First published in *Theory and Psychology*, 1991, 1, 383–388. In response to Chow (1991).

Conceptual Rigor: Where Is It?

Abstract

Conceptual rigor is indeed a desideratum worth dedicated pursuit; in fact, one might wish that Chow had pursued it somewhat more diligently in his present essay. I suggest that the approach to data interpretation he advocates here is an etch-a-sketch draft whose prospect for refinement into an operational logic of inference that professional scientists can live by appears minuscule.

It is refreshing to see an open-throttle challenge to currently prominent outlooks on data assessment, especially one that tries to dig deeper than mere appraisal of statistical models. Far too few psychologists today seem aware—or at least willing to acknowledge—that the logic by which rational thinkers can devise credible explanations for empirical observations is still profoundly problematic, not just in philosophical theory but in scientific practice as well. One can commend a paper for its forceful reopening of fundational issues even while strongly disagreeing with the positions it takes.

And disagree with Chow's (1991) perspective I most emphatically do. At the level of slogans, we are diametrically opposed on the epistemic merit of the Popperian hypothetico-deductive model of scientific inference and its statistical excrescence, null hypothesis significance testing. Whereas I have repeatedly argued that these are mindless abominations (Rozeboom, 1960, 1970, 1972, 1980, 1990), Chow wants us to intensify the doctrinaire extremity with which we practice them. Is this estrangement too vast for dialog to bridge? Not necessarily. Conflicting allegiances, once acknowledged, need not spoil pursuit of agreement on carefully restricted issues which have same prospect of eventual expansion into larger conciliations. In the small space available to me here, I would like to engage Chow on two such issues.

Although much of Chow's essay concerns appraising data by sampling-statistical models, which is virtually the only facet of data interpretation that ever surfaces in our methodological literature, he commendably emphasizes that this is just a first step toward the large goals of any serious science, namely, adjudicating explanatory theories of our data. (Chow calls these 'substantive' theories, but I suggest that 'explanatory' makes more explicit the sort of substance at issue.) Sampling-theoretic judgments about the population parameters to which sample statistics should converge as sample size becomes arbitrarily large are in epistemic principle (albeit admittedly not in practice) just an almost-trivial preliminary to our

attempted deciphering of what these population parameters might tell us about underlying reality. Any reputable account of how to interpret the data forthcoming from some particular empirical inquiry must above all envision some conclusions that could appropriately if tentatively be drawn were sample size large enough to make sampling uncertainty negligible. It makes good practical sense to supplement any large-sample inference model with some concern for how its guideline should be amplified or amended when the sampling noise in our results is appreciable. But unless we are clear on what to do with asymptotical large sample results, preoccupation with any type of sampling-theoretical assessments is on an intellectual par with the dog who incessantly chase cars but would not know what to do with one if he ever caught it.

On first impression, Chow's 'conceptual rigor' model of theory appraisal nicely fits this two-phase inference process. First, we are to derive from some attractive explanatory hypothesis H_0 a statistical expectation for some experiment's outcome, and then judge H_0 to be confirmed or rejected according to whether the observed outcome is suitably similar to that expectation. But major obscurities—indeed, remarkable deficits in conceptual rigor—emerge when one attempts to pin down Chow's specifics. To request clarification, let us start with how, in his Step 4, we are to 'decide whether or not the outcome of the study, D, is similar to [theory-derived expectation] X. What counts as 'similar', and how is this decided? We are told that 'a test of significance is used to determine if D is, indeed, similar to X.' Does this mean that D's lying within the test's acceptance region, *Accpt*, is what *defines* similarity of D to X? (If so, no decision, needed; we simply observe whether or not D is indeed in *Accpt*.) But ordinary English insists that 'similarity' is by conceptual intent a matter of degree. And if, outraged commonsense notwithstanding, we were to accept construing D's similarity to X as a binary alternative defined by *Accpt*, this similarity would become importantly dependent on one's whims in choosing a significance level—surely not the intent of Popperian hypothetico-deductivism.

Moreover, even if we allow Chow to arrogate a fixed alpha level how does he reconcile his sense of similarity with the dependence of *Accpt* width on sample size? I shall not repeat here the common but still strong cogent objection to null-hypothesis significance testing that we can guarantee rejection of H_0 with virtual certainty by taking a sample size sufficiently large. Instead, envision the following scene: researchers Smith and Jones agree that a particular theory H_0 of mutual interest implies that in a certain experimental set-up, fully specified except for not-yet-chosen sample size, the expectation for outcome measure X (say a mean treatment difference) is zero. Smith and Jones both carry out this experiment in their respective labs, but Smith uses 100 times as many subjects as does Jones and gets an acceptance interval $Accpt_{\text{Smith}} = [-.92, +.92]$ in contrast to the $Accpt_{\text{Jones}} = [-10.7, +10.7]$ obtained by Jones at the same alpha level. Now

