This is the author version of the following article: Baltimore, Joseph A. (2010). "Defending the piggyback principle against Shapiro and Sober's empirical approach." *Synthese*, 175 (2), 151-168. The final publication is available at www.springerlink.com: http://dx.doi.org/10.1007/s11229-009-9467-2

Defending the Piggyback Principle against Shapiro and Sober's Empirical Approach

Introduction

According to Jaegwon Kim's "supervenience argument" (2005, 13-22, 32-45), mental causation is inconsistent with the following set of claims:

> *Mind-body supervenience*: Mental properties supervene on physical properties, in that necessarily, for any mental property, M, if anything has M at time t, there exists a physical base (or subvenient) property, P, such that it has P at t, and necessarily anything that has P at a time has M at that time.[1]
>
> *Causal closure of the physical domain*: If a physical event has a cause that occurs at t, then it has a physical cause that occurs at t.
>
> *Mental/physical property dualism*: Mental properties are not reducible to, and are not identical with, physical properties.
>
> *Causal exclusion principle*: No single event can have more than one sufficient cause occurring at any given time (unless it is a genuine case of causal overdetermination).

Focusing on mental-to-physical causation, I take the following to be a fair outline of Kim's supervenience argument (or "exclusion argument," as it is sometimes called):

[1] The version of supervenience here is known as "strong supervenience" (Kim 1998, 9; 2005, 33). And, according to Kim, the necessity of this mind-body supervenience "is standardly taken to be at least *nomological necessity*—so that if mind-body supervenience holds, it holds in all worlds that share with our world the same fundamental laws of nature" (Kim 1998, 39, his italics).

(1) Assume for *reductio*: An instance of a mental property, M, causes an instance of a physical property, P*.[2]

(2) By mind-body supervenience, M has a physical supervenience base, P.

(3) If M causes P* and M has physical supervenience base P, then P causes P*.

(4) Hence, P causes P*.

(5) By mental/physical property dualism, M is distinct from P.

(6) But P* is not causally overdetermined by two distinct causes, M and P.

(7) By the causal exclusion principle, either M or P must not be a cause of P*.

(8) By causal closure of the physical domain, it cannot be the case that M causes P* and P does not cause P*.

(9) Hence, M must not cause P*, i.e. (1) must be rejected.[3]

Although there are multiple points of interest in this argument, I want to focus on (6). Premise (6) sets up the application of the causal exclusion principle in (7) and, therefore, it is important to be clear on Kim's motivations for (6). Often, Kim is taken to be motivated by the concern that we must limit the amount of genuine causes a single event can have occurring at the

[2] Properties, as abstract objects, are not causally efficacious. Rather, it is objects instantiating properties, or property instances, that enter into causal relations. However, I will follow Kim here in often sacrificing explicit identification of property instances as such for the sake of simplicity. Nonetheless, however, talk of a property X causing another property Y is to be taken to mean that an instantiation, or instance, of X causes an instantiation, or instance, of Y. For simplicity as well, temporal references have been suppressed. Furthermore, I will follow Kim (2005, 42, footnote 9) in being concerned with an instance of *X* causing an instance of Y *in virtue of the former being an instance of X and the latter being an instance of Y*. The reason for this is that securing robust mental causation seems to require mental *qua* mental causation, or mental causation *as such*. Kim (1989) and others have, for example, levied such a requirement against Donald Davidson's (1970) token physicalism in order to argue that Davidson's account fails to secure a sufficiently robust mental causation. For Davidson's response to such charges as well as replies to his defense, see Davidson (1995), Kim (1995), McLaughlin (1995), and Sosa (1995).

[3] Kim (2005, 19-20, 39-41) does argue as well that, due to mind-body supervenience, mental-to-mental causation entails mental-to-physical causation. The claim there is that in order for an instance of a mental property, M, to cause an instance of another mental property, M*, the former must do so *by causing* a physical supervenience base of the latter. This additional stage of Kim's supervenience argument is indeed important, for it would combine with Kim's argument for mental-to-physical epiphenomenalism to equally yield mental-to-mental epiphenomenalism. However, for the purposes of this paper, focusing on Kim's argument for mental-to-physical epiphenomenalism will be sufficient.

same time, so as to avoid widespread causal overdetermination. That is, rejecting (6) in order to save mental-to-physical causation would result in causal overdetermination being too widespread, for then all of our everyday cases of mental-to-physical causation would be cases of causal overdetermination.[4]

However, Kim has offered the following in clarifying his motivations behind (6):

> To be a cause of P*, M must somehow ride piggyback on physical causal chains—distinct ones depending on which physical property subserves M on a given occasion . . . . And we may ask: In virtue of what relation it bears to physical property P does M earn its entitlement to a free ride on the causal chain from P to P* and to claim this casual chain to be its own? Obviously, the only significant relation M bears to P is supervenience. But why should supervenience confer this right on M? The fact of the matter is that there is only one causal process here, from P to P*, and M's supposed causal contribution to the production of P* is totally mysterious. In standard cases of overdetermination, like two bullets hitting the victim's heart at the same time, the short circuit and the overturned lantern causing a house fire, and so on, each overdetermining cause plays a distinct and distinctive causal role. The usual notion of overdetermination involves two or more separate and independent causal chains intersecting at a common effect. Because of *Supervenience*, however, that is not the kind of situation we have here. In this sense, this is not a case of genuine causal overdetermination, and *Exclusion* applies in a straightforward way. (Kim 2005, 47-8, his italics)

Here, Kim suggests that M and P cannot causally overdetermine P* because M supervenes on P in such a way that they cannot both be genuine causes of P*. In which case, Kim's motivation behind (6) is not to limit the amount of *genuine* causes one can posit but, rather, to ensure that *merely apparent* causes are not mistakenly counted as *genuine* causes.

