

Front Page
Presidential Column
Teaching Tips
Student Notebook
Department Profile
People
Members in the News
Announcements
Meeting Calendar
Browse all Classifieds
Search the Classifieds
Submit a Classified
Contact the Observer
Subscriptions

Reviewing for Risk: What's the Evidence That It Works?

BY JOHN H. MUELLER AND JOHN J. FUREDY

The authors consider the effectiveness of the process for reviewing research proposals in terms of risk to human subjects, as it has evolved in psychological research in North America. They raise a fundamental question: Is there any evidence that these reviews are effective at reducing risk to the public? In Part I, they define the situation and identify some irrelevant measures of effectiveness. In Part II, which will be published in the October 2001 Observer, they discuss approaches and benefits to answering this question.

THE SITUATION

Over the past several decades, behavioral and social science research proposals have come under increasing scrutiny by Institutional Review Boards (IRB) in the United States and Research Ethics Boards (REB) in Canada. The mandate whereby these review groups emerged was in the interest of protecting human participants from "extraordinary risks," with everyday risk being accepted as unavoidable.

From this reasonable base, which involved departmental-level review, a veritable industry has developed and expanded in several directions. Reviewing for risk to human research subjects now is obligatory for all proposals, not just those that seem problematic, and not just for federally-funded research but every project on campus. The review is no longer entrusted to the departmental level, but generally occurs at some campus-wide level, where expertise in the research area too often is secondary to self-expressed interest in "ethics" or "bio-ethics."

Consequently, review now often inappropriately extends beyond experimental design, plus "risk" has been redefined to include the more nebulous notion of "ethics." (Some of the issues that are raised today in reviews seem more properly labeled "etiquette" rather than ethics; certainly they are not "risk" in any common usage of the term.) A further complication is that the ethical issues which preoccupy medical researchers are presumed to be relevant to every department on campus.

As we begin to contemplate concerns such as "beneficence," "respect," "justice," and "liability," along with obligatory indoctrination workshops as a prerequisite to review, it is clear that the limiting horizon for this expansion is not yet in sight. Contrary to Adair (2001) and Puglisi (2001), we see no evidence that, if we just learn the rules and cooperate, then the regulators will cease to encroach on intellectual inquiry in the social sciences. Sadly, the pattern has been quite the opposite thus far.

In the U. S., the IRB situation has become so murky that the best advice some can give is "Don't talk to the humans" (Shea, 2000). In Canada, the status and scope of REBs has been expanded by the implementation of the Tricouncil Policy Statement in May, 1998, a joint statement developed by the three major funding agencies, the Medical Research Council (MRC), the Natural Sciences and Engineering Research Council (NSERC), and the Social Sciences and Humanities Research Council (SSHRC). In contrast to U.S. practice, this document was labeled a "code" rather than a set of guidelines, and hence attracted considerable international attention (see, e. g., Azar, 1997; Holden, 1997).

The current version of the Tricouncil Policy Statement lacks some of the original attacks on the basic epistemological function of research, such as the rule that if a subject, during debriefing after an experiment, finds the researcher's hypotheses offensive, then that subject can withdraw his or her data (e. g., Furedy, 1997, 1998). However, even the current version of the Tricouncil statement has been criticized for what some view as being unsuitable for application to psychological and sociological research on humans (e. g., Howard, 1998), and there are no guarantees that the next iteration will not try to reinstate such anti-intellectual requirements. Is this expanded review effort worth it? For that matter, was the review process working before recent expansions?

We do not attempt to provide a full cost/ benefit analysis of the review process in this series. Such an analysis would need to consider, among other things, the distinction between epistemological and ethical functions, and the potentially deleterious educational effect on young researchers who are increasingly trained in how to pass ethics reviews rather than being educated in the complex research problems of their discipline. Rather, we focus on issues related to a specific benefit, using the business model metaphor that we are advised is so relevant to the campus these days: Are we getting our bang for the buck? Specifically, what hard evidence is there that the review process does in fact reduce "problems" (i. e., untoward incidents during the experiment)? We suggest checking the key performance indicators to be sure that we are getting corresponding benefit in terms of reduced hazard to subjects as a return for our increased efforts in reviewing proposals. Of course, this question does not apply to just psychological research, because the issue of effectiveness and accountability pertains to the review process rather than the subject matter of the proposals.

ARE REVIEWS WORKING? HOW CAN WE TELL?

