Toward a More Objective Understanding of the Evidence of Carcinogenic Risk[1]

Deborah G. Mayo

Virginia Polytechnic Institute and State University

The field of quantified risk assessment is a new field, only about 20 years old, and already it is considered to be in a crisis. As Funtowicz and J.R. Ravetz (1985) put it:

> The concept of risk in terms of probability has proved to be so elusive, and statistical inference so problematic, that many experts in the field have recently either lost hope of finding a scientific solution or lost faith in Risk Analysis as a tool for decisionmaking. (p.219)

> Thus the 'art' of the assessment of risks ... is at an impasse. The early hopes that it could be reduced to a science are frustrated. ...[O]thers are tending to introduce the 'human' and 'cultural' factors. The question now becomes, to what extent should these predominate? Would it be to the reduction or exclusion of the 'scientific' aspects? For, ...if the perceived phenomena of 'risks' are interpreted as lacking all objective content or being merely a small part of some total cultural configuration, then there is no basis for dialogue between opposed positions on such problems. (pp.220-221)

The crisis of confidence in this new field comes from two directions: on the one hand it comes from the general challenge of philosophers and others as to whether there exist any objective, rational rules in science; and on the other hand there are many real cases where conflicting risk assessments are reached on the basis of the same data. It will be useful to consider throughout an example of such a risk assessment conflict. I take a recent case from the Environmental Protection Agency (EPA) as to the carcinogenic potential of the substance formaldehyde. On the basis of the very same data, the EPA in May of 1981 reached a different and opposite assessment from the one it had reached in September of 1981. My aim is to suggest how a more objective understanding of the evidence would help in resolving such a conflict.

I want to emphasize at the start that my approach is distinct from those appeals to "objective science" that deny the entry of value judgments in reaching risk assessments. Rather, my approach will be to show that despite the entry of these value judgments, it is possible to unearth what the data do and do not say about the actual extent of the risk involved.

1. Risk Assessment in the Case of Formaldehyde

The term risk assessment, as I am using it, covers the generation and analysis of data in order to characterize the extent to which an agent causes an increase in the incidence of a

---

PSA 1988, Volume 2, pp. 489-503

health condition, in our case, cancer, in humans, lab animals or other test systems. Data generation in arriving at risk assessments includes prospective randomized treatment-control experiments, and retrospective case-control studies. To examine whether formaldehyde increases the risk of cancer, prospective experiments on rats were conducted. Epidemiological studies on humans, in contrast, only allowed a retrospective analysis of cancer rates in various occupations. On the basis of the statistically significant increases in (nasal) cancer among formaldehyde-treated rats, the Chemical Industry Institute of Toxicology (CIIT) reached the assessment that formaldehyde is carcinogenic in laboratory rats and reported this to the EPA in November of 1980. A panel of eminent scientists convened by the National Toxicology Program confirmed this risk assessment, and concluded that "formaldehyde should be presumed to pose a risk of cancer to humans". (See *Hearing*[2], p. 191.) The lengthy document detailing the formaldehyde risk assessment was entitled the "Priority Review Level 1" (PRL-1) dated February, 1981.

Risk assessments form the basis of *risk evaluations* and *management*. These involve assessing how substantively important a risk is and what should be done about it.[3] This requires an explicit consideration of social, ethical, and economic considerations (e.g., against what level of risk should the public be protected? and what form should this control take?) The evaluation the EPA staff reached was that formaldehyde should be designated as a priority chemical under the EPA provision known as 4(f). To quote from the *Federal Register* notice:

> [T]he Agency finds that there may be a reasonable basis to conclude that formaldehyde presents a significant risk of widespread harm to humans from cancer. (Federal Register, May 1981, pp.5-6.)

This last sentence is important because triggering statute 4(f) requires only that *there may be a reasonable basis* to conclude that a significant risk exists, and not that there is a reasonable basis for such a conclusion. In itself 4(f) does not call for any regulation. It is simply a call for closer scrutiny based on an indication that there *may* be a significant cancer risk.

All of this was in 1980 and early 1981. Then there was a change in Administration; the Reagan administration entered, and along with it a new EPA Administrator (Ms. Gorsuch) and some new staff. In fact, formaldehyde was the first 4(f) recommendation brought before the new Administrator for signing. Instead of signing it members of the new EPA staff carried out a reassessment of the hazard data in the PRL-1. The new and revised version of the data became the Todhunter Memorandum, Dr. Todhunter being a new EPA Assistant Administrator. Some of the changes included blatant erasures of the highest risk estimates that had been given in the PRL-1. (See *Hearing*, pp. 349-365.) There are other, less blatant changes in the reassessment. Most significant was Todhunter's deemphasis of the positive rat studies and his emphasis of the negative epidemiological studies. Todhunter concludes, for example,

> There does not appear to be any relationship, based on the existing data base on humans, between exposure [to formaldehyde] and cancer. Real human risk could be considered to be low on such a basis. *(Hearing,* p. 260)

This hazard reassessment was then given as the basis of a changed hazard evaluation. On Sept. 11, 1981 the EPA staff recommendation to designate formaldehyde as a 4(f) priority chemical was *reversed* and the opposite hazard evaluation was made. (See *Hearing*, pp.192-193.) Whether or not this shift in hazard assessment was justified was the subject of enormous controversy. It led to a congressional hearing on formaldehyde, which I refer to throughout as *Hearing*.[4]

