Use-Novelty, Gellerization, and Severe Tests

Abstract

This paper analyzes Deborah Mayo's recent criticism of use-novelty requirement . She claims that her severity criterion captures actual scientific practice better than use-novelty, and that use-novelty is not a necessary condition for severity. Even though she is right in that there are certain cases in which evidence used for the construction of the hypothesis can test the hypothesis severely, I do not think that her severity criterion fits better with out intuition about good tests than use-novelty. I argue for this by showing a parallelism in terms of severity between the confidence interval case and what she calls 'gellerization'. To account for the difference between these cases, we need to take into account certain additional considerations like a systematic neglect of relevant alternatives.

Word Count: 4919

1. Introduction

Deborah Mayo's recent analyses of the notion of use-novelty (1996, especially chs. 6, 8 and 9) added interesting new insights on this issue. The conclusion of her analyses was that sometimes evidence used to construct a theory can also be used to test the theory severely. The purpose of this paper is to examine her analyses. First, I show that Mayo is absolutely right about this claim in the most extreme cases. Then I proceed to a more realistic case. Even here, there is one way to interpret severity and use-construction so that some evidence used to construct a theory tests the theory severely, but if she takes this interpretation, even gellerized hypotheses (the worst kind of use-constructed hypotheses) should be considered as severely tested. In the last section, I discuss briefly the implications of my analysis for the use-novelty requirement itself.

2. Use-novelty and Severity

The notion of novel prediction as a criterion of good testing has a long history, but Popper was the one who brought the notion into contemporary debate (e.g., Popper 1962, 241; see also Giere 1983 for the historical

background of this idea). Popper's disciples have tried many different ways to refine this idea, and the notion of use-novelty, proposed by Zahar (1973) and Worrall (1978a, 1978b, 1989), seems the most promising one.[1]

The basic idea of use-novelty is simple: if evidence e is used in the construction of a hypothesis H, i.e. if H is deliberately constructed so that H fits with e, then e does not count as a support for H. Even though there is some intuitive plausibility with this idea, there have been several attacks on the idea of use-novelty from various directions.[2] Among them, Mayo's criticism seems particularly forceful because she shares most of the background intuition with those who advocate use-novelty.

Mayo claims that her severity criterion captures Popper's intuition and scientists' practice better than use-novelty, and use-novelty is not a necessary condition for severity in her sense. So let us first look at Mayo's severity criterion. She proposes several different formulations of this notion (Mayo 1996, 180-181), but differences between them are not important for our present purposes. Basically, the severity of a test of a hypothesis H is represented as the conditional probability that H fails the test, given that H is false. For example, suppose that the tested hypothesis H is that a given coin is not 'fair' (fifty-fifty probability of heads and tails), and also suppose that we decide that H passes the test when we get more than 60 heads or less than 40 heads out of 100 coin-tossing trials. If H is false (i.e., if the coin is fair), there is a very high probability (.94) that the test procedure fails H.[3] This is a severe test in Mayo's sense. The value of the probability of failure is called the severity of the test, and ranges from 0 to 1. Needless to say, the greater the severity in this sense, the more severe the test is.

In terms of the relationship with evidence e, H passes the test when it fits well with e, but it is important to

---

1 Actually 'use-novelty' is not the term these authors use to describe their position. For example, Worrall seems to prefer 'heuristic account' 'or 'Zahar-Worrall criterion' as the name of their position (Worrall 1989, 148-155). I use the terms 'use-novelty' and 'use-construction' simply because these are the terms Mayo uses in her analyses.
I do not have enough space to discuss other proposed notions of novelty and to compare the relative merits of these notions. See, for example, the exchange between Musgrave (1978) and Worrall (1978b) for more about this.

2 For example, Brush (1989) revealed a historical case (acceptance of the general theory of relativity) in which scientists preferred non-novel evidence (the advance of the perihelion of Mercury, which, as Earman 1992 shows, was used in the construction of the theory) to novel evidence (star-light bending by the sun). Should we admit that these scientists were irrational? Another attack comes from Bayesians who argue that use-novelty is not a necessary condition for good evidence from the Bayesian point of view (Howson and Urbach 1993, 408-409).

3 To find the probability, you need to calculate the corresponding areas of the binomial distribution.

note that two hypotheses which fit equally well with e can be totally different in terms of the severity of the test.

In her criticism of use-novelty requirement, Mayo first shows that logically speaking use-novelty and severity are independent. Then she proceeds to some concrete cases (presumably examples of good scientific reasoning) in which use-novelty seems violated while the severity criterion is met. So let us look at her conceptual argument first.