Nearly 20 years ago, Ceci, Peters, and Plotkin (1985) briefly considered the cost of the IRB process. The evidence then at hand was largely estimates, and the cost was said to be "sound insurance" (p. 995). However, that evaluation was not focused on what we see as the key indicator of effectiveness: *concrete evidence of incident-avoidance*. In fact, the review industry seems reluctant to count incidents that arise, apparently finding comfort in the prospect that reviews were warranted if they avoid "even a single case of malfeasance" (p. 995). (More on this last point below.)

The enterprise has grown and changed considerably over the past 20 years; it is legitimate to ask whether the expansion per se is providing better protection to the public. Answering this question requires the use of appropriate indicators of effectiveness.

Evidence supporting effectiveness for the review process might come from something straight-forward, such as the number of incidents (e. g., subject complaints) arising from research were reported in 1950, 1960, and so forth, per decade. The question is whether those data show progressively fewer incidents per experiment conducted over the last 50 years, during which time there has been ever more aggressive screening of research proposals. This would hardly prove a causal connection, but it seems a minimal expectation that more review effort should result in fewer problems reported from the laboratory.

Having proposed this as an indicator of effectiveness, we now must point out its inadequacy. We doubt that the incident rate is going down, for two main reasons.

First, anybody can complain about anything these days. No matter how much screening and no matter how trivial the concern in absolute terms, they will still find someone to nurture them along for a legal fee. Review boards don't have any influence on this aspect of our increasingly litigious society.

Second, - and this gets back to the point about malfeasance - the "bad guys" are not going to come asking IRB/REB permission. The proposal review movement was stimulated by the hearings at Nuremberg, but there is no basis for the inference that the Holocaust would have been prevented had ethics review boards been in existence during the war. Neither Dr. Mengele or Dr. Frankenstein would have applied to an ethics review board, and their contemporary counterparts will not do so either. The review system cannot prevent acts of malfeasance, but this simple truth continues to be misunderstood. We say this judging by the persistent use of past incidents as justification for IRB reviews (e. g., the recent discussion of the Unabomber's experiences, Chase, 2000). Many of these incidents that now seem inappropriate were quite legal in their time and place, thus IRB review would have approved them. Others were performed by agents officially (e. g., military LSD research) or unofficially operating beyond the law, and these would still not be restrained by IRB review.

PROBLEM FINDING 101

At least one type of data can be dismissed as bogus evidence for the effectiveness of the review system: When an ethics reviewer alleges that something in a proposal is a problem, some people equate "a problem found" with "an incident avoided" and point to this as justifying the review system. But it doesn't work that way, and this assumption needs to be made explicit and rejected. "Revision requested by IRB" does not constitute a "problem" that would have occurred during the experiment.

By analogy, consider a company that is obliged to institute an accident prevention program for the workplace. Someone dutifully goes around and identifies alleged hazards, and amasses an impressive count of things "fixed." Is this relevant? No, and in the non-academic world it would seem preposterous to accept this hazard count as an indication of the success of the intervention. The only acceptable evidence would be whether the actual rate of accidents declined. Actual outcome measures are required for assessing IRB value as well.

For ethics reviews specifically, the problem-found count is flawed for a couple of reasons:

No consensus on definition of risk. That something is identified as a problem by an IRB reviewer does not mean the subject in the experiment will see it as a problem. There is far from perfect overlap between the "professional" and the "public" perception of a problem. This is supported by the fact that occasional incidents arise in projects that reviewers approved as clean. There is no reason to believe that this sword doesn't cut both ways; in other words, things that reviewers see as potential problems would be non-events to the public. In fact, the latter is increasingly likely as the nature of the reviewers' criteria become more nebulous and personal. "Revision requested by REB" may indicate something about the creative abilities of the reviewers, but it is not a realistic barometer of the success of the ethics review process at avoiding risk.

Worst case is not the norm. The review process seems to be dedicated to identifying a "worst case" scenario, and then proceeding as if the worst case will be the norm. In fact, this seems to have evolved to the point where the review assumes that the worst case will be not just the norm but the certainty, which of course is simply nonsense. Just because something "could" happen does not mean it "will." And when the worst case is an improbable event, this confusion becomes more wasteful and inappropriate.

To illustrate, one might be hit by a truck leaving the office, but it would be unwarranted for your spouse to book an appointment with the undertaker this afternoon on that possibility. You might win the lottery next weekend, but it would not be prudent to hit your boss in the face with a pie this afternoon.