But how, exactly, does M supervene on P in such a way as to loose any legitimate claim to causing P*? In the above quotation, Kim suggests that the answer is in M failing to have a causal role with respect to P* that is distinct from the causal role that P has with respect to P*. That is, Kim seems to require that M, in order to be a genuine cause of P*, make a casual

[4] Such interpretations do, at times, seem encouraged by Kim's defense of (6), especially when he defends (6) by arguing that its denial would result in violating causal closure of the physical domain. See Kim (1989, 44; 1998, 44-5). There, the threat of violating causal closure of the physical domain apparently requires taking M to be, in addition to P, a *genuine* cause of P*.

contribution to P* *additional* to the causal contribution made by P. Without such further contribution by M, the concern goes, P will be doing all the work in bringing about P*, while M merely rides piggyback on P's causal contribution. Thus, Kim's supervenience argument appears to rely on the following principle:

> *Piggyback principle*: If, with respect to an effect, E, an instance of a property, A, has no causal powers over and above, or in addition to, those had by its supervenience base, B, then the instance of A does not cause E (unless A is identical with B).[5]

In their "Epiphenomenalism: The Dos and the Don'ts," Larry Shapiro and Elliott Sober (2007) also detect such a principle underlying Kim's reasoning about epiphenomenalism. However, Shapiro and Sober (S&S) employ a novel empirical approach in order to reject the likes of the piggyback principle and, equally, any argument for epiphenomenalism like Kim's that similarly relies on it. Their empirical approach pulls from the empirical experiments of August Weismann regarding the inheritance of acquired characteristics. Through a detailed examination of Weismann's experiments, S&S extract lessons, i.e. the dos and the don'ts, in reasoning about the epiphenomenalism of a property. And according to these empirically drawn lessons, the piggyback principle is a don't. My primary aim in this paper is to defend the piggyback principle against S&S's empirical approach.[6]

---

[5] This parenthetical qualification is important, for if properties A and B are one and the same property, i.e. type identical, then the causal powers of B-instances would equally be the causal powers of A-instances, eliminating any concern about an A-instance piggybacking on, or otherwise failing to have legitimate claim to, the causal powers of a B-instance. Mere token identity between A and B, on the other hand, does not so easily escape the concerns of the piggyback premise, for there is the concern that an A-instance may be causally efficacious only in virtue of being a B-instance, or *qua* B-instance. And, as noted earlier (see note 2), the A-instance must be causally efficacious in virtue of being an A-instance, or *qua* A-instance, in order to secure a genuine causal role for A. That is, the piggyback principle is to be understood as providing a necessary condition for an A-instance causing E *in virtue of being an A-instance, or qua A-instance*.

[6] While it would be preferable to go beyond such a defense and also develop arguments securing the truth of the piggyback principle, that is beyond the scope of this paper. However, in defending the piggyback principle, some points do surface that could be used to promote the piggyback principle (e.g. pages 20-1, where it is shown how S&S's own account of epiphenomenalism appears to naturally extend to include the piggyback principle).

S&S's empirical approach to charging that the piggyback principle is a don't

S&S employ an empirical approach for determining the dos and the don'ts in reasoning about the epiphenomenalism of a property. In particular, they derive their model for reasoning about the causal efficacy of a property from Weismann's experiments regarding the inheritance of acquired characteristics:

> In one of his famous experiments, Weismann cut off the tails of newborn mice; when the mice grew up and reproduced, their offspring had tails as long as their parents had had prior to surgery. These results remained constant over many generations. . . . [Weismann's experiments] clearly provided evidence that acquired taillessness in mice parents failed to cause taillessness in mice offspring. (S&S 2007, 235)

Employing the language of genotypes and phenotypes, S&S characterize Weismann as testing for the causal efficacy of parental phenotypes with respect to offspring genotypes. And Weismann's experiments are taken to have provided evidence that parental phenotypes are not causally efficacious with respect to offspring genotypes. That is, Weismann's experiments with mice are understood as having yielded evidence that the parental phenotype of tail length does not causally influence the offspring's genotype responsible for tail length.

Now, of primary importance to S&S is what Weismann held fixed versus what Weismann did not hold fixed in his manipulation of the parental phenotype:

> We have gone into some detail about the logic of Weismann's experiment because we think it is important for philosophers to see clearly what Weismann did *not* do. As we have explained, Weismann manipulated the parental phenotype while holding fixed the parental genotype. He did not manipulate the parental phenotype while holding fixed the microsupervenience base of that phenotype. . . . The most important lesson we draw from Weismann is that investigating whether *X* causes *Y* involves figuring out whether wiggling *X* while holding fixed whatever common causes there may be of *X* and *Y* will be associated with a change in *Y*. It is not relevant, or even coherent, to ask what will happen if one wiggles *X* while holding fixed the microsupervenience base of *X*. (S&S 2007, 239-40, their italics)

As S&S emphasize here, Weismann held fixed the parental genotype, which was the common cause of the parental phenotype and the offspring genotype. But Weismann did not hold fixed the

supervenience base of the parental phenotype. With this highlighting of what Weismann did and did not hold fixed in his manipulation of the parental phenotype, S&S generate the following distinction between the right and wrong way of testing whether X causally influences Y:

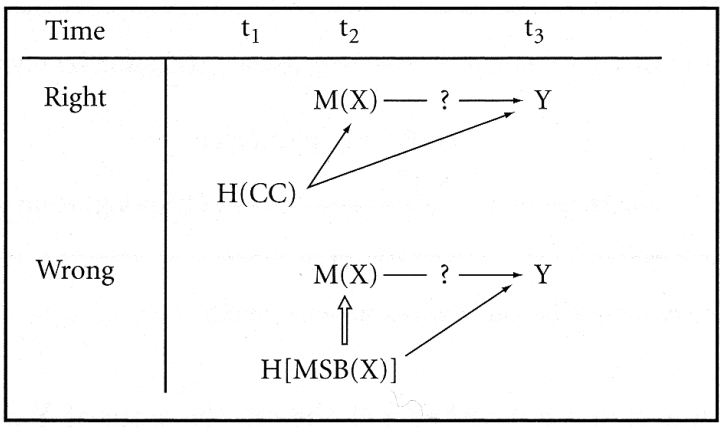

Fig. 13.5. The right and the wrong manipulations for determining whether *X* causally influences *Y*. M(*X*) means that *X* is manipulated by an intervention, H(*Z*) means that *Z* is held fixed, and MSB(*X*) is the microsupervenience base of *X*. (S&S 2007, 240)

S&S then elaborate what, precisely, is wrong with the wrong way of testing whether X causally influences Y:

> Because a supervenience base for *X* provides a sufficient condition for *X*, where the entailment has at least the force of nomological necessity, asking this question leads one to attempt to ponder the imponderable – would *Y* occur if a sufficient condition for *X* occurred but *X* did not? (S&S 2007, 240)

Thus, according to S&S, wiggling X while holding fixed its supervenience base is not a legitimate way of testing whether X causally influences Y, for such a manipulation is ruled out by supervenience. In particular, the *nomological* necessity of supervenience prevents one from successfully employing such a test and, therefore, such a test cannot be the right way of testing the causal efficacy of X with respect to Y.