Through the analyses of Worrall's (1989) and Giere's (1983) arguments for use-novelty, Mayo finds that in both cases some kind of reliability or severity of the test is operating as the intuition behind the use-novelty requirement (Mayo 1996, 263-271). Mayo's strategy is to show that this background intuition does not necessarily demand use-novelty. Mayo summarizes Giere's reasons why use-novelty is required in the following way (268-269; I slightly changed the way the argument is summarized):

1. A successful fit does not count as a good test of a hypothesis if such a success is highly probable even if the hypothesis is incorrect. (That is, a test of H is poor if its severity is low.)

2. If H is use-constructed, then a successful fit is assured, no matter what. Therefore its success is high ("near unity") even if it is false.

The proposition 1 is virtually the same as Mayo's severity criterion, and naturally Mayo has no complaint with it. But, according to Mayo, the first sentence of 2 does not support the conclusion. The successful fit is assured "no matter what the data are," but this is not the same as "no matter if H is true or false," which is required to draw the conclusion (269).[4] More formally, Mayo distinguishes the following two probabilities (270-271; emphasis in original)

A. The probability that (use-constructed) test T passes the hypothesis it tests

B. The probability that (use-constructed) test T passes the hypothesis it tests, *even if it is false*.

---

4 Actually Giere does claim that the evidence used for the construction does not constitute a good test because the test passes the hypothesis "no matter whether the corresponding hypothesis is true or false" (Giere 1983, 286). This suggests that Mayo misunderstands Giere's argument, but maybe it is Giere who misstate his own position. Here, I assume that Mayo correctly characterize Giere for the sake of argument.

If the hypothesis is use-constructed, probability A is 1 (or at least close to 1). However, according to Mayo, probability B can be less than 1 (or even small) when probability A is 1. Since severity is judged by probability B, not probability A, a use-constructed hypothesis can be severely tested. Here we should note that it is provable from axioms of probability theory that if probability A is 1, probability B (taken as ordinary conditional probability) is also 1 unless the probability that the hypothesis is false is 0 (see Appendix). This sounds a very limited case, but as we will see soon, there is an implicit condition behind A (and not behind B) which helps Mayo to avoid this problem. But let us first look at a case that does illustrate the conceptual difference successfully.

The example Mayo uses is the average SAT score of the students in a given class (Mayo 1996, 271-272). When you sum up all the SAT scores of the students in the class and divide the sum by the number of the students, you obtain a number, say, 1121. From this, you can construct a hypothesis that the average SAT score of the students in the class is 1121. Here you used the data to construct the hypothesis, and therefore probability A, the probability that the hypothesis fits with the data about the individual scores, is 1. But probability B, namely the probability that the hypothesis passes the test given that the average is not 1121, does not have to be 1. Actually, quite plausibly, the probability is 0; for, if the average is not 1121, then the result of the calculation cannot be 1121, so the hypothesis never fits with the evidence. Thus, in Mayo's terms, this is a maximally severe test. This conclusion is acceptable even given the above-mentioned proof. The chosen hypothesis cannot be false in this situation, so probability B can be less than 1 even when probability A is 1.[5]

Therefore, Mayo is right in that use-novelty and severity are conceptually independent, and in some cases use-constructed hypotheses can actually be severely tested. But one question arises here: is this result widely applicable? In the average SAT score case, the hypothesis becomes true by virtue of the construction procedure. This is certainly a very exceptional case, and if the difference matters only in this kind of case, the distinction between use-novelty and severity is not a big deal in real scientific practice.

3. Two meanings of use construction

In this section, our argument concentrates on two cases: one in which the use-constructed hypothesis is

5 Exactly speaking, Probability B should remain undefined in this case. See Appendix.

intuitively admissible, and claimed to be tested severely; the other in which the use-constructed hypothesis is intuitively inadmissible and claimed to be not severely tested. What I am going to show is that these two cases are actually parallel with respect to the severity of the test. This implies that the severity criterion cannot account for the difference in the intuitive admissibility between these cases.

3-1. Confidence Interval

Mayo's second example of severe testing of a use-constructed hypothesis is the confidence interval (Mayo 1996, 272). Suppose that a random sample of the U.S. population is polled and the result shows that 45 percent of the sample approves of the President with a margin of error of 3 percent. We can construct a hypothesis from the data.

H(e): the population proportion of approval of the President is in the interval between 42 percent and 48 percent.

