Robustness, Discordance, and Relevance

Jacob Stegenga†‡

Robustness is a common platitude: hypotheses are better supported with evidence generated by multiple techniques that rely on different background assumptions. Robustness has been put to numerous epistemic tasks, including the demarcation of artifacts from real entities, countering the "experimenter's regress," and resolving evidential discordance. Despite the frequency of appeals to robustness, the notion itself has received scant critique. Arguments based on robustness can give incorrect conclusions. More worrying is that although robustness may be valuable in ideal evidential circumstances (i.e., when evidence is concordant), often when a variety of evidence is available from multiple techniques, the evidence is discordant.

1. Introduction. Robustness is a recent term for a common platitude: hypotheses are better supported with plenty of evidence generated by multiple techniques that rely on different background assumptions. A simple example: Hacking (1983) argued that when a cellular structure is observed with different types of microscopes, we have more reason to believe that the structure is real. Salmon's (1984) "common-cause" argument is similar to the notion of robustness: Avogadro's number is consistently demonstrated using experiments based on different methodologies: Brownian motion, alpha particle decay, X-ray diffraction, blackbody radiation, and electrochemistry, and the common cause for this consistency is the existence of molecules. I have seen the term "robustness" first used as a methodological adage by the statistician George Box in 1953: a robust statistical analysis is one in which its conclusions are consistent despite changes in underlying analytical assumptions. In philosophy of

†To contact the author, please write to: Jacob Stegenga, Department of Philosophy, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093; e-mail: jstegenga@ucsd.edu.

‡Numerous people have given valuable commentary on this paper, including Nancy Cartwright, Craig Callender, Matthew Brown, Marta Halina, Kareem Khalifa, Eric Martin, Tarun Menon, Boaz Miller, Brendan Ritchie, Samuel Schindler, Léna Soler, Eran Tal, and audiences at the Canadian Society for the History and Philosophy of Science, the Philosophy of Science Association, and members of the University of California, San Diego Philosophy of Science Reading Group.

Philosophy of Science, 76 (December 2009) pp. 650–661. 0031-8248/2009/7605-0032$10.00

science I have seen the term first used with respect to models: results consistent across multiple models (with different background assumptions) are 'robust' and so more likely to be true (Levins 1966; Wimsatt 1981). Nearly every philosopher of science interested in evidence has, at least in passing, extolled the virtues of experimental robustness.[1] Despite this, the notion has received little explication or philosophical critique. In this paper I aim at least to begin such an investigation.

First, a definition:

> **Robustness**. The state in which a hypothesis is supported by evidence from multiple techniques with independent background assumptions.

Evidence from multiple techniques (or "multimodal evidence") provides greater support to a hypothesis than evidence from a single technique (or "monomodal evidence").[2] "Technique" is unspecified in this definition, since robustness can be a feature of statistical analyses, models, and experiments, though in this paper my focus will be on experimental robustness. Presumably robustness admits of degrees, but as I will argue, one of the key challenges facing robustness is the difficulty (or impossibility) of specifying the degree of robustness for any hypothesis.

Robustness is often presented as an epistemic virtue that ensures objectivity. Champions of robustness claim that robust evidence can demarcate artifacts from real entities, counter the "experimenter's regress," ensure appropriate data selection, and resolve evidential discordance. Consider the worry about artifacts: If a new technique shows x (an entity, process, relation, etc.), this might be due to a quirky aspect of the technique rather than a real description of x. *Response*: If x is observed with multiple methods, it is extremely unlikely that x is an artifact (Hacking 1983). Consider the "experimenter's regress": Good evidence is generated from properly functioning techniques, but properly functioning techniques are just those that give good evidence (Collins 1985). *Response*: This vicious experimental circle is broken if we get the same result from multiple techniques (Culp 1994). Consider the concern about data selection: Scientists use only some of their data, selected in various ways for various reasons, and the rest are ignored; but how do we know that the selection process gives true results? *Response*: Vary the selection criteria, and in-

1. Including Horwich (1982), Cartwright (1983), Franklin and Howson (1984), Howson and Urbach (1989), Achinstein (2001), Staley (2004), and Bechtel (2006). Several interesting papers were devoted to robustness at a workshop in June 2008 in Nancy, France.

2. I use "multimodal evidence" as shorthand for evidence of different kinds, or the total set of evidence from different "modes," that relate to the same hypothesis. Elsewhere I provide a more detailed discussion of the notion.

variant results are more likely to be true (Franklin 2002). Finally, consider discordant data: Multiple experimental results are not always coherent; which results should we believe? *Response*: Conduct more experiments until they yield convergent results.