With this lesson in testing the causal efficacy of a property, S&S turn to the problem of mental causation:

> How could believing or wanting or feeling cause behavior? Given that any instance of a mental property *X* has a physical microsupervenience base MSB(*X*), it would appear that *X* has no causal powers in *addition* to those that MSB(*X*) already possesses. The absence of these additional causal powers is then taken to show that the mental property *X* is causally inert. We call this the master argument for epiphenomenalism. (S&S 2007, 241, their italics)

Such reasoning, according to S&S, commits one to the error depicted in figure 13.5, which involves the nomologically impossible method of wiggling a supervenient property while holding fixed its supervenience base:

> The crucial mistake in this line of reasoning is that it requires one to consider a counterfactual situation that is in fact impossible and is, in any case, irrelevant to the question of whether the mental property *X*, or any other supervening property, is epiphenomenal with respect to an effect term *Y*. To see if *X* has an effect on *Y* that is additional to whatever effect MSB(*X*) has on *Y*, one would have to compare what would happen to *Y* if both MSB(*X*) and *X* were present with what would happen to *Y* if MSB(*X*) were present and *X* were absent. The master argument purports to evaluate this counterfactual and then concludes that the mental property *X* makes no contribution to *E* additional to the effect that MSB(*X*) has. The conclusion is then drawn that mental properties are causally inert. The principal fallacy is the thought that if *X* causes *Y*, then *X* must have an impact on *Y* additional to the impact on *Y* that MSB(*X*) has. (S&S 2007, 241)

S&S claim here that the primary mistake of the master argument is its requirement that, in order to cause Y, X must make a causal contribution to Y additional to the causal contribution made by MSB(X). That is, S&S reject the piggyback principle. And the piggyback principle is said to be wrong because its application depends on the wrong way of testing, for one must wiggle X while holding fixed MSB(X) in order to see if X has an impact on Y additional to the impact on Y that MSB(X) has.

Now, S&S explicitly employ counterfactual evaluation in their above criticism of the piggyback principle. But since S&S began with empirical tests that concern what *does* happen when certain interventions *are* made, one might wonder why S&S move to what *would* happen if

certain interventions *were* made. That is, how, precisely, does counterfactual analysis link the piggyback principle to the wrong way of testing?

The answer seems to be this. The nomological necessity of supervenience prevents one from *empirically* performing the wrong way of testing. Consequently, the wrong way of testing can, at best, be employed through *counterfactual* evaluation that mirrors the wiggling method of the wrong way of testing. Now, such counterfactual evaluation can be a legitimate alternative, for counterfactual truths are often a good indication of casual truths. One need not, for instance, do the empirical work of eating five hamburgers in order to know that doing so causes a belly ache. It is sufficient to determine that if you *were* to eat five hamburgers, then you *would* get a belly ache. However, in the case of employing the wrong wiggle test in its counterfactual form, such counterfactual evaluation can never result in a substantive conclusion, for the nomological necessity of supervenience ensures that asking that counterfactual question only "leads one to attempt to ponder the imponderable." Therefore, the wrong way of testing is wrong because it is incapable of yielding any substantive conclusion, whether that method of testing is used in its empirical or counterfactual form. But application of the piggyback principle requires using the wrong way of testing (in either its empirical or counterfactual form) and, hence, the piggyback principle is deeply mistaken—a don't in reasoning about the epiphenomenalism of mental properties.

With this challenge to the piggyback principle presented, there are some minor terminological issues to smooth out before moving on to evaluate S&S's argument. First, for ease of discussion, it is common practice to move from talk of a "microsupervenience base property" to simply "base property." I will follow that practice here and use "B(X)" in the place of "MSB(X)," where the former has the same meaning that S&S intend for the latter. Second, it

would also help to have more neutral ways of referring to the methods of testing in Figure 13.5. So, instead of "right manipulation (intervention, wiggle, etc.)" I will use "H(CC)-manipulation (-intervention, -wiggle, etc.) and, similarly, instead of "wrong manipulation (intervention, wiggle, etc.)" I will use "H[B(X)]-manipulation (-intervention, -wiggle, etc.)." With our terminology somewhat more streamlined and less bias, we are now ready to evaluate S&S's argument.

An alternative testing method: causal isolation

According to S&S, application of the piggyback principle requires using the H[B(X)]-wiggle test. And since the empirical form of such a test is ruled out by the nomological necessity of supervenience, one's next best option is analyzing the following sort of counterfactual:

CF: If there were an H[B(X)]-wiggling of X, then there would be a change in Y.

But, S&S argue, CF cannot be evaluated because the nomological necessity of supervenience makes the antecedent of CF impossible (one cannot have B(X) present without also having X present).

Consider, however, the following counterfactual:

CF*: If there were a change with respect to X being causally isolated from Y, while B(X) remained free of such causal isolation, then there would be a change in Y.

The antecedent of CF* does not require any wiggling of X; both B(X) and X can remain present. Instead, X is causally isolated from Y, i.e. environmental conditions occur that prevent any causal powers X might have with respect to Y from reaching Y. Thus, the antecedent of CF* is not ruled out by B(X) being nomologically sufficient for X.