If the margin of error is set for 95 percent confidence, this result means that the probability that the true population proportion is inside this interval is 95 percent.[6]

This is a widely used method, and, intuitively speaking, this sounds admissible. At the same time, this is a case in which the most relevant evidence is used in the construction of the hypothesis. Then the question is whether the evidence consists a severe test. Given the above proof about probability A and probability B, the answer should be no, because in this case the probability that the tested hypothesis (namely the hypothesis that the population proportion is either lower than 42 percent or higher than 48 percent) is false is not 0, which necessitates that probability B be 1, given that probability A is 1. But, as I mentioned above, this is not what she means to say. Let her explain why this is a case of a severe test (273): "the fit with the resulting hypothesis (the interval estimate H(e)) is given by the specified margin of error δ, say 2 standard deviation units. It is rare for so good a fit with H(e) to occur unless the interval estimate H(e) is true (i.e. unless the population proportion really is within 2 standard deviation units of e). So the severity in passing H(e) is high."

---

6 Actually this is a bit inaccurate. Exactly speaking, a 95 percent confidence interval means that whatever the true population value is, the probability that the estimate produced with this procedure excludes the true value is less than 5 percent. See Mayo 1996, 273 n.14.

To understand her argument better, it is convenient to distinguish two senses of use-construction. Compare the following two hypotheses:

H1(e): The population proportion of approval of the President is in the interval between 42 percent and 48 percent.

H2(e): Whatever the confidence interval obtained through this procedure is, the population proportion of approval of the President is in that interval.

These hypotheses coincide in their empirical contents in the actual world, but it is easy to imagine a possible world in which they do not coincide. I would like to call H1(e) an *accidentally use-constructed* hypothesis, and H2(e) a *genuinely use-constructed* hypothesis. I call H2(e) genuinely use-constructed because its empirical content is essentially based on the data, and actually it is empirically vacuous before we gather the data. On the other hand, there is no such essential tie between the empirical content of H1(e) and the data. You can construct H1(e) and predict certain result even before you start to gather the data.[7] Of course, in the actual course of events, H1(e) *is* constructed using the data, and this warrant our calling it use-constructed, but this is quite accidental in the sense that if the outcome of the poll were not exactly 45 percent (and this is highly probable), we would not construct H1(e) from the data. This is why I call H1(e) accidentally use-constructed.

In terms of severity, H1(e) is severely tested while H2(e) is not. Mayo's above argument nicely applies to H1(e), so I don't think I have to repeat it here. The same argument doesn't apply to H2(e), because H2(e) changes its empirical content when the data changes. The condition that H2(e) is false, i.e. the true population proportion is not inside the interval obtained from the data, simply means that the distance between the true proportion and the sample proportion are more than two standard deviation units. Even under this condition, whatever the sample proportion is,

---

7 And there is even a systematic way to do so. If we know the size of the sample to be investigated,we can simply construct all possible hypotheses of the form H1(e) before the poll, and reject all other hypotheses after the result. For, given the sample size, there are only finite number of possible outcomes of the poll, and for each outcome, there is an associated estimate of the confidence interval.
This is an enough reason not to call H1(e) use-constructed for those who take a logical view of evidential relationship, but as Worrall (1978a, 1978b) and Mayo (1996, especially ch. 10) argue, we can not take the logical view for granted. Worrall's later writing (1989) seems to show an inclination toward more logicalistic understanding of use-novelty, though.

H2(e) obviously fits with the data because the empirical content of H2(e) is given by the very data. Thus the severity of the test of H2(e) is 0.

I think it is clear now that Mayo has H1(e) in mind when she says that the hypothesis is tested severely. But here is one problem. I do not think that probability A here, i.e. the probability that H1(e) passes the test, is 1, if it is taken literally (and this is expected from the proof in Appendix). Whatever the true population proportion is, the probability that the result of poll is exactly 45 percent is fairly low. As I understand, when the result is not 45 percent H1(e) doesn't pass the test, so probability A is also fairly low. There is one way Mayo can claim that probability A for H1(e) is 1, namely to add a condition to A:

A'. The probability that (use-constructed) test T passes the hypothesis it tests, *given the actual data*.

This condition makes probability A' for H1(e) unity. But she cannot add the same condition to probability B, because, as is clear from the above quote, her argument is based on the fact that the actual result of poll is unlikely under the condition that H1(e) is false. By confining the case to actual one, this low probability disappears.[8]

So now we have arrived at a consistent interpretation of Mayo's position. When use-novelty requirement is violated, probability A' (not probability A) is 1. But even in such cases, hypotheses like H1(e) can be severely tested in the sense that probability B (without assuming the actual evidence) for H1(e) can be low.