suppose that the salient empirical outcomes in these two experiments are respectively $D_{\text{Smith}} = 1.01$ and $D_{\text{Jones}} = 5.02$. Does Chow really want to hold that result $D_{\text{Jones}} = 5.02$ is confirmationally similar to $X = 0.0$ but $D_{\text{Smith}} = 1.01$ is not? How could anyone not conclude instead that Smith's result is much more similar to the theoretical expectation under H_0 than is of Jones's result and, indeed, if 1.01 is quite small in comparison to the X-values that are reasonable to anticipate if H_0 is false, that Smith's result is rather strongly supportive of H_0, much more so than is Jones's. For surely not even Popperians are required to condemn a theory that is almost but not quite correct in a certain respect as intolerably inferior to one that gets this respect exactly right. (Popper's own notion of similitude' speaks to the contrary.) More briefly, what would become of Chow's recommendations for this first stage of data interpretation in an alternative universe wherein empirical research is always required to use sample sizes so enormous that not only is sampling error always negligible but sampling-theoretic statistics have never even been invented?

Let me wrap this point by asking Chow to imagine that he edits an experimental journal whose publication space is extremely tight. If he were ruthless in rejecting submissions that are padded with outcome details that make no useful contribution to accumulated scientific knowledge, would he forbid authors to describe more than their null hypothesis, their experimental set-up, and whether or not their data called for rejection of H_0 at some mandated significance level? Or would editor Chow also tolerate reports of observed effect size, sample variance and perhaps even power analyses or confidence-interval estimates? But if he does concede that the latter may have some modest scientific value, what role is envisioned for them in his account of scientific inference?

I now shift to a far more troublesome issue. Suppose, within Chow's framework, that sample statistic D is so close to its expectation X under H_0 (i.e. an $A.E_1$ or maybe K in Chow's Table 3) that only a churl could cavil at viewing this result as triumphant verification of H_0's prediction. Some propositions underlying X have thus gained credibility, that is, have received confirmation; and almost any plausible philosophic model of rational belief change will concede that if H_0 logically entails the verified prediction, confirmation in this case extends to H_0 as a whole.[1] But confirming H_0 *as a whole* does *not* confirm every prepositional constituent of H_0, that is, everything implied by H_0. The significant problem for adjudicating theory H_0 in light of confirmatory evidence D is to discriminate, with selective care, between constituents of H_0 whos plausibilities are appreciably enhanced by D and those to which D is indifferent or even disconfirmatory.

[1] In Rozeboom (1980) I show that if H_0 entails only a conditional If-A-then-X. as usual in experimental research, observation of X under condition A does not necessarily confirm H_0. But this complication is not the issue here, especially since circumstances which allow this confirmation to fail seldom if ever hold in research practice.

Popperian hypothetico-deductivism in general, and Chow's version in particular, is blind to this vital epistemic responsibility and offers no clue whatever to its effective management. To his credit, Chow has not ignored my past arguments on this matter; but he has attempted here to insulate his position against them by throwing up a fog of obfuscation which intimates, without at all clarifying, that these arguments are formalistic artificialities with no bearing on genuine 'non formalistic psychological theory [which cannot generally] be decomposed and reconstituted at will'. Although I have no idea what might distinguish formalistic theories from non-formalistic ones, or how it helps Chow's case not 'to identify theoretical implication with logical entailment,' or in what sense a theory should be 'retained only when its supporting data have internal and external validity' (oh, that elusive conceptual rigor!), I try to illustrate here the problem of constituent confirmation with commonsensical informality.

My stock example of how confirming a whole need not confirm all its constituents is the imaginary case of Popper's paranoid disciple who, attempting to adjudicate his heretofore unsubstantiated suspicion of his wife's infidelity, constructs the theory T: 'My wife is unfaithful and the sun rises every morning', notes T's implication that the sun will rise tomorrow and becomes violent with jealousy when the sun's rising tomorrw confirms his wife's infidelity via its confirmation of a theory that entail this, namely T. Commonsensically, this inference seems utterly absurd (even though standard hypothetico-deductivism recognizes no rational principle that repudiates it) in that T's solar-kinematics constituent, which yields the verified prediction, has no evident linkage with T's sexual transgression part. But now suppose further that, when this oddball is brought to trial for wife abuse, he defends his inference hypothetics deductively by appeal to his conjecture (T^*) that not only does the suaris every morning but also prolonged exposure to regular cycles of strong light/dark alternation triggers sexual promiscuity in mammalian femaales. Commonsense still considers it crackpot to infer infidelity of one's wife from observing a sunrise, causal hypothesis T^* notwithstanding; but how does Chow disown this? Should not his position accept at least a modicum of T^*-mediated flow of credibility from observed sunrise to wifely infidelity, even if the confirmation is rather weak?