Furthermore, such causal isolation seems like a perfectly fine alternative to the H[B(X)]-wiggle test in determining whether X has an impact on Y additional to that had by B(X). Take, for example, an empirical investigation into the causal powers of a certain colored paint with

respect to a person's mood. Exposure to this paint results in feelings of anxiety. And this paint, in order to get its specific color, has to be mixed in such a way that the paint has a noticeable odor. Suppose further that we already know that the paint's color causes anxiety. Now, the question of our current investigation is whether the odor has any additional causal contribution to the feeling of anxiety. And to answer this question, it seems entirely appropriate to causally isolate the paint's odor property by, say, placing a transparent, glass barrier between the paint and the test subjects, blocking any causal influence that the odor might have on the subjects. Should it turn out that the subjects feel anxiety in this case just as they did when the odor was not causally isolated, then we will have good evidence that the odor does not make a causal contribution to feelings of anxiety additional to the causal contribution made by the color. While, on the other hand, if it turns out that the subjects feel less anxiety than they did without the transparent barrier in place, then we will have good evidence that the odor does make a causal contribution to feelings of anxiety additional to the causal contribution made by the color.

Also, notice that it does not matter if one could conduct an alternative experiment where the paint has the color but not the odor. Even if such a wiggling of the odor while holding fixed the color were available, causal isolation would remain an appropriate means of testing whether the odor makes a causal contribution to feelings of anxiety additional to the causal contribution made by the color. The point of the example is not that causal isolation is *necessary* but, rather, that it is *sufficient* for testing whether the odor makes a causal contribution to feelings of anxiety additional to the causal contribution made by the color.[7]

[7] It is worth noting that, as a general form of testing, causal isolation appears to be present in accepted, scientific practice. In biochemistry, for example, it is common practice to employ an enzyme inhibitor in order to test for whether a particular enzyme activates a certain drug. While there are a variety of inhibitors, one option is using a competitive inhibitor, which binds to the active site of the enzyme and can thereby block causal contributions the enzyme might make with respect to activation of the drug. If the presence of the enzyme and its competitive inhibitor is associated with a lesser degree of drug activation than the presence of the enzyme without the inhibitor, then there is evidence that the enzyme contributes to the activation of the drug. In this sort of experiment, where the

There is, then, an acceptable alternative to the H[B(X)]-wiggle test (in either its empirical form or by evaluating CF) that one can appeal to in applying the piggyback principle, namely, causal isolation (in either its empirical form or by evaluating CF*). And this alternative method of testing is not vulnerable to the alleged problems of the H[B(X)]-wiggle method, for it does not require that B(X) occur without X occurring.

Now, in order to accommodate this point, S&S can reply the following way. There is an important difference between, on the one hand, the causal pathways of color and odor and, on the other hand, the causal pathways of X and B(X). In the case of color and odor, there are two, distinct causal pathways. And these distinct causal pathways are what enable one to causally isolate the paint's odor without thereby also causally isolating the paint's color. In the case of X and B(X), however, there are not two, distinct causal pathways. The supervenience relation between X and B(X) involves a robust dependency relation, so that X is instantiated *in virtue of* B(X) being instantiated. And this significant anchoring of X in B(X) ensures that, for whatever causal contribution X might make with respect to Y, the causal contribution will employ the same causal pathway as whatever causal contribution B(X) might make with respect to Y. On functionalism, for instance, X will be a functional property, i.e. the property of having some property that performs a certain causal role. And X is instantiated in virtue of B(X) because B(X) plays the functional role associated with X and thereby functionally realizes X. But it seems clear that this grounding of X in B(X) prevents one from causally isolating X without thereby also causally isolating B(X); one cannot block any causal contribution a functional property might make without thereby also interfering with the causal contribution of its realizer. Thus, due to the robust anchoring of supervenient properties in their base properties, the antecedent of CF* is just

---

introduction of the competitive inhibitor does not destroy the enzyme, the competitive inhibitor is apparently used to causally isolate the enzyme in order to determine whether it activates the drug.

as impossible as the antecedent of CF and, therefore, causal isolation is, in the case of X and B(X), no more plausible a method of testing than the H[B(X)]-wiggle method.

Notice, though, that S&S now require more from supervenience than originally suggested. Originally, S&S suggested that the nomological sufficiency of B(X) for X is enough to generate the inability to test for whether X has causal powers over and above those had by B(X). However, as we have just seen, in addition to B(X) being nomologically sufficient for X, it must also be the case that X is anchored in B(X) to an extent that causal isolation of X nomologically necessitates causal isolation of B(X).

Now, in the context of Kim's supervenience argument, this might not be a problem for S&S. Kim (2007, 34) takes mind-body supervenience to entail more than merely covariation between mental and physical properties. According to Kim, mind-body supervenience also involves an "existential" dependence of the mental on the physical, where that sense of dependence "justifies saying that a mental property is instantiated in a given organism at a time because, or in virtue of the fact that, one of its physical 'base' properties is instantiated by the organism at that time" (Kim 2007, 34). But, as functionalism suggests, physicalist accounts of this "existential" dependency relation may very well anchor mental properties in their physical bases to the extent that causal isolation of the former nomologically necessitates causal isolation of the latter. Thus, in the context of Kim's supervenience argument, S&S might not have too much trouble accommodating the alternative testing method of causal isolation.[8]

[8] I say "might" here because although non-reductive physicalists do heavily favor the functionalist account of the existential dependency relation, it is only one account. Consequently, there remains room to investigate more fully whether or not there have been, or could be, physicalist accounts of the existential dependency relation that anchor the mental in the physical to a lesser extent, so that it is nomologically possible to causally isolate a mental property without similarly restricting its physical base. But such an investigation would require significant analysis of both physicalism and existential dependency relations, for which the present paper, unfortunately, does not have space.

Nonetheless, however, this section has revealed the following about S&S's argument. First, S&S's empirical approach employs too narrow a view on how one might test for whether X has causal powers over and above those had by B(X). In addition to the H[B(X)]-wiggle method, there is also the causal isolation method. Second, this alternative method forces S&S to assume a stronger account of supervenience than what they initially appear to require for their argument. In addition to B(X) being nomologically sufficient for X, it must also be the case that X is anchored in B(X) to an extent that causal isolation of X nomologically necessitates causal isolation of B(X). Now, although these points provide insight into key assumptions of S&S's argument, we have not yet undermined their argument, especially in the context of Kim's supervenience argument. Time, then, to turn to more penetrating concerns with S&S's argument.