Indeed, robustness has been used as an argument for realism. The canonical example is Perrin's arguments for the reality of atoms (described in Nye 1972 and discussed in Cartwright 1983, Salmon 1984, and Mayo 1996). Given the variety of epistemic tasks placed on robustness and given the frequency with which the notion is appealed to, it has received surprisingly little direct attention. This paper discusses several problems with robustness in an attempt to provide needed constraints on the concept. Robustness is valuable in ideal evidential circumstances, when all available evidence is concordant. The difficulty for robustness is, of course, that most evidence is not concordant. When multimodal evidence is available for a given hypothesis, the evidence is usually discordant; that is, evidence from various modes displays varying degrees of consistency, congruity, and intensity. Given the vicissitudes of evidence, scientists must choose sets of evidence that they deem most relevant to their given tasks. Evidence of varying degrees of quality is more or less confirming of and more or less relevant to a particular hypothesis. The value of robustness is mitigated by the problem of discordant evidence; discordance entails the problem of choice, or of relevance. We lack systematic methods for assessing and combining multimodal evidence, and without such methods, robustness is limited to a qualitative or intuitive notion.

2. Three Easy Problems. Prior to discussing what I consider to be the 'hard' problem of robustness, I discuss three 'easy' problems. First, scientists do not always have multiple techniques; second, knowing that multiple techniques are independent is difficult (as is knowing in what way multiple techniques should be independent); and finally, concordant multimodal evidence will not necessarily give a correct conclusion. None of these problems taken alone completely repudiates the value of robustness. Indeed, it is a (trivially) important epistemological notion and methodological strategy. However, the value of robustness is heavily mitigated upon consideration of the following three problems.

Generating concordant multimodal evidence is difficult. Scientists need different kinds of evidence generated by independent techniques, but they do not always have multiple techniques to study the same subject. New techniques are introduced into scientific practice for a good reason: they give evidence on a new subject, or on a smaller or larger scale, or in a different context, than existing techniques. Even if multiple techniques do exist, it is not always clear that the techniques are independent. Bechtel (2006) argued that often new techniques are calibrated to existing tech-

niques, and so even when both techniques provide concordant results, the techniques cannot be said to be independent. Furthermore, determining what criteria should be used for the independence between techniques is a difficult problem; elsewhere I call this the "individuation problem" for multimodal evidence (Stegenga n.d.). Simply put, *having* independent modes of evidence and *knowing* that they are properly independent are difficult; since robustness requires multiple modes of evidence, an incomplete or vague individuation of evidential modes will leave robustness as an incomplete or vague notion, and hence robustness-style arguments will be vague or inconclusive.

This is not to claim that robustness is a useless methodological strategy—Perrin's arguments for the existence of molecules, the canonical example based on concordant multimodal evidence, were convincing—it is simply to state what scientists already know: generating multimodal evidence is difficult.

One might think that multiple invalid arguments that reach the same conclusion give no better reason to believe this conclusion than a single invalid argument reaching the same conclusion. Similarly, multiple methodologically mediocre experiments, or multiple epistemically unrelated experiments, or multiple experiments with implausible background assumptions give no better reason to believe a hypothesis than a single experiment does (let alone a single well-performed experiment with plausible background assumptions). A detailed case study by Rasmussen (1993) provided an instance of this problem: multiple methods of preparing samples for electron microscopy demonstrated the existence of what is now considered an artifact.[3] The fact that such evidential diversity was used as an argument for the reality of an artifact mitigates the epistemic value of robustness. The problem demonstrated by Rasmussen can be generalized: concordant multimodal evidence can support an incorrect conclusion. Robustness is a type of no-miracles argument: it would be miraculous if multiple independent experiments showed x (where x is an entity, or a process, or a constant, or a relation) and x was not real. You may not be swayed by no-miracles arguments. You might, perhaps, think that the no-miracles argument commits the base rate fallacy (Magnus and Callender 2004). For now I remain agnostic on the general importance of no-miracle arguments and worry only about difficulties with robustness.

Somewhat facetiously, I have called the three problems of robustness discussed in this section *easy*. They are, of course, anything but. Scientists need multiple techniques that give evidence on the same subject, while ensuring that such techniques are sufficiently independent. Scientists are

3. This case study was controversial; see responses by Culp (1994) and Hudson (1999) and the counterresponse by Rasmussen (2001).

adept at grappling with these challenges. However, the problem raised by Rasmussen indicates that robustness alone can lead to an incorrect conclusion. Robustness requires having multiple modes of evidence, knowing that multiple modes of evidence are independent, and knowing how they should be independent, and yet remains fallible.