Before we move on to the next case, let me point out one counter intuitive aspect of the case. H1(e) and H2(e) are radically different in terms of the severity of the test, but do we really think that H1(e) is a much better hypothesis than H2(e)? Especially, in the real world they always coincide in their empirical contents given the testing procedure. Can Mayo give us an intuitively understandable reason why we should judge these hypothesis that differently?

---

8 This modified version is a departure from Mayo's own interpretation of Giere. According to Mayo, Giere objects to use-constructed hypotheses because they pass the test "no matter what the data are" (Mayo 1996, 269). This interpretation is clearly inconsistent with the modification, and this suggests that H1(e) is not the kind of use-constructed hypothesis Giere or other people had in mind. Again I postpone the question for the sake of argument. As for a suspicion about Mayo's interpretation itself, see the above note 4.

3-2. Gellerization

When we interpret Mayo in this way, however, a further problem occurs. Many other intuitively inadmissible use-construction procedures can meet this requirement! Let me show this using what Mayo calls 'gellerization' as an example.

Gellerization is a typical erroneous way of creating "maximally likely alternatives" (Mayo 1996, 200). A maximally likely alternative is "one constructed after the fact to perfectly fit the data in hand" (ibid.). More precisely, evidence e makes hypothesis H maximally likely when P(e|H) = 1. For example, when we do curve-fitting with some sample points as data, a curve drawn deliberately to connect all the points fits the data perfectly. This is a case of use-construction, and we are not impressed with most of these alternatives, but the question is whether Mayo's severity criterion (under the interpretation in the previous section) can be used to rule out these maximally likely alternatives.

Let us take her own example of a gellerized hypothesis test to illustrate her point. Her example is coin-tossing trials (201-202). Suppose that we are testing a hypothesis H_0 that a certain coin is fair. Suppose also that we obtain the results of H,T,T,H from four trials. We can gellerize a hypothesis that the probability of heads is 1 on trials one and four, 0 on two and three. Of course this hypothesis is maximally likely. In general, we can construct a maximally likely alternative hypothesis G(e) for any result (202):

G(e): the probability of heads equals 1 on just those trials that result in heads, 0 on the others.

Is this not a better hypothesis than H_0? According to Mayo, it is not. Let us think about the following test procedure (202): "Fail (null) hypothesis H0 and pass the maximally likely hypothesis G(e) on the basis of data e." Mayo claims: "There is no probability that test T would *not* pass G(e), even if G(e) were false. Hence the severity is minimal (i.e., 0)" (ibid.; emphasis in original). Thus, even though G(e) fits better with e, H0 is a better hypothesis in the sense that H0 is tested more severely than G(e).

So, Mayo arrives at a radically different conclusion from the case of the confidence interval. But are they that different in terms of severity? To clarify the point, let us distinguish two hypotheses again:

G1(e): The probability of heads is 1 on trials one and four, 0 on two and three.

G2(e): Whatever the result of the trials are, the probability of heads is 1 on trials that result in heads, 0 on the others.

Again, G1(e) is accidentally use-constructed and G2(e) is genuinely use-constructed. Mayo's argument applies to G2(e) fairly well, and her argument is basically the same as my above argument that H2(e) is not severely tested. But what about G1(e)? It seems to me that G1(e) is tested severely. Suppose that G1(e) is false and H0 is true.[9] Then the probability that G1(e) passes the test is 1/16 (roughly .06), because G1(e) passes the test only when the obtained sequence is H,T,T,H. Thus probability B for G1(e) is fairly low, which makes severity fairly high. Though I do not get into details, probability A for G1(e) is fairly low, and probability A' for G1(e) is unity.

Thus, the situation is parallel between G1(e) and H1(e), and G2(e) and H2(e). The parallelism may break down if we have some reasons to believe that those who use confidence interval always test accidentally use-constructed hypotheses, and those who does gellerization always test genuinely use-constructed hypotheses. But I do not think we can assume this, especially because these hypotheses (accidentally use-constructed ones and genuinely use-constructed ones) are empirically equivalent in the actual world.