Are empirically based tests even required in applying the piggyback principle?

As we saw, S&S argue that the piggyback principle is wrong because its application requires the H[B(X)]-wiggle test, in either its empirical or counterfactual form. And we can see this line of argument again in S&S's following criticism of Kim:

> In asking whether *M* should be given a distinct causal role in the production of *P**, Kim is inviting us to consider whether *M* has an effect on *P** additional to the effect that *P* has. The obvious method to use in answering this question involves holding *P* fixed while wiggling *M*. But this, of course, is just to commit the error that figure 13.5 depicts. One cannot manipulate a macroproperty while holding fixed the microproperties on which it supervenes. (S&S 2007, 244-5, original variables altered to match those in this paper's introduction)

Here, S&S charge that determining whether a supervenient mental property, M, has a causal role distinct from that of its physical base, P, leads one to the H[B(X)]-wiggle test, which S&S claim is the wrong sort of test. And, consequently, Kim's reasoning about whether M is epiphenomenal with respect to P is said to go wrong insofar as Kim applies the piggyback principle to M, for doing so involves engaging in the wrong method of testing.

To an extent, we have already challenged this claim that application of the piggyback principle requires the H[B(X)]-wiggle test, for we identified an alternative test that one could employ in applying the piggyback principle, i.e. causal isolation. But we saw how S&S might adjust to accommodate such an alternative test, especially in the context of Kim's supervenience argument. Also, both the H[B(X)]-wiggle test and causal isolation test are empirically based tests in the following sense: They are clearly empirical/scientific methods for testing, with their counterfactual forms being justified insofar as they reflect their corresponding empirical/scientific methods. And it is now time to question whether any such empirically based test is required in applying the piggyback principle, thereby challenging the heart of S&S's basic assumption that application of the piggyback principle requires the H[B(X)]-wiggle test.

When evaluating *what a theory entails* regarding whether X has causal powers in addition to those had by B(X), empirically based tests may be unnecessary. Take, for example, a theory that explicitly states that all mental properties supervene on physical properties in such a way that the former have no causal powers in addition to those had by the latter. We can clearly judge that, if this theory is true, M does not have causal powers in addition to those had by P. There is no need to apply either the H[B(X)]-wiggle test or causal isolation test to M, for it is obvious that, based on the fundamental commitments of our theory, M cannot have causal powers in addition to those had by P. Hence, when judging *what a theory entails* regarding whether X has causal powers over and above those had by B(X), empirical methods (or their counterfactual substitutes) for determining whether X has such additional causal powers may very well be unnecessary.

But this is precisely the sort of context in which Kim employs the piggyback principle. As we saw in the introduction, Kim's concern is with the *consistency* of mental-to-physical

causation with the joint acceptance of certain metaphysical claims, i.e. mind-body supervenience, causal closure of the physical domain, mental/physical property dualism, and the causal exclusion principle. Thus, Kim's supervenience argument attempts to show that a certain metaphysical view of the world *entails* mental-to-physical epiphenomenalism. And the metaphysical view is that of non-reductive physicalism. Kim (2005, 22) takes mind-body supervenience and causal closure of the physical domain to be commitments of non-reductive physicalism as a version of physicalism, while mental/physical property dualism is taken to define the non-reductivism of non-reductive physicalism. Now, the causal exclusion principle is, for Kim, a general, metaphysical truth (Kim, 2005, 22). And as we observed in the introduction, Kim's supervenience argument appears to rely on the piggyback principle in order to secure premise (6), which is required for applying the causal exclusion principle in (7). Therefore, Kim's supervenience argument appears to argue that certain metaphysical commitments of non-reductive physicalism *entail* that M supervenes on P in such a way that M cannot have any impact on P* additional to the impact that P has on P*. Now, of course, evaluating this entailment of non-reductive physicalism is not as simple as the evaluation of our sample theory above. The metaphysical commitments of non-reductive physicalism do not explicitly state that M fails to have any causal powers in addition to those had by P. However, in the context of Kim's supervenience argument, the issue is still whether the metaphysical claims central to non-reductive physicalism force the non-reductive physicalist to deny M any causal contribution to P* additional to the effect that P has on P*. And evaluating those entailments of non-reductive physicalism does not clearly require any special testing by way of empirically based methods, for certain metaphysical considerations may be sufficient.

Indeed, S&S themselves appear to allow that certain metaphysical considerations, independent of any empirically based tests, are sufficient for claiming that, *given non-reductive physicalism*, M has no causal powers in addition to those had by P. The leading version of non-reductive physicalism tends to be functionalism, and S&S say the following about functional properties:

> We can grant that a functional property has no causal powers beyond those of its realizer. . . . Functional properties are generally conceived as second-order properties – the property of having some property that fills some functional role. Believing that martinis should be shaken, for instance, is the property of having some property (perhaps neural, or silicon-based, or . . .) that plays the role that defines such a belief. But how could a second-order property – the property of having some first-order property – have powers beyond those of a first-order property? (S&S 2007, 245)

Here, just by analyzing the concept of a functional property, S&S grant that a functional mental property cannot have causal powers beyond those of its physical realizer. That is, on the basis of purely a priori, metaphysical considerations of functionalism, S&S allow that functionalism entails that M has no causal powers in addition to those had by P. Again, then, it seems that evaluating whether, on non-reductive physicalism, M has causal powers in addition to those had by P need not always require applying the H[B(X)]-wiggle test, or the causal isolation test, or any such empirically based test. The metaphysical commitments of non-reductive physicalism can potentially settle the issue.