The suspicion which led to these hearings was that the agency was altering the widely accepted standards for carcinogenic risk assessment, and that it was doing so in order to

garner "expert scientific support" for furthering the aim of anti-regulation. One of the main reasons for this suspicion was that the new Administration did not base its decision against a 4(f) designation on any new data beyond the PRL-1 document which had been the basis for the original, and opposite, recommendation (though, as mentioned, it did *conceal* some evidence supporting a 4(f) designation). Rather, the new Administration proceeded to hold a series of secret meetings restricted only to certain scientists and lawyers from the Formaldehyde Institute, the Formaldehyde Trade Association and EPA staff. In these meetings they reinterpreted the data and, without the usual peer review, came to the opposite conclusion than that endorsed by numerous eminent scientists and agencies. As one attorney with the Natural Resources Defense Council (Dr. Warren) put it:

> There are no new data to support the reversal, only a reinterpretation which has been advocated by and is quite favorable to the interests of the formaldehyde industry. Those new assumptions, as we have heard, depart radically from accepted principles of cancer risk assessment... . In our view, this has been an effort to get the Government off the back of the formaldehyde industry. (*Hearing*, p.188)

The disagreement was not about the level of risk required before triggering 4(f) i.e., for judging that a substance may pose a significant or serious human risk. Todhunter maintains that he was holding the same range of risk which agencies have tended to deem of public concern.[5] Nevertheless, on the same evidence, different conclusions are reached. If going from data to risk assessments was a matter of applying a single universally accepted best method, then this difference in resulting assessment could not occur. That disagreement does occur shows there is no such algorithm for risk assessment. The fact that assessments are nevertheless reached is typically taken as grounds to conclude that extra-scientific, cultural, social, ethical or other contextual values must be entering. For some, this conclusion shows that something is wrong and that we need to try to avoid or somehow neutralize the entry of non-scientific interests. In the formaldehyde case—which came at a time during which such politicization at the EPA was rampant—this attitude resulted in the above mentioned hearing to determine if the EPA was altering the standards for carcinogenic risk assessment. Later, the National Academy of Sciences issued a report stressing the need to separate the science of assessment from the social and policy values that enter at the level of risk management (National Research Council). But a growing body of risk literature questions the possibility that scientific risk assessment can ever be free of the policy values appropriate to risk management. To this group, policy in science is not a violation, but rather is inevitable. However, there are very different grounds for reaching such a view, and it is important to separate them.

2. The Sociological Relativist View

On one set of views, which I will call the *sociological relativist view*, scientific risk assessment—indeed science generally—is inevitably a product of, if not entirely constructed from—socio-cultural values. An example is the influential socio-cultural theory of risk assessment of Douglas and Waldavsky. According to Douglas and Waldavsky:

> The risk assessors offer an objective analysis. We know that it is not objective so long as they are dealing with uncertainties and operating on big guesses. They slide their personal bias into the calculations unobserved. (Douglas and Waldavsky 1980, p.80)

Risk assessment, on their view, is totally determined by socially constructed methods and judgments; they are social constructs. This view provides an explanation of conflicting risk assessments in terms of the different policy judgments and competing 'world-views' of different assessors.

Granted one can find competing political interests to explain the conflicting risk assessments in the case of formaldehyde—as searching the fascinating testimony shows.

The EPA and Todhunter were influenced by the political commitment to anti-regulation, and one can explain the Todhunter Memo, as one attorney argued and as we have already noticed, as "an effort to get the Government off the back of the formaldehyde industry" (*Hearing*, p.188). Likewise, in defense of Todhunter one witness (Mr. Walker) maintained that the opposition were "a few disgruntled employees of the EPA who [simply because they want to place formaldehyde under 4(f)] feel justified in waging guerrilla warfare against the Agency and those in positions of authority" (*Hearing*, p.4).

But however well a story about background interests, social values, and negotiation may explain a risk assessment, and however much our assessment tools are products of social beings and institutions, the question whether or not a risk assessment is warranted is not a matter of social values; it is a matter of what the risk actually is. Sociological relativists are led to consider "objective" physical risks either unattainable or unimportant. However, they hold an overly stringent conception of objectivity—one that is precluded by the need to make inferences under uncertainty without algorithms. Such relativists consider the entry of any and all judgments subjective and biased; as we saw in the quote from Douglas and Waldavsky.

3. Risk Assessment Policy (RAP) and RAP Relativism

While the sociological challenge to objectivity may be countered by denying that objectivity requires neutral algorithms or freedom from uncertainty, there is a different basis for challenging the objectivity of risk assessments—one which does not turn on an overly narrow conception of objectivity and does not deny the importance or possibility of measuring physical risks. This second view stems from the nature of the judgments and decisions that are required in order to carry out risk assessment estimates. For example, one must decide what data to collect, how large a sample size to take, what levels of reliability (e.g., statistical significance) to use, how to weigh studies with different results (e.g. whether they should be weighed according to statistical power), what models should be used to extrapolate from animals to humans, etc. Because these judgments involve choices with no unequivocal scientific answers—at least at present—and since these choices have policy implications, they are intertwined with policy. Thus they are not just a matter of objective science. These judgments may be called risk assessment policy (RAP) judgments.