3. Discordance. Discordant data is a fourth problem for robustness. If multiple independent experimental techniques provide greater epistemic support to a hypothesis, it is unclear what support is provided to a hypothesis in the more common situation in which multiple techniques give results that are inconsistent or incongruent. Franklin (2002) recently raised the problem of discordance and suggested that it can be readily solved by various methodological strategies, which prominently include the strategy of generating more data from independent techniques. While I think Franklin is correct to identify discordance as a problem for what he calls the "epistemology of evidence" and his appeal to a plurality of reasoning strategies is valuable, I argue that what he considers as a solution to the problem of discordance is better construed as the source of a problem.

Discordance is really two separate problems of evidence: inconsistency and incongruity. Inconsistency is straightforward: petri dishes suggest x and test tubes suggest $\neg x$. In the absence of a methodological metastandard, there is no obvious way to reconcile various kinds of inconsistent data. Incongruity is even more troublesome. How is it even possible for evidence from different types of experiments to cohere? Evidence from different types of experiments is often written in different 'languages'. Petri dishes suggest x, test tubes suggest y, mice suggest z, monkeys suggest $0.8z$, mathematical models suggest $2z$, clinical experience suggests that sometimes y occurs, and human case control studies suggest y while randomized control trials suggest $\neg y$. To consider multimodal evidence as evidence for the same hypothesis requires more or fewer inferences between evidential modes. The various 'languages' of different modes of evidence can be translated into other languages with the right background assumptions. If techniques actually have independent background assumptions, they might simply generate incommensurable data. The background assumptions necessary for such translations have varying degrees of plausibility. If they are not plausible, then it is hard to see how multimodal evidence provides greater epistemic support to a hypothesis than a single mode of evidence does.

Another dimension of discordance is the degree of "intensity" or salience of results. Consider Galison's (1987) distinction between golden-event experiments and statistical experiments within the particle physics community. Golden-event experiments give "intense" evidence for particularly rare events. In contrast, statistical experiments measure more fre-

quent but less striking events. If different kinds of evidence with different intensities support opposite conclusions, there is no obvious way to compare or combine such evidence in an orderly or quantifiable way, let alone to compare such a combination of evidence to evidence from a single kind of experiment. Philosophers have long wished to quantify the degree of support that evidence provides to a hypothesis. At best, the problem of discordance implies that robustness is limited to a qualitative notion. But if robustness is a qualitative notion, how do we demarcate robust from nonrobust evidence? At worst, the problem of discordance implies that evidence of different kinds cannot be combined in any coherent way.

One might respond that discordance is not a problem for robustness, since by definition robust evidence is generated when multiple independent techniques give the *same* result on the *same* hypothesis. To appeal to discordant data as a challenge for robustness simply misses the point: robustness just is concordance of multiple kinds of evidence, so no one would say that evidence that is discordant could also be robust. Fair enough: the problem of discordance is not a knockdown argument against the value of robustness per se. Rather, discordance demonstrates an important constraint on the value of robustness. Robustness and its corresponding methodological prescription—get more data! (of different kinds)—are trivially valuable. However, the prescription to get more data, from different kinds of experiments, is not something that scientists need to be told; they already follow this commonsense maxim. I share the intuition that multimodal evidence *does* (often) provide greater epistemic support to a hypothesis than monomodal evidence does—at least when all independent techniques are concordant. Unfortunately, multimodal evidence, when available, is rarely concordant. Further, robustness-style arguments presuppose a principled and systematic method of assessing and amalgamating multimodal evidence, and without such methods of combining evidence, robustness arguments are merely intuitive or qualitative.

That multiple independent techniques often display discordant evidence is an empirical claim. Some might think this a weakness of the above argument. However, the opposite is, of course, also an empirical claim—that multiple independent techniques often display concordance—and this is an empirical claim that seems false. History of science might occasionally provide examples of apparent concordance, but concordance is easier to see in retrospect, with a selective filter for reconstructions of scientific success. Much history of science tends to focus on the peaks of scientific achievement rather than the winding paths in the valleys of scientific effort: at least, the history of science that *philosophers* tend to notice, such as Nye's account of Perrin's arguments for atoms, is history of scientific success. Philosophers have focused on the peaks of scientific success, but

the lovely paths of truth in the valleys of scientific struggle are often discordant.