S&S appear, then, to fail to appreciate the structure of Kim's argument regarding mental epiphenomenalism. Kim is not attempting to show that mental properties are epiphenomenal in the *actual* world. If Kim were seeking to discover the causal role of mental properties in the actual world, then empirically based tests might very well be central to his investigation. And S&S's (2007, 241, 259) insistence that science, rather than armchair philosophy, determine whether mental states are epiphenomenal would make good sense. However, Kim is concerned

with whether the metaphysical commitments central to non-reductive physicalism make it incompatible with mental causation. According to Kim, "The aim of the supervenience argument is to clarify the options available to the physicalist: If you deem yourself a physicalist, you must choose between [mental epiphenomenalism] or [reductive physicalism]" (2005, 54-5). And asserting this incompatibility between non-reductive physicalism and mental causation is entirely consistent with remaining neutral about whether there is any mental causation in the *actual* world. Consequently, in the context of Kim's supervenience argument (or in the context of "the master argument for epiphenomenalism" insofar as it is similarly concerned with *what a theory entails* regarding the causal role of X), empirically based tests may not be essential to the argument. Therefore, in such contexts, one can, as S&S's own analysis of functionalism shows, apparently apply the piggyback principle without relying on empirically based tests in order to determine whether X has an impact on Y additional to that had by B(X).

S&S might reply that I have underestimated the connection they take to hold between the H[B(X)]-wiggle test and whether X has an impact on Y additional to that had by B(X). I have, according to this line of reply, focused too much on how the former can play an epistemic role with respect the latter, for there is a deeper, mind-independent, connection as well. This further connection comes in the form of an interventionist theory of causation, according to which causal claims themselves are to be understood as counterfactual claims involving idealized, non-anthropocentric, interventions.[9] That is, S&S might offer the following, *non-epistemic* account of how the piggyback principle depends on the H[B(X)]-wiggle test in its counterfactual form: CF must, *whether we can analyze it or not*, have a substantive truth value in order for their to be a substantive truth of the matter regarding whether X has an effect on Y additional to the effect

[9] S&S do note the following: "What we do claim is that *X* causes *Y* if and only if a suitably defined intervention on *X* would be associated with some change in *Y* (or in the probability of *Y*)" (2007, endnote 4). Also, S&S reference the interventionist theory of causation advanced by Woodward (2003).

that B(X) has on Y. Thus, the piggyback principle, insofar as it specifies a significant condition for X causing Y, requires that CF have a substantive truth value.

However, as we observed, S&S take the nomological impossibility of CF's antecedent to be sufficient for their argument. And it is not clear that the nomological impossibility of CF's antecedent prevents it from having a substantive truth value. Granted, if the antecedent of a counterfactual is a metaphysical (or broadly logical) impossibility, then there is a clear threat of the counterfactual failing to have any substantive truth value. According to the standard Lewis-Stalnaker (Lewis 1973; Stalnaker 1968) analysis of counterfactuals, a claim of the form "If P were the case, then Q would be the case" is true in a world, W, if and only if (i) there is no possible P-world, or (ii) some (P&Q)-world is closer (more similar) to W than is any (P&~Q)-world. Now, if condition (i) is satisfied, then the counterfactual is typically understood to be vacuous, or only trivially true. Thus, if the antecedent of CF is metaphysically impossible, then there is the clear threat of CF not having any substantive truth value in virtue of it meeting condition (i). However, such an obvious threat of vacuous truth value is no longer present if the antecedent of CF is only nomologically impossible, for then there are metaphysically possible worlds in which the antecedent of CF is true.

And notice that S&S cannot simply beef up their assumed supervenience, so that it extends beyond nomological necessity to include metaphysical necessity. First, such a move is contentious in the context of Kim's supervenience argument. Kim's argument is concerned with mind-body supervenience as a feature of non-reductive physicalism. But Kim (2005, 49) takes mind-body supervenience with metaphysical necessity to be reductionist. Second, S&S presumably want their argument to apply to functionalist accounts of how X is instantiated in virtue of B(X). Now, on functionalism, any functional role associated with X is defined in terms

of causally relating certain inputs to certain outputs. But given that the laws of causality hold with nomological, but not metaphysical, necessity, then B(X), as a realizer of X, will ensure the presence of X with nomological necessity and not with metaphysical necessity.

Furthermore, on the proposed, non-epistemic connection between CF and whether X has an impact on Y additional to that had by B(X), S&S's own position on functional properties entails that CF has a substantive truth value. As we saw, S&S argue that, on functionalism, X cannot have any causal impact on Y additional to whatever impact its functional realizer, B(X), has on Y. And their argument doesn't employ any cheap trick that would make the truth of their conclusion trivial. But if their conclusion can be substantively true only if CF is substantively true, then S&S have also shown that CF can have a substantive truth value.

Nor would it help to take S&S as adopting a verificationalist theory of meaning. On such an interpretation of S&S, the piggyback principle depends on the H[B(X)]-wiggle test the following sort of way: In order for the piggyback principle to be meaningful, it must be possible to empirically determine whether X has an impact on Y additional to that had by B(X). And the H[B(X)]-wiggle test appears to be the only way of determining such a thing. But the H[B(X)]-wiggle test is impossible to perform and, therefore, the piggyback principle is meaningless.

Again, S&S's argument concerning the causal contributions of functional properties undermines this alternative approach. S&S cannot claim that the question of whether X has any causal impact on Y additional to that had by B(X) is meaningless, for they claim the question has a meaningful answer in the case of functionalism. Moreover, with the notorious problem of verificationalism being self defeating, i.e. meaningless on its own terms, S&S would only undermine the plausibility of their empirical approach by resting it on such a discredited view.

Let us summarize the work of this section. S&S criticize the piggyback principle by saying its application requires the H[B(X)]-wiggle test (in either its empirical or counterfactual form), which they claim is the wrong sort of test. However, we have shown that application of the piggyback principle may not require employing the H[B(X)]-wiggle test, or any other empirically based test, especially when it is applied, as it is by Kim, in order to see *what a theory entails* about X. Furthermore, we handled potential replies S&S might offer in an attempt to defend a problematic connection between the piggyback principle and the H[B(X)]-wiggle test. Therefore, S&S fail to tie the piggyback principle to the H[B(X)]-wiggle test, or any other empirically based test, in a way that reveals the piggyback principle to be a don't.