The view that risk assessment is necessarily entwined with policy because of the inevitability of making RAP judgments may be called *RAP relativism*—to distinguish it from other more extreme relativistic views (e.g., sociological relativism). Among the first to articulate a version of RAP relativism was Alvin Weinberg, who placed what I am calling RAP judgments under his term *trans-science*—"questions which can be asked of science and yet *which cannot be answered by science*" (1972, p.209). Now the view is fairly widespread. A congressional report from the National Academy of Science in 1983, which resulted from the suspicion that agency science was being politicized (e.g., as represented by the formaldehyde case), offers a very useful delineation of over 50 junctures at which RAP judgments enter in the course of making risk assessments.[6] Carl Cranor has recently provided a clear statement of what I am calling RAP relativism in regards to the judgments required in specifying methods of risk estimation:

> [T]he supposedly objective scientific studies used for estimating risks to human health...are considerably more controversial and political than most people think. ...a wise and conscientious scientist with perfect test data *cannot* help but make moral and policy judgments in order to interpret an epidemiology study and to produce the risk numbers that are the outcome... . ...the moral and policy judgments are forced by the statistical equations themselves and the choice of variables employed in them.[7] (Cranor 1987)

The basis for Cranor's allegation is that each RAP choice has policy implications. That is, each choice influences the chance that the substance will be considered a significant risk to humans; in other words, it affects the protectiveness of the risk assessment.

The conflicting assessments in the formaldehyde case stem from different choices of RAP options. In particular, while the CIIT report and the PRL-1 accorded high weight to the positive animal results and denied that negative epidemiological studies warranted concluding no increased human risk, the Todhunter Memo did just the reverse. The Todhunter Memo deemphasized the positive rat studies and—most importantly—took the negative epidemiological studies to indicate low increased human risk or none.[8] And the reason a result was considered negative was itself a result of a particular choice of statistical analysis—also a RAP judgment. With an EPA purged of scientists save those tending to favor anti-regulation, there was plenty of leeway for the Agency to consistently choose the inference option least likely to have a protective outcome.[9] Under the guise of demanding stringent scientific evidence, these policy choices made it extremely unlikely that a substance would be claimed to pose a significant human risk. In contrast, those who endorsed the original PRL-1 made a more protective RAP choice. On the RAP relativist account, therefore, the conflicting risk assessment results from the difference in the view of those concerned as to how protective regulations should be; it is a policy conflict.

Although RAP relativism is less threatening to objectivity than sociological relativism and sociological reductionism, it shares some of the same implications for risk assessment. First, it implies that risk assessment judgments (at least where there is uncertainty) inevitably reflect policy judgments, and risk assessment disagreements largely reflect disagreements about policy —including moral, social, economic or other values typically considered "non-scientific". Thus, science is given little role in an unbiased adjudication of disagreements over risk assessments. If interpreting scientific results is necessarily colored by social and political contexts, it is impossible for science to provide risk assessment oversight that is fully objective.

Second, if it is true that in reaching a risk assessment (based on RAP judgments) one cannot help making an ethical choice about how protective one should be, then it does not seem that risk assessment is an appropriate business for scientists. After all, scientists are not elected to make social policy choices about acceptable risk. (It was precisely the fact that EPA scientists were guilty of politically motivated assessments that led many to decry the politicization of EPA science during the time of the formaldehyde reassessment.) If risk assessment judgments are policy judgments, then dealing with the uncertainties involved in RAP choices should be performed by policy makers and ethicists, not scientists.

However, such a practice allows policymakers to fall into all manner of misinterpretations of the assessment evidence. (See, for example, Silbergeld.) If RAP judgments are made by non-scientist policymakers, they are divorced from the original issues and uncertainties underlying the different risk estimates. At the same time, the scientist is limited to presenting possible RAP choices, but is involved neither in making them nor in bringing out the implications for protectiveness. For example, if the scientific work ends after reporting two possible estimates that may be used, say, a maximum likelihood estimate and an upper 95% confidence bound, then the scientist will not be around to explain how far off each of these estimates is likely to be from the actual risk and why. This permits the assessor to make the final choice (e.g.,about which estimate to use) without acknowledging what standard of protectiveness he is effectively requiring in choosing a given option (e.g., that the maximum likelihood estimate is less protective than the upper 95% confidence bound).

Without investigating further the consequences of RAP relativism here,[10] I want to consider whether the need to make decisions in applying statistical risk assessment methods really does have this relativistic consequence—i.e., the consequence that interpreting results necessarily requires policy judgments. I shall argue that this conclusion may be avoided.

It may be admitted that the conflicting assessment of formaldehyde is explainable by different (more and less protective) choices of RAP options (i.e., whether or not to accord higher weight to the positive animal results than to the negative epidemiological studies). But this would not lead to RAP relativism unless adjudicating between them is itself relative to a stance on how protective we should be. I shall argue it is not so relative. Granted, there will be latitude for choice among possible RAP options. Granted also that each choice has a policy implication—influencing the likelihood that a substance will be judged to pose a significant hazard to human health. Nevertheless, this does not preclude objective scrutiny as to whether a given assessment is warranted by the evidence. Whether given evidence warrants a given risk estimate is a matter of scientific not social acceptability—that is, it is a question of how well the inferred estimate reflects what is really the case about the causal effect of the substance in question. The latitude in choosing RAP options does not preclude the objective scrutiny needed to answer this question.

Focusing on the RAP judgments involved in interpreting statistical tests, I shall argue that risk estimates are necessarily policy judgments only under misuses of the statistical tests involved—ones which, unfortunately, are encouraged by the manner in which tests are often formulated and taught.