If this Case of the Paranoid Popperian seems insufficiently non-formalistic to Chow, consider a different example. A recent acquaintance, Monica, has chanced to mention being the youngest child in her family. Since this reveals that she has older siblings, should you be tempted to construe her evident femininity (call this datum FM) as confirming that her oldest sibling is also female (call the latter possibility FO)? For both FM and FO are straightforward consequences of the theory (T) that all children of Monica's parents are female, which is splendidly confirmed by observation FM. (For example, conditional on the background assumption that Monica's parents have conceived just two children with each having

had independent .5 probability of being a girl, T's credibility should go from .25 before determination of Monica's gender to .50 afterward.[2]) But does FM also confirm Ts additional consequence TO? Common sense is disposed to deny this—until it considers the prospect (T^*) that Monica's father may have a genetic defect (call this 'Boybane') that blocks mobility in spermatozoa carrying a Y-chromosome. T^* too is confirmed by your observation FM; and now, considering possibility T^*, it is no longer so commonsensically plain that FM should not be taken to confirm FO as well. What does Chow want to say in this case? Does his model of scientific inference *mandate* confirmation of FO by FM through their mutual entailrnent by a respectable medical hypothesis with explanatory force? Or is he able to agree with me, instead, that we simply do not understand the deeper logic of non-demonstrative inference well enough as yet to be dogmatic about what confirms what in contexts like this.[3]

Lest Chow be overly quick to assure us that observation FM does indeed increase FO's plausibility because it is genuinely possible, even if unlikely, that Monica's father is afflicted with Boybane, consider that any two logically compatible propositions, S_1 and S_2, are joint consequences of many theories $\{T_i\}$ that do not merely juxtapose S_1 and S_2 by 'formalistic' conjunction but contrive some sort of explanatory cohesion for them. (If no more intriguing possibilities come to mind, we can always fall back on some version of 'Almighty God requires both S_1 and S_2 in His grand scheme of things'. Popperian hypothetico-deductivism is not at all abashed if these theories seem rather low in plausibility: the bolder the better.) But for every such T_i we can also create a contrastive counterpart T_i' that implies both S_1 and $\sim S_2$ (i.e. Not-S_2). So under what circumstances does verification of S_1 confirm S_2, rather than disconfirming it or leaving its credibility unaltered? Surely not just when some Popperian is actively appraising an explanatory theory of which S_1 and S_2 are both consequences while neglecting also to contemplate alternative explanations for S_1 that entail $\sim S_2$. I expect Chow to confront this challenge by appeal to repeated testing of the salient theory, a stratagem that does not work but which I cannot properly attack until Chow deploys his arguments. To close, I merely submit once again that trusting hypothetico-deductive models

[2]I am conflating credibility and objective (statistical) probability in this quantitative illustration; but it makes the point while evading the extensive digression an honest account would require.

[3]Even if serious contemplation of T^* seemingly justifies taking FM to confirm FO via T^*, is it then also rational to view FM as confirming FO when we are actively aware that both are consequences of T but have no T^*-like notion of why T might be true? If not, on what grounds do we reject the latter inference while accepting the former? There is much more to be said about this example: simple as it at first seems, it readily unfolds into deep unresolved problems in the foundations of scientific inference and probability theory. I do not have satisfactory answers to all these puzzles; but for anyone interested, I will happily conduct a tour of disturbing exhibits in this chamber of horrors from which certain important insights into the nature of real-world scientific inference can be derived.

of scientific inference to guide data interpretation in research practice is like investing all your savings in a stock promotion that promises 200 percent annual return on your investment.

References

Chow, S. L. (1991). Conceptual rigor versus practical impact. *Theory & Psychology*, *1*, 337–360.

Rozeboom, W. W. (1960). The fallacy of the null-hypothesis test. *Psychological Bulletin*, *57*, 416–428.

Rozeboom, W. W. (1970). The art of metascience, or, What should a psychological theory be? In J. R. Royce (Ed.), *Toward unification in psychology.* Toronto: Toronto University Press.

Rozeboom, W. W. (1972). Scientific inference: The myth and the reality. In R. S. Brown & D. J. Brenner (Eds.), *Science, psychology, and communication: Essays honoring William Stephenson.* New York: Teachers College Press.

Rozeboom, W. W. (1980). Nicod's criterion: Subtler than you think. *Philosophy of Science*, *47*, 638–643.

Rozeboom, W. W. (1990). Hypothetico-deductivism is a fraud. *American Psychologist*, *45*, 555–556.