Does the H(CC)-wiggle test present any problems for the piggyback principle?

Up till now, we have focused on how S&S use the H[B(X)]-wiggle test to criticize the piggyback principle. Yet, to address S&S's empirical approach fully, we must also consider their H(CC)-wiggle test and whether it presents any problems for the piggyback principle.

S&S claim that "what Weismann *did* do should serve as our model for what good arguments for epiphenomenalism should be like" (S&S 2007, 240, their italics). And what Weismann did do was use the H(CC)-wiggle test in order to argue that parental phenotypes of tail length are epiphenomenal with respect to offspring genotypes responsible for tail length. And, according to S&S, the H(CC)-wiggle test should serve as a model for good arguments for epiphenomenalism because, as the case of Weismann's experiments shows, the H(CC)-wiggle test does a good job of uncovering cases of epiphenomenalism. In particular, it captures those clear cases of epiphenomenalism in which X and Y are correlated merely by being joint effects

of a common cause.[10] In such cases, S&S (2007, 258-9) say X is "screened off" from Y by a common cause, CC.

Notice, however, that the success of the H(CC)-wiggle test in establishing epiphenomenal conclusions in no way conflicts with the piggyback principle. The piggyback principle is entirely consistent with counting X as epiphenomenal with respect to Y when the H(CC)-wiggling of X is not associated with any change in Y. On S&S's view, such results of the H(CC)-wiggle test indicate that X is screened off from Y by CC. But whenever X is screened off from Y by CC, the piggyback principle won't conflict with judging X to be epiphenomenal with respect to Y. Rather, the piggyback principle will simply entail that there is an additional sort of case in which X can be epiphenomenal with respect to Y, namely, when X fails to have an impact on Y additional to that had by B(X). Thus, the H(CC)-wiggle test can yield good arguments for epiphenomenalism without hindering the piggyback principle from doing so as well—both can be do's in arguments for epiphenomenalism.

Furthermore, S&S's account of when X is screened off from Y can be viewed as naturally extending to include the piggyback principle. Again, according to S&S, X is screened off from Y when the correlation between X and Y is due *merely* or *solely* to their being joint effects of a common cause, CC. And notice that CC apparently screens off X from Y because X fails to make a causal contribution to the production of Y in addition to that made by CC. For example, in the case of the correlation between parental and offspring eye color, the former fails to have any impact on the latter over and above that had by their common cause (i.e. the parental

[10] In addition to cases of correlation between parental phenotypes and offspring genotypes (and phenotypes), S&S (2007) consult the following, standard cases of epiphenomenalism: (i) a circle of colored light on the ceiling of the Astrodome at one time not affecting the shape or color of a circle of light on the ceiling an instant later, where both circles of light are effects of a rotating spotlight aimed at the ceiling (236-7, 258-9), (ii) a barometer reading not affecting a storm, where both are effects of barometric pressure (237, 258-9), and (iii) an image in the mirror at one time not affecting an image in the mirror an instant later, where both are effects of the object in front of the mirror (243, 258-9).

genotype responsible for eye color). And such failure on the part of parental eye color with respect to offspring eye color seems to be precisely why their correlation is due *merely* or *solely* to their being joint effects of a common cause.

Now, in the case where B(X) causes Y, it seems that B(X) can similarly threaten to screen off X from Y. Although the relation between B(X) and X is assumed by S&S (2007, 239) not to be a causal relation, the instantiation of B(X) ensures the instantiation of X with nomological necessity nonetheless. Thus, given that B(X) both causes Y and determines the instantiation of X with nomological necessity, it seems a natural extension of S&S's notion of epiphenomenalism that B(X) be a potential candidate for screening off X from Y. That is, just as the correlation between X and Y can be due *merely* to CC, the correlation between X and Y can be due *merely* to B(X). Furthermore, recall that when the correlation between X and Y is due *merely* to CC, it is apparently because X fails to make a causal contribution to the production of Y additional to that made by CC. Therefore, S&S's notion of epiphenomenalism naturally extends to include X as epiphenomenal with respect to Y when X fails to have an impact on Y additional to the impact that B(X) has on Y. And this further sort of epiphenomenalism is precisely what the piggyback principle captures.

S&S, of course, argue that the piggyback principle is tied to the H[B(X)]-wiggle test and, therefore, this additional sort of epiphenomenalism aught to be rejected. However, we have already thoroughly dealt with that line of argument and, consequently, are focused on showing that the piggyback principle can accommodate the epiphenomenal conclusions of the H(CC)-wiggle test.

S&S will likely respond that the H(CC)-wiggle test is capable of establishing not just epiphenomenalism but also causation. That is, if the H(CC)-wiggling of X is associated with a

change in Y, then there is good evidence that X is a cause of Y (or at least just as good evidence as there would be for the claim that X is epiphenomenal with respect to Y, were the intervention on X not associated with a change in Y). And, S&S might claim, the piggyback principle cannot accommodate this ability of the H(CC)-wiggle test to establish that X causes Y, for the H(CC)-wiggling of X can be associated with a change in Y even though X fails to have an impact on Y additional to that had by B(X).

But notice that while the H(CC)-wiggle test may establish causation, it does not clearly have the fine tuned conclusion that X causes Y. Suppose that one performs the H(CC)-wiggle test, and the intervention on X is associated with a change in Y. Now, in performing this test, one does not do an intervention on just X but, rather, on B(X) as well. Recall that, for S&S, the H(CC)-wiggling of X does not involving holding fixed B(X). Instead, one wiggles the two properties together. Therefore, the intervention is not a finely tuned manipulation of X alone but, rather, an intervention on both X and B(X). Consequently, the conclusion of the H(CC)-wiggle test cannot be so precise as to identify X as the cause of Y. Instead, we at best get the conclusion that either X or B(X) causes Y.[11] Granted, this disjunctive conclusion is inclusive, so that there is room for both X and B(X) to cause Y.[12] But the point remains that the H(CC)-wiggle test does not itself yield such a conclusion but, rather, the weaker conclusion that at least one of the two properties are causally efficacious with respect to Y.[13]

---

[11] Now, in the context of non-reductive physicalism, where X is a mental property and B(X) its physical base, the causal conclusion that either X or B(X) causes Y may very well entail a more precise conclusion. This is because the physicalist's commitment to the causal closure of the physical domain may kick in to ensure that B(X) causes Y. Such a conclusion, of course, does not have X as a cause of Y but, rather, helps set the stage for the concern that X is causally excluded by B(X).