4. Neyman-Pearson (NP) Tests

The type of statistical test standardly used in reaching risk assessments is the Orthodox or Neyman-Pearson test (NP test), often in combination with Fisherian significance tests. The test considers hypotheses, typically assertions about a property of some population: a *parameter*. In the formaldehyde case, the hypotheses are assertions about the parameter which I will call Δ, the increased cancer risk in the population—humans. The NP test splits the possible parameter values into two: one representing the *test (or null) hypothesis* H, the other the set of *alternative hypotheses* J. The test hypothesis H in the formaldehyde case asserts that formaldehyde does not cause an increase in a person's risk of dying from cancer of a give type. That is, it asserts that there is a 0-increase in the hazard rate, i.e., $\Delta = 0$. The alternative hypotheses assert that formaldehyde causes a positive increase, i.e., $\Delta > 0$. Since here one is looking for positive discrepancies from 0, this is a *one-sided* test, which I call test T+:

Test T +: Test (null) hypothesis H asserts $\Delta = 0$ (no increased risk)[11]
Alternative hypothesis J asserts $\Delta > 0$ (a positive increased risk).

The test considers a *test statistic* that describes an aspect of the outcome of interest. One statistic in testing formaldehyde is D, the difference in cancer rates between the subjects exposed and those unexposed to formaldehyde:

Test Statistic D: the difference in cancer rates between the subjects exposed and those unexposed to formaldehyde.

Corresponding to each observed difference is its level of statistical significance, defined as follows:

The Statistical Significance Level of an Observed Difference D_{obs} is the probability with which so large a difference arises assuming the null hypothesis H is true, i.e., Prob($D \geq D_{obs}$, given that H is true).

A good way to see significance levels is as standard measures of distance from H, except with this inversion: the larger (and more significant) the distance, the smaller its significance level.

An NP test consists of a rule which specifies, before the observation is made, how statistically significant (i.e., how improbably far) an observed difference must be before it should be taken to reject H. The maximum significance level chosen beyond which D_{obs} is taken to reject H is called the *size* of the test, and is denoted by α. Thus, test T+ with size α consists of the following rule:

Test T+ with size α: Reject H if and only if observed difference D_{obs} is statistically significant at level α.

Observed differences which are not large enough to reach this preset size are taken to accept H. In this way the test maps the possible outcomes—the *sample space*—into either reject H (and accept J) or accept H. The partitioning that results from the test is illustrated

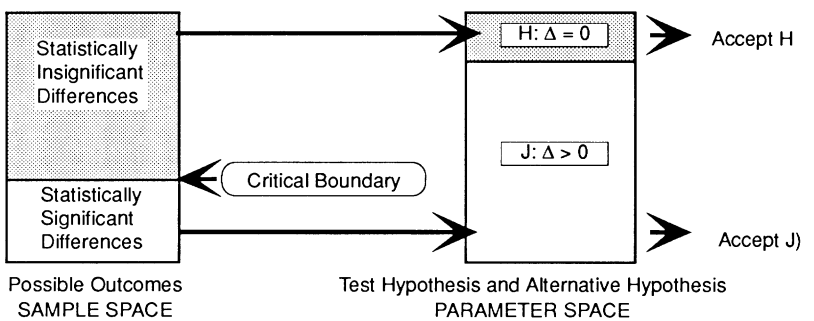

Figure 1. Neyman-Pearson Test T+ as a Mapping rule

below:

As long as there is variability in the effect (e.g., not all who are exposed get cancer, and not all who get cancer are exposed), and as long as only a sample from the population is observed, there is a chance the test will make an error. Two types of errors are considered: First, the test leads to reject H (accept J) even though H is true (the Type I error); second, the test leads to accept H although H is false (the Type II error). A test with size α rejects H when in fact H is true—i.e., it commits a Type I error—with probability no more than α. The smaller the test's size α, the less frequent the Type I error. But by making α smaller the test suffers an increase in the frequency with which it accepts H even when in fact H is false (and so should be rejected)—i.e., an increase in the frequency of a Type II error. The probability of a Type II error is denoted by ß. α and ß are the test's error probabilities:

Error Probabilities: α is the probability of an erroneous rejection of H (Type I error); ß, the probability of an erroneous acceptance of H (Type II error).

Since these two error probabilities cannot be simultaneously minimized, the NP model instructs one to first fix α, the size of the test, at some small number, such as .05 or .01. (In other words, the test is specified so as to ensure it is very improbable for the test to reject H when the hypothesis H is true.) One then seeks out the test which at the same time has a small ß. 1 - ß is the corresponding power of the test. Because in our case alternative hypothesis J contains more than a single value of the parameter, i.e, it is composite, the value of ß varies according to which alternative in J is assumed true. The "best" NP test of a given size α (if it exists) is the one which at the same time minimizes the value of ß (i.e., the probability of type II errors) for all possible values of Δ under the alternative J. I shall refer to a specific simple alternative hypothesis as $J{:}\Delta = \Delta'$.

We have now to ask: Are test specifications matters of policy values?

