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

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

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

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

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

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

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

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

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

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