Finally, one might ask: Since we're all fallibilists, doesn't robustness seem like a valuable methodological strategy? After all, what else can we do but investigate something with as many modes of investigation as possible? There is something intuitively appealing about the methodological strategy, and I have not meant to argue against this intuition. Rather, I have raised the problem of discordance, or discordant multimodal evidence, as an indication of an important constraint on robustness. Appeals to robustness are often philosophical cheap tricks. How useful robustness is as a methodological strategy depends on what we can actually do with it; and without systematic ways of assessing and amalgamating discordant multimodal evidence, I do not think there is much we can do with it; the more a particular body of evidence is discordant, the less useful the methodological strategy of robustness is.

Another way of putting this is as follows. Concordant multimodal (robust) evidence for x is sufficient, but not necessary, for a high probability of x. Now, notice two problems that stem from this vague formulation. First, specifying the high probability of x depends on principled methods of quantifying concordance and assessing and amalgamating multimodal evidence, which we lack, and thus, we cannot specify the high probability of x. That x even has a high probability is merely an intuition. Second, x might be true despite a failure of robustness, but robustness-style arguments do not tell us what to believe in situations of evidential discordance.

For most of the twentieth century, philosophy of science considered idealizations of evidence. Carnap (1950), for example, developed confirmation theory "given a body of evidence e" without worrying about what constitutes a "body." In ideal evidential circumstances, robustness is a valuable epistemic guide. Real science is almost never in ideal evidential circumstances; recent historical and sociological accounts of science have reminded philosophers of this messy detail. The following example illustrates the problem, though it should hardly be needed: discordance is ubiquitous.

Epidemiologists do not know how the influenza virus is transmitted from one person to another. The mode of infectious disease transmission has been traditionally categorized as either "airborne" or "contact."[4] A causative organism is classified as airborne if it travels on aerosolized particles through the air, often over long distances, from an infected individual to the recipient. A causative organism is classified as contact if

4. I simplify for purposes of exposition.

it travels on large particles or droplets over short distances and can survive on surfaces for some time. Clinicians tend to believe that influenza is spread only by contact transmission. Years of experience caring for influenza patients and observing the patterns of influenza outbreaks has convinced them that the influenza virus is not spread through the air. If influenza is an airborne virus, then patterns of influenza transmission during outbreaks should show dispersion over large distances, similar to other viruses known to be spread by airborne transmission. Virtually no influenza outbreaks have had such a dispersed pattern of transmission. Moreover, nurses and physicians almost never contract influenza from patients, unless they have provided close care to a patient with influenza.

Conversely, some scientists, usually occupational health experts and academic virologists, believe that influenza could be an airborne virus. Several animal studies have been performed, with mixed conclusions. One prominent case study often referred to is based on an airplane that was grounded for several hours, in which a passenger with influenza spread the virus to numerous other passengers. Based on seating information and laboratory results, investigators were able to map the spread of the virus; this map was evidence that the influenza virus was transmitted through the air. More carefully controlled experiments are difficult. No carefully controlled human experiments can be performed for ethical reasons. However, in the 1960s, researchers had prisoner 'volunteers' breathe influenza through filters of varying porosity; again, interpretations of results from these experiments were varied but suggested that influenza could be airborne. Mathematical models of influenza transmission have been constructed, using parameters such as the number of virus particles emitted during a sneeze, the size of sneeze droplets upon emission, the shrinking of droplet size in the air, the distance of transmission of particles of various size, and the number of virus particles likely to reach a 'target' site on recipients. The probability of airborne influenza transmission is considered to be relatively high given reasonable estimates for these parameters.

Even when described at a coarse grain, the various types of evidence regarding the mode of influenza transmission illustrate the problem of discordance. Some scientists argue (using mathematical models and animal experiments) that influenza is transmitted via an airborne route, whereas others argue (based on clinical experience and observational studies) that influenza is transmitted via a contact route. Such discordance demonstrates the poverty of robustness: multiple experimental techniques and reasoning strategies have been used by different scientists, but the results remain inconclusive. A single case study does not, of course, demonstrate the ubiquity of discordance; rather, the case study is merely meant as an illustration of what is meant by discordance.

Franklin (2002) suggests that robustness helps resolve discordant data, but I have argued the converse: discordant data diminish the epistemic value of robustness. Epistemic guidance is needed most in difficult cases, when multiple independent techniques produce discordant evidence. In such cases robustness is worse than useless, since the fact of multiple modes of evidence is the source of the problem. Real science is almost always confronted with the problem of discordance; in such circumstances scientists must decide which evidence is most relevant. We could call this the "new-new problem of induction": are there principled methods to amalgamate discordant multimodal evidence, and is there a way to *justify* methods of evidence amalgamation?