The test's error probabilities, α and ß, are objective in that they refer, not to subjective degrees of belief, but to relative frequencies of error in sequences of applications (whether similar or dissimilar) of a given experimental test.[12] And NP tests are objective in the sense that they control the error probabilities of tests regardless of what the true, but unknown, value of Δ is. However, the error probability specifications themselves go beyond the objective formalism of NP tests. As Neyman and Pearson note:

> From the point of view of mathematical theory all that we can do is to show how the risk of the errors may be controlled and minimized. The use of these statistical tools in any given case, in determining just how the balance should be struck, must be left to the investigator. (Neyman and Pearson 1933, p.146)

As a result, NP tests were developed in a certain decision theoretic framework where there would be a clear basis for the test specifications. Neyman called the resulting theory of tests an objective theory of *inductive behavior*. Here tests are formulated as mechanical rules or "recipes" for reaching one of two possible decisions: "act as if H were true" or "act as if H were false", according to whether H is accepted or rejected. Such "machinery" produces automatic acceptance or rejection of H. For example, rejecting H in our formaldehyde case may be associated with a decision to trigger 4(f). For each decision there are certain losses and costs associated with acting on it when in fact H is true. By considering such consequences the scientist is, presumably, able to specify the risks he can "afford". (It is imagined that the scientist first specifies α as the maximum frequency with which he feels he can afford to reject H erroneously, and then seeks to minimize the value of ß.)

However, this opens the door to the RAP relativist's concern. For such considerations of consequences—in our case social, ethical,and economic—are clearly policy matters; so it appears that specifying a test *is* tantamount to making a policy decision—just as the RAP relativists contend. But if the domain of a scientist is objectively finding out what is the case as opposed to setting policy goals, then he does not seem to be in the position of making the needed test specifications. And if he is left to make these value judgments, the results necessarily reflect, not just what is the case about the cause, but what he believes about when a substance ought to be regulated—e.g., placed under 4(f)—and that is precisely the RAP relativist allegation. If this is true, then it is impossible for an assessment to be wholly objective—where by "objective" I mean reflecting what is the case about the risk and not one's policy preferences.

How are we to avoid this conclusion and with it the charge that risk assessors necessarily make ethical and policy decisions in reaching and interpreting risk estimates? The answer is to be found in rejecting the automatic use of tests where the null hypothesis is accepted or rejected according to whether the preset significance level is reached, without any reflection on the evidential meaning of the specific observed result.

It is worth noting that the threat to objectivity caused by the automatic use of NP tests has been recognized since the tests were first advanced in the 1930's—most notably by R.A. Fisher. While Fisher's ideas formed the basis of NP tests, by couching them in a behavioral-decision framework he felt Neyman and Pearson had given up the ideal of objectivity in science. Although Egon Pearson, one of the co-founders of NP tests, responded to Fisher, as did others who reject the automatic use of statistical tests,[13] these automatic uses and the misinterpretations they encourage are still problematic enough in epidemiology to have given rise to a recent movement in that field to "cleanse its literature" of statistical tests and α values altogether. (Fleiss 1986, p.559. See also Walker 1986.) The misinterpretations which have led to this are: automatically equating rejections of H (statistically significant differences) with finding substantively important dis-

crepancies from H, and failures to reject H with finding 0 or unimportant discrepancies. However, the alternative methods recommended (confidence intervals) are open to analogous automatic uses. To avoid misinterpretations we need a more objective understanding of the statistical results, which I shall now consider.

4. A More Objective Understanding of NP Statistical Tests

a. Understanding Rejections of H:

Ideally, the policy question of what counts as a substantively important increase in cancer risk is answered at the start, so that the test may be specified in order to have appropriately high probabilities of detecting all and only those increases. Substantively important increases in the formaldehyde case are those increases in cancer risk deemed serious enough to trigger 4(f)—a policy judgment. However, regardless of how the test has been specified, whether based on policy or other values, knowledge of the test's error probabilities, I claim, allows interpreting objectively what the data do and do not indicate about the increased risk—i.e, about Δ.

What is the objective import of a rejection of hypothesis H (a positive result) in the context we are considering? The proper construal of a rejection of our test hypothesis H of 0-increase is an assertion to the effect that: "This test detected an increase Δ of at least such and such". The task is to determine the approximate *lower bound* for Δ.

To accomplish this, consider how one interprets a failing test score on an (academic or physical) exam. If it is known that such a score frequently arises when a subject's knowledge is deficient only to a given degree, say δ, then one would deny that such a rejection indicated the existence of a deficiency *in excess of* δ. For example, suppose a student obtains a failing score on an examination. But suppose such a failing score arises very frequently among students who have a deficiency δ of only 10% of the material being tested (i.e.,they know 90% of the material). Then such a failing score is not good grounds to infer that the student has a deficiency, say, of 40% (i.e., that he or she knows only 60% of the material). Such a test is *too severe* for that inference. The situation is analogous in interpreting statistical tests, and the reasoning can be made precise in what we may term the *severity function.*

Let us focus on the test T+ used in our formaldehyde example. The test result is a difference in risk rates, D_{obs}. For any hypothesized value of the increased risk one can ask: what is the probability of a difference as large as D_{obs}, if in fact some hypothesized value Δ' were the true increased risk? I call the answer to this question the severity of observed difference D_{obs} against an increase Δ':

The severity of observed difference D_{obs} against the alternative hypothesis that $\Delta = \Delta'$
equals the Prob(such a large difference, given that $\Delta = \Delta'$)
i.e.,Prob($D \geq D_{obs}$, given that $\Delta = \Delta'$)

for Δ' ranging ranging over the possible values of the parameter Δ. By "such a large difference" I mean one as large as or larger than the one observed, D_{obs}. (Note that for the case where $\Delta' = 0$, the severity of the difference equals its significance level.) The higher the test's severity against positive values of Δ', the higher its chance to detect Δ' (by rejecting H).