[12] S&S (2007, 256) argue for this conclusion on the grounds that an intervention on X will be associated with a change in Y only if an intervention on B(X) will be associated with a change in Y. As I have just indicated, however, S&S's favored sort of intervention on X is not fine tuned enough to get the conclusion that X causes Y but, rather, that X or B(X) causes Y.

[13] Notice that it doesn't help here to insist that the *instances* of X and B(X) are identical with one another. Granted, the causal conclusion of the H(CC)-wiggle test can then be extended to the claim that an instance of X causes an

Furthermore, this leaves just the right amount of room for the piggyback principle. X can be rendered epiphenomenal according to the piggyback principle without conflicting with the conclusion that either X or B(X) causes Y. Should the piggyback principle end up counting X as epiphenomenal with respect to Y, the piggyback principle can nonetheless leave B(X) as a cause of Y and thereby not conflict with the proper causal conclusion of the H(CC)-wiggle test.

Now, of course, S&S might argue that there is no such room for the piggyback principle because it depends on the H[B(X)]-wiggle test, which is the wrong test. But, again, we have already thoroughly dealt with that line of reasoning.

In this section, then, we have seen that the H(CC)-wiggle test does not pose a threat to the piggyback principle. The H(CC)-wiggle test can be successful in establishing conclusions for *epiphenomenalism* without conflicting at all with the piggyback principle being similarly successful. Indeed, the sort of epiphenomenalism that S&S take the former to indicate can, we saw, be seen as naturally extending to include the sort of epiphenomenalism captured by the latter. And when it comes to the H(CC)-wiggle test's ability to establish conclusions for *causation*, we saw that, once the proper extent of the casual conclusions is clarified, it too is entirely consistent with the piggyback principle. Thus, the H(CC)-wiggle test is no more a threat to the piggyback principle than is the H[B(X)]-wiggle test.[14]

---

instance of Y. However, recalling that we are interested in a non-reductive context, the instance of X will also be an instance of a property distinct from X, i.e. B(X). Consequently, the X-instance may cause Y in virtue of being a B(X)-instance, or *qua* B(X)-instance, and not in virtue of being an X-instance, or *qua* X-instance. But the latter sort of causation is what we are concerned with securing. (See note 2).

[14] In addition to lessons regarding methods of testing, another lesson S&S draw from Weismann's experiments is that arguments for epiphenomenalism should be limited: "They should aim to show that one class of properties does not affect a second class, not that the first has no effects at all" (S&S 2007, 241). Now, as noted earlier, Kim's argument is so limited, to a certain extent, in virtue of its two stages; one stage shows how M is epiphenomenal with respect to P and the other stage shows how this leads to M being epiphenomenal with respect to a different sort of property as well, namely, M*. Furthermore, the piggyback principle is formulated regarding the causal efficacy of a property *with respect to a specific effect* and, hence, can clearly be employed in arguments for epiphenomenalism that are as limited in scope as S&S desire.

Conclusion

S&S's make a distinction between two ways of empirically testing for the causal efficacy of a property, i.e. the H(CC)-wiggle test and the H(B(X)]-wiggle test, and judge the former to be a do and the latter a don't. And from this empirical standpoint, S&S claim that arguments like Kim's supervenience argument are flawed insofar as they employ the piggyback principle. However, I have provided an extensive defense of the piggyback principle against S&S's empirical approach. Therefore, S&S have failed to show that Kim's supervenience argument rests on a don't in reasoning about the epiphenomenalism of a property.

References

Davidson, D. (1970), Mental Events. In L. Foster and J. Swanson (eds.), *Experience and Theory*. Amherst: University of Massachusetts Press, 79-101. Reprinted in D. Davidson, *Essays on Actions and Events*, Oxford: Clarendon Press, 1980, 207-27.

Davidson, D. (1995). Thinking Causes. In J. Heil and A. Mele (eds.), *Mental Causation*. Oxford: Clarendon Press, 3-17.

Kim, J. (1989). The Myth of Nonreductive Materialism. *Proceedings of the American Philosophical Association*, 63, 31-47. Reprinted in J. Kim, *Supervenience and Mind*, New York: Cambridge University Press, 1993, 265-284.

Kim, J. (1998). *Mind in a Physical World*, Cambridge, MA: MIT Press.

Kim, J. (1995). Can Supervenience and 'Non-Strict Laws' Save Anomalous Monism? In J. Heil and A. Mele (eds.), *Mental Causation*. Oxford: Clarendon Press, 19-26.

Kim, J. (2005). *Physicalism or Something Near Enough*, Princeton: Princeton University Press.

Lewis, D. (1973). *Counterfactuals*, Oxford: Blackwell Publishers.

McLaughlin, B. (1995). On Davidson's Response to the Charge of Epiphenomenalism. In J. Heil and A. Mele (eds.), *Mental Causation*. Oxford: Clarendon Press, 27-40.

Shapiro, L. and Sober, E. (2007). Epiphenomenalism: The Dos and the Don'ts. In P. Machamer and G. Wolters (eds.), *Thinking about Causes: From Greek Philosophy to Modern Physics*. Pittsburgh: University of Pittsburgh Press, 235-264.

Sosa, E. (1995). Davidson's Thinking Causes. In J. Heil and A. Mele (eds.), *Mental Causation*. Oxford: Clarendon Press, 41-50.

Stalnaker, R. (1968). A Theory of Conditionals. In N. Rescher (ed.), *Studies in Logical Theory, American Philosophical Quarterly Monograph Series*, No. 2. Oxford: Blackwell, 98-112.

Woodward, J. (2003). *Making Things Happen*. New York: Oxford University Press.