Is there any other way to justify Mayo's discrimination between two cases? Maybe Mayo maintains that H1(e) and G1(e) are severely tested and H2(e) and G2(e) are not, while she forgot to mention G1(e) and H2(e) in the text. Even if we take this interpretation, there is still an intuitive objection. It looks intuitively obvious that G1(e) and

---

9 I need to explain what I am doing here. Obviously there are many other alternative hypotheses of the form "probability of heads is p ($\neq$.5)," or even other possible maximally likely hypotheses like "The probability of heads is 1 on trials one and two, 0 on three and four." How can we calculate probability B when "G1(e) is false" means that one of these alternative is true? This is particularly problematic for Mayo because she rejects assigning prior probabilities to hypotheses (Mayo 1996, 75). Mayo is answer is that her severity criterion "requires that the severity be high against *each such alternative*" (195; emphasis in original). This sounds too strict a requirement, but she does not mean all logically possible alternatives. As her argument in chapter 5 of the book shows (and as she pointed out to me in her private communication), it is enough for her if severity is high against all relevant alternatives specified by the research design. Therefore, since the question here is the comparison of H0 and G1(e), I think I am warranted to concentrate on H0 where the severity criterion (formally) requires to consider all other alternatives.
By the way, if Mayo does not add this qualification about the relevant alternatives, then it can be shown that almost alway there is at least one maximally likely alternative which is incompatible with the tested hypothesis and still assures high probability of passing result of the hypothesis (though I do not have enough space to prove it here). This makes the severity criterion almost vacuous.

G2(e) are equally bad hypotheses. If Mayo's severity criterion judges them with such a radical difference, maybe we should question the validity of the severity criterion itself. Or, maybe she can find a different interpretation of severity so that H1(e) and H2(e) are severely tested while G1(e) and G2(e) are not, but this does not sound very promising given the strong parallelism between these two cases.

4. Concluding Remarks

My assessment of Mayo's argument is mixed. First, she is successful in showing that use-novelty and severity are conceptually independent, and she also succeeds to give us a case in which a use-constructed hypothesis is indeed severely tested by the same evidence (where the procedure assures the correctness of the hypothesis). Second, she shows that some accidentally use-constructed (and, intuitively speaking, admissible) hypotheses can be severely tested. On the other hand, thirdly, there is a suspicion that accidentally use-constructed hypotheses are not what those who endorse use-novelty have in mind when they blame use construction (see notes 4 and 8 above). Moreover, there are certain accidentally use-constructed hypotheses which seem (intuitively speaking) bad hypotheses and still are tested severely.

If my assessment is valid and I can rely on my intuition (which is hopefully shared with the reader) about good hypotheses, then it seems to me that neither severity nor use-novelty provide a satisfactory account of a good test.[10] Then, how can we account for the difference between the confidence interval and gellerization? Let me conclude this paper with one attempt of such an account. In the case of confidence interval, the class of all hypotheses that can be accidentally use-constructed (the class of all hypotheses of the form "The population proportion of approval of the President is in the interval between x percent and y percent" with concrete numbers in x and y where y-x=6) is exhaustive in the sense that at least one of the hypotheses is true. Thus, simply eliminating implausible alternatives can lead us to a reliable hypothesis. The same thing does not apply to gellerization. As Mayo recognizes, the test procedure rejects H0 come what may (Mayo 1996, 202) and this makes the class of possible accidentally

---

10 The extension of this conclusion to other cases --- like Mayo's "hunting" (Mayo 1996, ch 9), various historical cases of ad hoc theory changes (Howson and Urbach 1993, 147-149 has a nice list), or successful scientific cases like Fresnel's wave theory of light (Giere 1983; Worrall 1989), the special theory of relativity (Zahar 1973), or the general theory of relativity (Brush 1989; Earman 1992) -- is an interesting topic, but obviously I do not have space here.

use-constructed hypotheses non-exhaustive. I would like to call this feature of gellerization a *systematic neglect of relevantalternatives*. I think this may be a key to refine the notions of severity and use-novelty: namely, a theory construction and testing procedure is not justified if the procedure involves a systematic neglect of relevant alternatives. This may be represented by an unfairly low probability of accepting a member of certain class of relevant alternatives given any possible outcome. What is the relationship between this rule and the other notions? What exactly means to say "unfairly low"? And how general is this rule applicable? These are questions yet to be answered through further analyses of the above and other cases.

Appendix

Here is the proof that probability B is 1 if probability A is 1 and probability that H is false is not zero. Let us call the event "(use-constructed) test T passes the hypothesis it tests" PASS, and call the tested hypothesis H.