We can discriminate between legitimate and illegitimate construals of a statistical result by considering the values of the severity function for various values of Δ in the parameter space. For the same reasons we noted in our examination analogy above, a rejection of H only indicates that the increase Δ exceeds some value Δ', if it is improbable that Δ' brought about so large an observed difference. This may be formulated as the following rule for understanding rejections, [RR]:

[RR]: *A* T+ *rejection with observed difference* D_{obs} *indicates that* $\Delta > \Delta'$ to the extent that so large a difference is improbable were Δ no greater than Δ'—i.e.,to the extent that the severity of D_{obs} against the hypothesis that $\Delta = \Delta'$ is *low*.

So reasonable lower bounds are alternative hypotheses (i.e., positive Δ-values) against which the observed difference has low severity.

Out of a desire to obtain a test with high severity against an alternative of interest, say Δ', it has sometimes been suggested (e.g., by Cranor) that the test's size (i.e., the α level required to reject H) be raised. The above reasoning should make clear why such a suggestion fails to accomplish its aim. While the resulting test (with raised α) may now classify a previously negative result as one significant enough to reject H, that rejection will not indicate the existence of the increase of interest. The reason stems from the following consequence of rejection rule (RR):

(From [RR]) D_{obs} is a poor indication that Δ exceeds Δ' if a difference as large as D_{obs} would occur frequently even if Δ were no greater than Δ'.

Note that if we choose a small size, the test's severity against 0 is low. So the reason one wants a small size (in a non-automatic use of tests) is not the desire for low long-run frequency of error, but the desire that each *particular* rejection of the null hypothesis warrants inferring that the increase exceeds 0. Otherwise the result is misleading.

b. Understanding Failures to Reject H (i.e., Acceptances of H):

The problem in the formaldehyde conflict was not the interpretation of the positive rat studies (where hypothesis H was rejected), but the interpretation of the negative epidemiological ones (where H was not rejected). This problem—how to interpret negative statistical results—is one that particularly plagues epidemiological studies for estimating cancer risk. For here sample sizes are typically small relative to incidence rates of the cancer in question. Even the best epidemiological studies can rarely detect increases in cancer risk of less than 1 in 10 (one additional cancer per 10 individuals), while smaller increases are typically of interest. (See, *Hearing*, p.763 and the study reported in Freiman *et. al.*, 1978.)

One of the main questions that was raised at the formaldehyde hearings was this: Do the failures to reject the hypothesis H of 0-increase indicate that there is little or no risk? The many scientists and organizations endorsing the PRL-1 document say no. Indeed, Todhunter's own epidemiologist on the staff responsible for this work wrote:

> Before leaving [the EPA], I would again like to emphasize that the available epidemiologic data from studies on formaldehyde exposure are inconclusive and *not supportive of no association*, as purported by the formaldehyde Institute. (*Hearing*, p.137, emphasis added)

But Todhunter and the Formaldehyde Institute say yes, claiming:

> There is a limited but suggestive epidemiological base which supports the notion that any human problems with formaldehyde carcinogenicity may be of low incidence or undetectable. ...[The ranges of risks] are of from low priority to no concern... .(*Hearing*, p.253)

Consider a study that was cited in defense of the Todhunter interpretation (*Hearing*, pp.137-138). In a mortality study of Du Pont workers, the relative risk of dying from cancer among those in the study exposed to formaldehyde was not statistically significantly greater than among those not so exposed: the null hypothesis H was not rejected.[14] Du Pont concluded that

> ...the data suggested that cancer mortality rates in the company's formaldehyde exposed workers were no higher than the rates among nonexposed workers. (*Hearing*, p.284)

They are inferring, in other words, that the increased risk Δ equals 0. The error in such an interpretation is this: Failure to reject the null hypothesis of 0-increased risk is not the same as having positive evidence that the increased risk is 0. For such negative results may be common (i.e., probable) even if the underlying increase in risk is greater than 0. In fact, the Du Pont study had a very small chance of rejecting null hypothesis H (i.e., of having H "fail the test") even were the actual increase in risk to exceed 0 by substantial amounts. For example, the Du Pont study had only a 4% chance of rejecting H, even if there were a twofold increase in cancer of the pharynx or of the larynx in those exposed to formaldehyde. Thus, failing to reject H does not rule out twofold increases in these types of cancers. The situation was even worse with nasal cancers and not much better with the others. This is indicated in the following chart adapted from a review of the Du Pont study by the National Institute for Occupational Safety and Health (NIOSH) (*Hearing*, p.549):

	Lung	Pharynx	Larynx
# of cases	181	7	8
Power to Detect Odds Ratio = 2*	37%	4%	4%
Least Significant Odds Ratio Detectable**	2.9	57.5	42.5

* Assumes $\alpha = .05$ (1 tail)
** Assumes $\alpha = .05$ (1 tail) and Power $(1-\beta) = .80$

Although I recommend interpreting tests by considering the severity function rather than the usual NP power function as employed in this chart, this does not alter the present point because high power entails high severity.[15]

More generally, failing to reject H does not rule out increases as large as Δ', if there is a small probability of rejecting H even if the increased risk were as large as Δ' (i.e., even if the severity against Δ' is low). All of this follows very familiar reasoning. If we are unlikely to hear a fire alarm (to get a rejection of H) even if there is a bad fire, then not hearing the alarm is not grounds for thinking there is no fire.

