

1 We have addressed the reviewers' comments by running seven new experiments, which shed useful new light on some of
 2 these issues. We have updated the main text and appendices to respond to each comment, reporting the new experiments
 3 and adding further discussion where appropriate.

4 *R2: The authors propose that MCMCP/GSP estimate utility, whereas Sanborn & Griffiths proposed that MCMCP*
 5 *estimates subjective probability; however, it is not clear whether utility and subjective probability are equivalent, and*
 6 *which should be preferred. A:* To investigate whether these possibilities can be differentiated, we reran the MCMCP
 7 and GSP color tasks with three different questions probing different constructs. However, we found no clear effects of
 8 question type, suggesting that the three constructs can be equated here (Fig. A).

9 *R2: How about aggregated MCMCP? We reran the color experiment with aggregated MCMCP and found that*
 10 *aggregated GSP performed worse than both forms of GSP (Fig. B).*

11 *R2: GSP seems intuitively dependent on parametrization, can you discuss? To address this issue we reran the face*
 12 *experiment comparing three parametrization methods (PCA, sparse PCA, ICA), as well as low-variance components*
 13 *of PCA as a control. We found good performance in each case except the low-variance components, suggesting that*
 14 *several data-driven methods can recover sufficiently psychologically meaningful dimensions for GSP (Fig. C).*

15 *R3: Does the benefit of aggregation disappear once you take into account the number of responses required? A:* We find
 16 that non-aggregated GSP performs best for short chains, but aggregated GSP performs best for long chains (Fig. S12).

17 *R3: How do the experimenters avoid subjects merely making the same response 10 times? A:* The across-participants
 18 algorithm ensures that no subject sees the same question twice; multiple chains are run in parallel (Fig. S5), and each
 19 participant only visits the same chain once.

20 *R3: It would be worth discussing how the technique differs from e.g. multidimensional methods of adjustment in*
 21 *psychophysics. A:* Our revised paper explains how our adaptive procedure differentiates GSP from slider paradigms
 22 common in psychophysics, which are tailored to identifying perceptual limits in low-dimensional spaces.

23 *R5: The theory section discusses the trade-off between mode seeking and stochastic sampling, but this trade-off is*
 24 *neglected when discussing Study 1. A:* We conducted a new experiment to derive a ground truth for the utility function,
 25 and designed a new analysis to examine the trade-off between mode-seeking and stochastic sampling. We confirm that
 26 GSP is more mode-seeking than MCMCP, but nonetheless recovers the utility function more reliably (Fig. D).

27 *R5: The authors changed the MCMCP trial question slightly from Sanborn & Griffiths (2008), I'm curious to see*
 28 *whether this had any effect on behavior. We reran MCMCP and GSP on the color task, using three different questions*
 29 *including one closely resembling the Sanborn & Griffiths question. There was no systematic difference in outcomes*
 30 *here, supporting the notion that all these questions probe a common utility function (Fig. A).*

31 *R5: The authors do not sufficiently acknowledge how biases in the GAN's training data affect the samples generated*
 32 *by the GSP process. To draw conclusions from the results, I would want to see additional results using a different*
 33 *training dataset. Our revised manuscript acknowledges this by reframing GSP as a tool for navigating and interpreting*
 34 *the parameter space of generative models using participant judgements. To illustrate this, we conduct two further*
 35 *experiments, one manipulating the dataset (portraits vs. photographs, Fig. E) and one manipulating the participant group*
 36 *(male vs. female, Fig. F).*