4. Relevance. If evidence for a particular hypothesis from all types of techniques is concordant, then scientists do not need to choose which mode of evidence is more or less relevant to the hypothesis. At least they are not faced with contradictory evidence. But given discordant evidence, scientists are faced with choices: data from some techniques support a hypothesis, while other data do not (inconsistency) or, worse, data from some techniques simply require too many implausible assumptions to consider them as evidence for the same hypothesis as evidence from other techniques (incongruity). Indeed, the basis of many scientific controversies is the problem of relevance: one group of scientists believes that evidence from some techniques is relevant to a hypothesis while another group of scientists believes that evidence from other techniques is more relevant.

Cartwright (2007) has argued that modes of evidence are of varying quality and are of varying relevance to a given policy; but of course, the problem of differential quality and relevance of evidence also applies to hypotheses. There is often a trade-off between quality and relevance. Evidence of high quality could be generated from experiments with low relevance; such experiments Cartwright calls 'clinchers'. Evidence of high relevance could be generated from experiments that are of low quality; Cartwright calls these experiments 'vouchers'. The challenge facing policy makers is as much a challenge for any scientist considering a hypothesis of relatively general scope: some evidence must be selected as relevant from discordant data generated by multiple kinds of techniques. Which kinds of data are most relevant to the hypothesis? Which kinds of data are high-quality? Scientists and policy makers can consider data from all kinds of experiments, or only data from high-quality experiments, or only data from one particular kind of experiment. How should they choose?

Scientists lack universal criteria for making decisions regarding relevance, though particular disciplines do have criteria for determining what counts as high-quality evidence. As Galison (1987) argues, one tradition in physics considers an image of a "golden event" to be high-quality

evidence. The evidence-based medicine movement rank orders various kinds of experiments, with the randomized control trial (RCT) considered the highest quality of evidence; prospective cohort studies, case control studies, observational studies, and case reports normally follow RCTs in descending order of quality. However, since high-quality evidence is not necessarily the most relevant to a given policy (or hypothesis), Cartwright has argued that multiple kinds of evidence must be considered (and not just 'clinchers'; see also Worrall 2002). Deliberation about a policy or a general hypothesis should use evidence that is ideally both high-quality and relevant, but this is not often available. To illustrate, I will continue the example based on the mode of influenza transmission, though, again, such an example should hardly be necessary since the problem of relevance is ubiquitous.

Despite ignorance about the mode of influenza transmission, policy makers in public health jurisdictions around the world have been expected to develop guidelines regarding the type of mask that should be provided to health care workers in the case of an influenza pandemic. If influenza is spread through the air, then active filtration masks are necessary; these masks are cumbersome, must be custom-fitted to the face of health care workers, and are relatively expensive. If influenza is spread by contact transmission, then surgical masks are sufficient; surgical masks are cheap and readily accessible and do not require custom fitting. The guideline written by the U.K. Department of Health is exemplary: the authors claim that the "balance of evidence points to large droplet and direct and indirect contact as the most important routes of transmission" of influenza (2007, 6).[5] The trouble is that what "balance of evidence" means in this guideline is unspecified. The policy makers had no principled method to amalgamate the discordant multimodal evidence regarding the mode of influenza transmission, nor did they have criteria to determine which particular kinds of evidence were most relevant to the question of what kinds of masks should be recommended to health care workers during an influenza pandemic. Vague appeals to balancing evidence are perhaps all that can be done, given discordance. Unfortunately, as argued in Sections 2 and 3, arguments that appeal to a balance of evidence can support incorrect conclusions.

Determining what evidence is relevant could mitigate the problem of discordance. A discordant evidential situation could be rendered more concordant if some of the discordant evidence were deemed less relevant. At the beginning of this section I suggested that perhaps many scientific controversies are controversies just because different scientists consider

5. Hedging their bets, or worried about massive lawsuits, the writers of this guideline also claim: "Airborne, or fine droplet transmission, may also occur."

different kinds of evidence relevant. Conversely, scientific controversies could be closed by determining what evidence is relevant to the undecided hypothesis. However, this just shifts the difficulty from the problem of discordance to the problem of relevance.