Then

(1) $P(H| B) + P(\sim H| B) =1$

where B stands for information about background settings, including the experimental design, decision rules, theory construction procedures etc.

Also, since H and ~H are mutually exclusive and exhaustive,

(2) $P(PASS|B)=P(PASS\&(H\vee\sim H)| B) = P((PASS\&H)\vee(PASS\&\sim H)| B)$

$=P(H| B)*P(PASS|H\&B)+ P(\sim H| B)*P(PASS|\sim H\&B)$

Here, the left hand side, P(PASS|B), stands for Mayo's probability A, thus $P(PASS|B)=1$.

As for the right hand side, since $0\leq P(PASS|H\&B)\leq 1$ and $0\leq P(PASS|\sim H\&B)\leq 1$,

$0\leq P(H| B)*P(PASS|H\&B)\leq P(H| B)$ and $0\leq P(\sim H| B)*P(PASS|\sim H\&B)\leq P(\sim H| B)$

Therefore

(3) $P(H| B)*P(PASS|H\&B)+ P(\sim H| B)*P(PASS|\sim H\&B)\leq P(H| B) + P(\sim H| B)\ (=1)$

Assuming that neither $P(H| B)$ nor $P(\sim H| B)$ is 0, the left side and the right side of (3) are equal only when both $P(PASS|H\&B)$ and $P(PASS|\sim H\&B)$ are 1, and $P(PASS|\sim H\&B)$ stands for Mayo's probability B. Thus, if probability

A is 1, then by virtue of (2), probability B should be also 1. However, if P(~H| B) is 0, i.e. if the probability of H's falsity given the background setting is 0, then the value of P(PASS|~H&B) does not matter for the equality. Exactly speaking, if P(H| B) is zero, then P(PASS| H&B) should remain undefined, because that makes the denominator of the following equation 0:

P(PASS| H&B) =P(PASS&H| B)/P(H| B)

Please note that I have to assume little about the interpretation of probability. As far as the axioms of probability theory are accepted, the interpretation of these probabilities (subjectivist, frequentist or physical) matters little for the validity of this argument. For example, P(H| B) can be a frequentist probability that the given theory construction procedure yields a true hypothesis, rather than the degree of belief that H is true given information B.

References

Brush, S.G. (1989), "Prediction and Theory Evaluation: The Case of Light Bending", Science 246, 1124-1129.

Earman, J. (1992), Bayes or Bust?: A Critical Examination of Bayesian Confirmation Theory. Cambridge, Mass: The MIT Press.

Giere, R. (1983), "Testing theoretical hypotheses", in John Earman (ed.), Testing Scientific Theories, Minnesota Studies in the Philosophy of Science vol. X. Minneapolis: University of Minnesota Press, pp. 269-298.

Howson, C. and P. Urbach (1993), Scientific Reasoning: The Bayesian Approach, second edition. La Salle, IL: Open Court.

Mayo, D. G. (1996), Error and the Growth of Experimental Knowledge. Chicago: The University of Chicago Press.

Musgrave, A. (1978), "Evidential Support, Falsification, Heuristics, and Anarchism", in G. Radnitzky and G. Andersson (eds.), Progress and Rationality in Science, Boston Studies in the Philosophy of Science vol. LVIII. Dordrecht: Reidel, pp.181-201.

Popper, K. (1962), Conjectures and Refutations: The Growth of Scientific Knowledge. New York: Basic Books.

Worrall, J. (1978a), "The Ways in Which the Methodology of Scientific Research Programmes Improves on Popper's Methodology", in G. Radnitzky and G. Andersson (eds.), Progress and Rationality in Science, Boston Studies in the Philosophy of Science vol. LVIII. Dordrecht: Reidel, pp. 45-70.

--. (1978b), "Research Programmes, Empirical Support, and the Duhem Problem: Replies to Criticism", in G. Radnitzky and G. Andersson (eds.), Progress and Rationality in Science, Boston Studies in the Philosophy of Science vol. LVIII. Dordrecht: Reidel, pp. 321-338.

--. (1989), "Fresnel, Poisson and the White Spot: The Role of Successful Predictions in the Acceptance of Scientific Theories", in D. Gooding, T. Pinch and S. Schaffer (eds.), The Use of Experiment: Studies in the Natural Sciences. Cambridge: Cambridge University Press, pp.135-157.

Zahar, E.G. (1973), "Why Did Einstein's Programme Supersede Lorentz's?", The British Journal for the Philosophy of Science 24: 93-123 and 223-262.