Dear Mr. Goodhill,

Attached you will find our revised version of the manuscript entitled, "Are mental properties supervenient on brain properties?" Please excuse the delay in our revisions, as you will see, the revisions have been somewhat substantial due to the very helpful suggestion of the two reviewers. Below, we respond to each specific comment (our responses are in bold).

Reviewer #1 (Technical Comments):

Superficially: I find the gesture towards history in the introduction a bit amateurish. Plato, Descartes, and Davidson are used to represent the whole of the philosophy of mind. The introduction as it stands suggests implicitly a kind of unfamiliarity with the literature which is probably best not to flag. (No philosophy of mind pre-Plato? Nothing between Plato and Descartes? Or Descartes and Davidson? And why Davidson rather than Strawson or Armstrong or Kim or Ryle? So to avoid raising those questions, one might chose to introduce the paper differently. It might be better to start off thinking about what supervenience is and why it is an important idea.

Thank you for this suggestion. We have tried to reformulate the introduction with these comments under consideration.

So establishing supevenience on empirical grounds would not settle the dualism question with which we started out. And so a formal way of specifying the supervenience hypothesis does not "render the mind-brain dualism debate a hypothesis." The debate will continue even if minds supervene on brains.

Indeed, we have removed claims of resolving the mind-brain dualism debate.

The authors might acknowledge that their view of supervenience is of the local rather than global variety.

We now explicitly note that we are interested in a local version supervenient, in the following two places:

Introduction: "For example, neural network theory and artificial intelligence often implicitly assume a local version mind-brain supervenience \cite{Haykin2008,Ripley2008}."

Statistical supervenient: a definition: "To facilitate both statistical analysis and empirical investigation, we convert this local supervenience relation from a logical to a probabilistic relation. "

Small point. On the bottom of p. 2, the thesis is stated that two different minds must supervene on two different brains.

Yes, we changed that sentence to read:
"If minds statistically supervene on brains, then if two minds differ, there must be some brain-based difference to account for the mental difference."

I am not clear how brain types are to be specified in this apparatus, or mental types, for that matter.

We have tried to modify the text to correct this confusion, as follows:

Statistical supervenient: a definition: "Let M={m_1, m_2,…} be the space of all possible minds and let B={b_1,b_2,…} be the set of all possible brains. M includes a mind for each possible collection of thoughts, memories, beliefs, etc.B includes a brain for each possible position and momentum of all subatomic particles within the skull. "

The existence and construction of a consistent statistical test for epsilon-supervenience: "We therefore must restrict our interest to a mind/brain *property* pair. A mind (mental) property might be a person's intelligence, psychological state, current thought, gender identity, etc. A brain property might be the number of cells in a person's brain at some time t, or the collection of spike trains of all neurons in the brain during some time period t to t'."

Small point: I don't know what it means to "transcend the evolutionary hierarchy" (p. 4).

At the end of the paper, I feel the need to emphasize again that the proposed method does not serve as a test for the existence of souls. I wish it did, but it doesn't. As described above, souls might supervene on brains. It is also certainly not a way of testing whether souls have causal roles over and above the physical. If M and P are the same whenever effect E happens, how does one tell whether it was M that did the work, P that did the work, or both? Well, again we seem to be in the territory of metaphysics rather than experimentation. If supervenience holds, we can never experimentally intervene to change M without ipso facto (by definition) changing P. They are experimentally indistinguishable. And so we will not be designing experiments to sort these metaphysical possibilities.

We have substantially modified the Discussion based on these points to remove extraneous (and perhaps inaccurate) claims.

Medium point: I'm not sure how much work is done by the assumption that the space of all possible minds is finite or of how it would change if the space of possible minds is infinite.

The results all hold when the finite assumption is relaxed. We have tried to highlight that more clearly in the current version of the text:

Gedankenexperiment 1: "Furthermore, the proof holds for other matrix norms (which might speed up convergence and hence reduce the required n), and the regression scenario where I\mMI is infinite (again, see Methods for details)."

Methods: "Stone's original proof also applied to the scenario when IMI was infinite, resulting in a universally consistent regression algorithm as well."

I hope that these comments prove useful to the authors in thinking about how to pitch their interesting theoretical work.

Very!

Reviewer #2 (Remarks to the Author):

On the whole, the article is very clear, and I enjoyed reading it.

Thank you!

- The central terms "mental property" and "physical property" were not, as far as I could tell, given a clear definition within the text.

We have clarified the definitions of these properties in two sections: "Statistical supervenient: a definition" and "The existence and construction of a consistent statistical test for epsilon-supervenience." Please see above for details.

- The main results, the methods and supplementary information are currently slightly disorganised

Thank you, we have re-organized to reflect these comments. Each subsection now builds on the previous one to establish a new claim. The first paragraph of the discussion highlights this organization:

"This work makes the following contributions. First, we define statistical supervenience based on Davidson's canonical statement (Definition 1). This definition makes it apparent that supervenience implies the possibility of perfect classification (Theorem 1).
We then prove that there is no viable test against supervenience, so one can *never* reject a null hypothesis in favor of supervenience, regardless of the amount of data (Theorem 2). This motivates the introduction of a relaxed notion called epsilon-supervenience (Definition 2), against which consistent statistical tests are readily available. Under a very general brain-graph/mental property model (*Gedankenexperiment* 1), a consistent statistical test against epsilon-supervenience is always available no matter the true distribution F_{MB} (Theorem 3). In other words, the proposed test is guaranteed to reject the null whenever the null is false, given sufficient data, for any possible distribution governing mental property/brain property pairs. "

- The discussion in the supplementary information that 'supervenience does not imply identity' is somewhat confusing.

We have completely re-written the non-identity claim, starting with the non-causality claim:

"Second, supervenience does not imply causality. For instance, consider an analogy where M and B correspond to two coins being flipped, each possibly landing on heads or tails. Further assume that every time one lands on heads so does the other, and every time one lands on tails, so do the other. This implies that M supervenes on B, but assumes nothing about whether M causes B, or B causes M, or some exogenous force causes both.

Third, supervenience does not imply identity. The above example with the two coins demonstrates this, as the two coins are not the same thing, even if one has perfect information about the other. "

- Again, in the supp. info: correlation does not seem well-defined in this context

We have tried to clarify the definition of correlation and the distinction between supervenient and correlation:

"Note that statistical supervenience is distinct from statistical correlation. *Statistical correlation* between brain states and mental states is defined as rho_{MB}=E[(B-mu_B)(M-mu_M)]/(sigma_B sigma_M), where mu_X and sigma_X are the mean and variance of X, and E[X] is the expected value of X. If rho_{MB}=1, then both MsB and BsM. Thus, perfect correlation implies supervenience, but supervenience does not imply correlation. In fact, supervenience may be thought of as a generalization of correlation which incorporates directionality, can be applied to arbitrary valued random variables (such as mental or brain properties), and can depend on any moment of a distribution (not just the first two)."

Minor comments:
- Ref 28 should be "Nielsen" rather than "Nielson"
- last sentence of the 3rd paragraph of the discussion, "neurcognitive" should be "neurocognitive".
- "Stone" is mentioned several times in the supplementary information, but no reference is given.

All fixed.

Thank you again for all your hard work. We firmly believe that our manuscript is vastly improved due to your comments.