5. Conclusion. Though great epistemic weight has been placed on robustness, it has received little philosophical explication. I have illustrated several difficulties with robustness and argued that it is constrained in the following ways: multiple independent techniques are required to generate robust evidence; robust evidence must be consistent and congruent, that is, multimodal evidence must support the same hypothesis; certain kinds of evidence must be selected as more or less relevant to a hypothesis; finally, robust evidence is a qualitative notion. Scientists do not always have multiple techniques. Even if they do, multiple techniques can give incorrect results. Discordant evidence mitigates the value of robustness, since scientists lack criteria for combining inconsistent or incongruent evidence. Finally, the fact of discordance raises the problem of relevance: from among discordant data, scientists much choose which data are relevant to their given tasks.

REFERENCES

Achinstein, Peter (2001), *The Book of Evidence*. New York: Oxford University Press.
Bechtel, William (2006), *Discovering Cell Mechanisms*. Cambridge: Cambridge University Press.
Box, George (1953), "Non-normality and Tests on Variances", *Biometrika* 40: 318–335.
Carnap, Rudolf (1950), *Logical Foundations of Probability*. Chicago: University of Chicago Press.
Cartwright, Nancy (1983), *How the Laws of Physics Lie*. Oxford: Clarendon.
——— (2007), "Are RCTs the Gold Standard?", *Biosciences* 2: 11–20.
Collins, Harry (1985), *Changing Order: Replication and Induction in Scientific Practice*. Chicago: University of Chicago Press.
Culp, Sylvia (1994), "Defending Robustness: The Bacterial Mesosome as a Test Case", in David Hull, Micky Forbes, and Richard M. Burian (eds.), *PSA 1994: Proceedings of the 1994 Biennial Meeting of the Philosophy of Science Association,* vol. 1. East Lansing, MI: Philosophy of Science Association, 46–57.
Franklin, Allan (2002), *Selectivity and Discord: Two Problems of Experiment*. Pittsburgh: University of Pittsburgh Press
Franklin, Allan, and Colin Howson (1984), "Why Do Scientists Prefer to Vary Their Experiments?", *Studies in History and Philosophy of Science* 15: 51–62.
Galison, Peter (1987), *How Experiments End*. Chicago: University of Chicago Press.
Hacking, Ian (1983), *Representing and Intervening*. Cambridge: Cambridge University Press.
Horwich, Paul (1982), *Probability and Evidence*. Cambridge: Cambridge University Press.
Howson, Colin, and Peter Urbach (1989), *Scientific Reasoning: The Bayesian Approach.* LaSalle, IL: Open Court.
Hudson, Robert G. (1999), "Mesosomes: A Study in the Nature of Experimental Reasoning", *Philosophy of Science* 66 (2): 289–309.
Levins, Richard (1966), "The Strategy of Model Building in Population Biology", *American Scientist* 54: 421–431.

Magnus, P. D., and Craig Callender (2004), "Realist Ennui and Base Rates", *Philosophy of Science* 71 (3): 320–338.

Mayo, Deborah (1996), *Error and the Growth of Experimental Knowledge*. Chicago: University of Chicago Press.

Nye, Mary Jo (1972), *Molecular Reality: A Perspective on the Scientific Work of Jean Perrin*. London: Macdonald.

Rasmussen, Nicolas (1993), "Facts, Artifacts, and Mesosomes: Practicing Epistemology with the Electron Microscope", *Studies in History and Philosophy of Science* 24 (2): 221–265.

——— (2001), "Evolving Scientific Epistemologies and the Artifacts of Empirical Philosophy of Science: A Reply Concerning Mesosomes", *Biology and Philosophy* 16: 629–654.

Salmon, Wesley (1984), *Scientific Explanation and the Causal Structure of the World*. Princeton, NJ: Princeton University Press.

Staley, Kent (2004), "Robust Evidence and Secure Evidence Claims", *Philosophy of Science* 71: 467–488.

Stegenga, J. (n.d.), "Multimodal Evidence", unpublished manuscript.

U.K. Department of Health (2007), *Pandemic Influenza: Guidance for Infection Control in Hospitals and Primary Care Settings*. London: U.K. Department of Health.

Wimsatt, William (1981), "Robustness, Reliability, and Overdetermination in Science", in Marilyn B. Brewer and Barry E. Collins (eds.), *Scientific Inquiry and the Social Sciences: A Volume in Honor of Donald T. Campbell*. San Francisco: Jossey-Bass, 124–163.

Worrall, John (2002), "*What* Evidence in Evidence-Based Medicine?", *Philosophy of Science* 69 (Proceedings): S316–S330.