While a failure to reject does not indicate that the increase is 0, it does permit an inference about the likely *upper bound* of the unknown increase Δ. That is to say, a failure to reject H provides reason to say "the data provide good grounds that the increased risk is no greater than such and such (upper bound)". To find plausible upper bounds requires determining the extent of risk increase that with high probability would have resulted in a rejection of H, i.e, the increase against which the observed difference had *high severity*. In the Du Pont study, the test had a high probability (.8) of rejecting H, if the risk of cancer of the larynx were 42 times higher among exposed than unexposed workers. Hence a failure to reject H *does* indicate that the actual increased risk is not as high as 42-fold. (This follows from the fact that if the test has high power against an alternative, it has high severity against it.)

This leads to a general rule for understanding an acceptance of H with respect to test T+ (converting talk of ratios to differences) rule [RA]:

[RA]: A T+ acceptance with observed difference D_{obs} indicates that the actual increased risk Δ is less than Δ' to the extent that a larger difference would be probable, were the

increase as great as Δ'—i.e., to the extent that the *severity* of D_{obs} against the hypothesis that $\Delta = \Delta'$ is *high*.

So reasonable upper bounds are risk increases (Δ-values) that yield high severity values. Correspondingly, the smaller the test's severity against Δ', the *less well* a T+ acceptance indicates that $\Delta < \Delta'$.

As indicated on the chart above, the Du Pont study had a fairly good chance (80%) of detecting a 3-fold increase in the relative risk of lung cancer, and a 57-fold increase in the risk of cancer of the pharynx. So, even ignoring some methodological difficulties with the study, its negative statistical results at most indicate that the various cancers are no more than 3 or 57, etc. times as likely among workers exposed to formaldehyde. They clearly do not warrant the conclusion reached by Du Pont and others, that the study supports the claim of no increase in (relative) cancer risk among formaldehyde workers. The study does not even support the claim of low increased risk, given what Todhunter himself claimed the EPA counted as low.

We can summarize informally the interpretation of both positive and negative results in terms of what a difference D_{obs} indicates:

[RR]: An observed difference (in incidence rates) D_{obs} only indicates that the increased risk (in the population), i.e., Δ, exceeds those values that would rarely have resulted in so large an observed difference.

[RA]: An observed difference (in incidence rates) D_{obs} only indicates the nonexistence of those population risk increases (Δ values) that would frequently have resulted in a *larger* observed difference.

But how, it might be asked, should "rarely" and "frequently" be specified? We can get a feel for the increase indicated by using benchmarks such as .9 or .95 for very frequent, and .1 or .05 for rare. But by interpreting tests along the lines suggested in [RR] and [RA] the use of statistical tests should no longer be a matter of pre-specified error probabilities altogether. Instead we can understand what the actual result D_{obs} indicates more or less well by calculating all (or several) of the upper and lower bounds for different degrees of severity. This would yield *severity curves*. (While each pair of upper and lower bounds of a given degree of severity is mathematically equivalent to formulating the confidence interval at the corresponding level of confidence, the difference is that not all values within the interval are treated on par. It most closely corresponds to forming a series of confidence intervals, one for each confidence level.[16])

The criticism of the EPA assessment was based on the reasoning incorporated in the rule for interpreting acceptances, i.e., rule [RA]. A number of scientists concluded that "in order to justify its failure to address formaldehyde under 4(f)...EPA has rewritten both the science and the law." (*Hearing*, p.195). (See also Ashford, *et.al.*, 1983). Because of the criticisms of the science underlying the EPA risk assessment, the EPA ultimately did place formaldehyde under the 4(f) category in 1985.

5. Conclusion

The conflict in the formaldehyde example, we said, arose from a difference in RAP judgment. By means of the above understanding of the actual extent of risk indicated by a statistical test result, one can determine the protectiveness of a given RAP judgment. It allows one to answer the question: according to the standard being required, what extent of risk must be fairly clearly indicated before it is taken as grounds that there may be a significant human risk? It allowed critics to ascertain that—contrary to what Todhunter maintained—the Todhunter assessment reflected a change in the standard required for

triggering 4(f). Thus, even granting that the judgments required in reaching statistical assessments may reflect policy values, conventions, etc., I have argued, it does not follow that the task of evaluating *whether a given risk assessment is warranted by the evidence* need also be infected with subjective policy values. This task is an empirical one that may often be accomplished objectively, in the sense of reflecting what is actually the case regarding the risk, regardless of what anyone thinks is or ought to be the case.

Notes

[1]A portion of this research was carried out during tenure of a National Endowment for the Humanities Fellowship for College Teachers; I gratefully acknowledge that support. I would like to thank Marjorie Grene for numerous useful comments on earlier drafts.

[2]*Formaldehyde: Review of the Scientific Basis of EPA's Carcinogenic Risk Assessment.* Hearing Before the Subcommittee on Investigations and Oversight of the Committee on Science and Technology, U.S. House of Representatives, May 20, 1982. All pages in parentheses following *Hearing* refer to this report.

[3]This delineation of risk assessment and risk management is in accordance with such documents as National Research Council, *Risk Assessment in the Federal Government: Managing the Process.* Other uses of "risk assessment", in contrast, take it to include risk management.

[4]See Note 2.

[5]Carcinogens may be considered problematic if they increase the risk of cancer by 1 case in 10,000 or 1 case in 1,000.

[6]National Research Council (pp.29-33). My use of the term "risk assessment policy" comes from this report.

[7]He expressed essentially the same point in his oral presentation at this session of the PSA 1988.

[8]The Todhunter formaldehyde assessment endorsed other less protective choices, such as holding to the existence of a threshold for carcinogenicity of formaldehyde, discounting benign tumors, and preferring maximum likelihood estimates over upper confidence level estimates.

