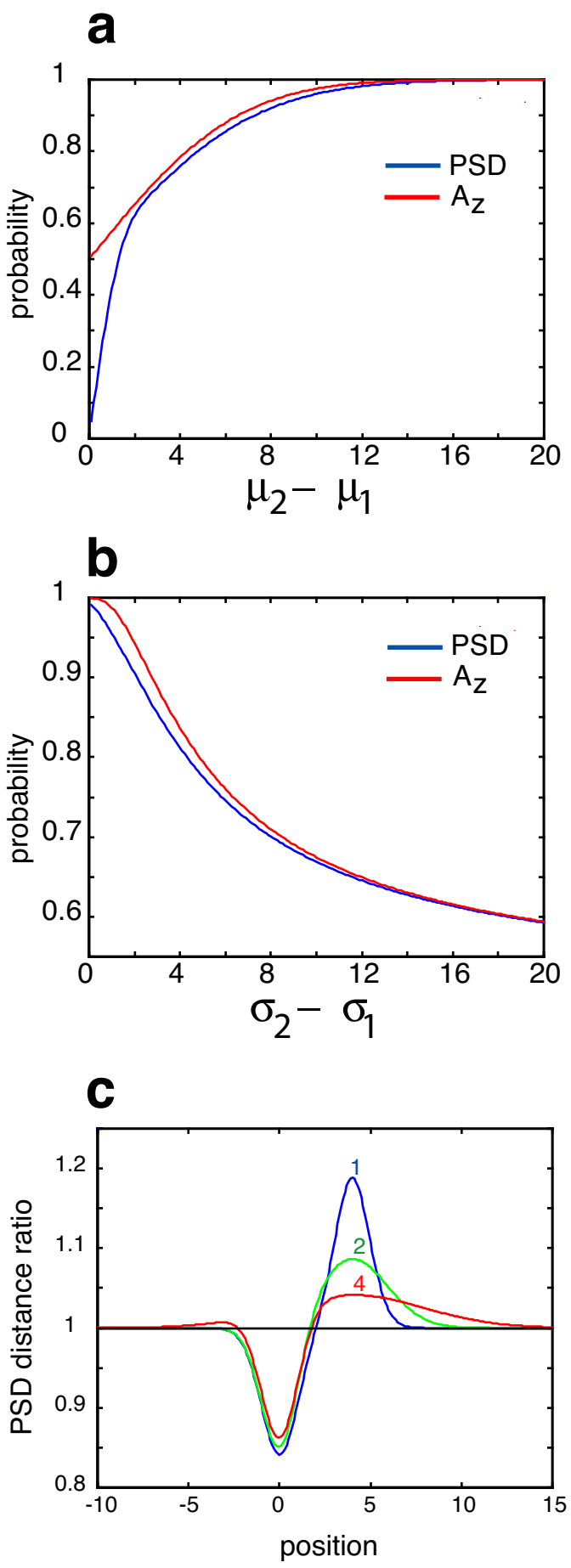




hoffman_fig2



hoffman_fig3



hoffman_fig4