[9]For a discussion of the blacklisting of scientists and "hit lists" at the EPA during this period, see Lash, et. al.,(1984).

[10]I do so in "Sociological vs. Metascientific Philosophies of Risk Assessment".

[11]The same test would be used were the null hypothesis to assert that $\Delta \leq 0$.

[12]Since within the context of NP tests parameter Δ is viewed as fixed, hypotheses about it are viewed as either true or false. Thus, since a probability is interpreted as a relative frequency, it makes no sense to assign such hypotheses any probabilities other than 0 or 1.

[13]A discussion of these responses occurs in Mayo (1985).

[14]In the Du Pont study, 481 cases of male cancer deaths among employees between 1957-1959 constituted the cases. These were matched on relevant factors with controls

who did not die of cancer. The statistic observed was the relative odds ratio, the ratio of the odds of having been exposed to formaldehyde among cases and controls. For simplicity, I refer here to the risk rather than the relative risk.

[15]The difference between power and severity is that while severity is a function of the particular observed difference D_{obs}, the power is a function of the smallest difference judged significant by a given test. Let D* be the smallest difference test T+ judges significant. (i.e., D* is the *critical boundary* shown in Fig. 1 beyond which the result is taken to reject H.) Then, power is defined as follows:

The power of test T+ against alternative $\Delta = \Delta'$ equals the probability of a difference as large as D*, given that $\Delta = \Delta'$.

The severity, in contrast, substitutes D_{obs} in for D*. The advantage of the severity function, I claim, is that it affords an understanding that reflects the difference that has actually been observed.

[16]For further discussion of this relationship, see Mayo (1985). Poole (1987).makes use of what are essentially severity curves in interpreting statistical results. Such curves are also employed by Kempthorne and Folks (1971).

References

Ashford, N.A., Ryan, C.W. and Caldart, C.C.(1983), "A Hard Look at Federal Regulation of Formaldehyde: A Departure from Reasoned Decisionmaking", *Harvard Environmental Law Review* 7:297-370.

Cranor, C. (1987), "Some Public Policy Problems with Epidemiology: How Good is the 95% Rule?". Paper presented at the Pacific Division meeting of the American Philosophical Association, March 1987.

Douglas, M. and Wildavsky, A. (1982), *Risk and Culture*. Berkeley: University of California Press.

Fisher, R.A. (1955), "Statistical Methods and Scientific Induction", *Journal of the Royal Statistical Society* (B) 17:69-78.

Fleiss, J.L. (1986), "Significance Tests Have a Role in Epidemiologic Research: Reactions to A.M. Walker", *American Journal of Public Health* 76 (No.5, May 1986): 559-560.

Formaldehyde Federal Register Notice, May 1981.

Freiman, J.A., Chalmers, T.C., Smith Jr., H. and Kuebler, R.R. (1978), "The Importance of Beta, the Type II Error and Sample Size in the Design and Interpretation of the Randomized Control Trial, Survey of 71 'Negative' Trials", *The New England Journal of Medicine* 299 (No.13):690-694.

Funtowicz, S.O. and Ravetz, J.R. (1985), "Three Types of Risk Assessment: A Methodological Analysis", in *Risk Analysis in the Private Sector*, C. Whipple and V.T. Covello (eds.). New York: Plenum Press, pp.217-231.

Kempthorne, O. and Folks,L. (1971), *Probability, Statistics, and Data Analysis*. Ames: Iowa State University Press.

Lash, J., Gillman, K. and Sheridan, D. (1984), *A Season of Spoils: The Reagan Administration's Attack on the Environment*. New York: Pantheon Books.

Mayo, D. (1985), "Behavioristic, Evidentialist, and Learning Models of Statistical Testing", *Philosophy of Science* 52:493-516.

_____., "Sociological vs. Metascientific Philosophies of Risk Assessment", in *Acceptable Evidence: Science and Values in Risk Management*, D. Mayo and R. Hollander (eds.). Forthcoming, Oxford.

National Research Council, *Risk Assessment in the Federal Government: Managing the Process*. Washington, D.C.: National Academy Press, 1983.

Neyman, J. and Pearson, E. S. (1933), On the Problem of the Most Efficient Tests of Statistical Hypothesis", *Philosophical Transactions of the Royal Society* A 231: 289-337. (Reprinted in Joint Statistical Papers, Berkeley: University of California Press, 1967, pp.276-283).

Pearson, E.S. (1955),"Statistical Concepts in Their Relation to Reality", *Journal of the Royal Statistical Society* (B) 17:204-207.

Poole, C. (1987), "Beyond the Confidence Interval", *American Journal of Public Health* 77 (No.2, Feb. 1987):195-199.

Silbergeld, E.K., "Risk Assessment and Risk Management—An Uneasy Divorce", in *Acceptable Evidence: Science and Values in Hazard Management*, D. Mayo and R. Hollander (eds.). Forthcoming, Oxford.

U.S. House of Representatives, *Formaldehyde: Review of the Scientific Basis of EPA's Carcinogenic Risk Assessment*. Hearing Before the Subcommittee on Investigations and Oversight of the Committee on Science and Technology, 97th Congress (second session), May 20, 1982.

Walker, A.M. (1986), "Reporting the Results of Epidemiologic Studies", *American Journal of Public Health* 76 (No.5, May 1986):556-558.

Weinberg, A. (1972), "Science and Trans-Science", *Minerva* 10:209-222.