If the demand in an IRB review is conceived as "find any problem" (zero risk), the demand can be satisfied; a problem will be found. However, this practice encourages the identification of outcomes that are highly unlikely and/or inconsequential in terms of harm, and even though a problem is found its correction offers no benefits commensurate with the effort, relative to everyday risk.

That's why the original concept of "everyday risk" was useful. Unfortunately, the goal of achieving "zero risk" seemingly has replaced the rational acceptance of everyday risk, with no evidence that this policy offers additional protection for research subjects.

Discrete incidents. An appropriate accident metaphor is flight insurance. The experiment occurs in a discrete interval of time, like an airplane flight. So the question is, does a problem occur during that specific interval of time? Life insurance for your lifetime involves an unfortunately high and definite probability of death, whereas flight insurance is whether you die during a discrete interval of time. Most financial advisers have long considered flight insurance to be grossly over-priced, similar to the argument we are making about the ethics review process. Confusion of different kinds of risk is quite useful to the insurance industry, but expensive to the consumer. For whom is it useful to confuse varieties of risk in the IRB review process?

(And, no, considering institutional risk to be the collection of all experimenters working does not convert it to a cumulative risk. Each experiment (flight) is an independent risk.)

In short, "revision requested" cannot be a metric for the success of the IRB review process at avoiding harm in the experimental setting. Its inappropriateness increases to the point of rendering it an utter waste of time when the alleged risk in question is low probability. As emotionally satisfying as discovering a "problem" might be, such identifications do not constitute evidence with regard to documenting review effectiveness in protecting research subjects.

Also in the category of bad evidence: It is possible to imagine a situation in which a letter goes around campus stating something to the effect that "We had no complaints from experimental subjects this year, thanks to the diligent efforts of our ethics reviewers." We hope that survivors of Statistics 101, if not Psychology 101, can see the problem with such a causal attribution. It is quite possible that no complaints would have been received even without the review. We are always wise to heed the maxim that correlation is not causation, and further acknowledge that superstitious behavior is generally inefficient or wasteful of effort.

In Part II of this commentary, we will describe other approaches to assessing review effectiveness, based upon documenting the rate of incidents actually arising in the experiment.

Part II will appear in the October 2001 issue of the Observer.

REFERENCES

Adair, J. G. (2001). Ethics of psychological research: New policies; Continuing issues; New concerns. *Canadian Psychology*, 42, 25-37.

Azar, B. (1997). Ethics-code changes may dampen research efforts. *APA Monitor*, 28(March), 27.

Ceci, S. J., Peters, D., & Plotkin, J. (1985). Subjects review, personal values, and the regulation of social science research. *American Psychologist*, 40, 994-1002.

Chase, A. (2000). Harvard and the making of the Unabomber. *Atlantic Monthly*, 285 (6), 41-65. URL: <http://www.theatlantic.com/cgi-bin/o/issues/2000/06/chase.htm>

Furedy, J. J. (1997). "An interpretation of the Canadian proposed tri-council ethics code: Epistemological crime and cover-up", in a symposium on "Social policy masked as ethics hurts science: Some working scientific perspectives," at Society for Neuroscience meeting, New Orleans, November, 1997.

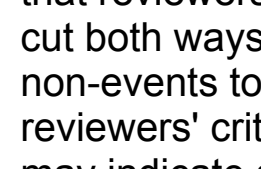
Furedy, J. J. (1998). Ethical conduct of research from code to guidelines: A shift in the Tricouncil approach? *Society for Academic Freedom and Scholarship Newsletter*, 18, 6.

Holden, C. (1997). Draft research code raises hackles. *Science*, 274, 1604.

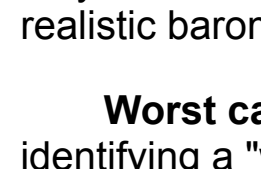
Howard, R. E. (1998). Letter to Nina Stipich, Senior Policy Analyst, SSHRC. Personal communication, February 20, 1998. In Furedy, J. J., Eminent scholar's concern about tri-council statement, *Society for Academic Freedom and Scholarship Newsletter*, 20, 3-8.

Puglisi, T. (2001). IRB review: It helps to know the regulatory framework. *APS Observer*, 14, (5), 1.

Shea, C. (2000). Don't talk to the humans: The crackdown on social science research. *Lingua Franca*, 10 (6), 26-34. URL: <http://www.linguafranca.com/print/0009/humans.html>.



JOHN H. MUELLER, a Fellow of APS, received his PhD in experimental psychology from St. Louis University in 1968, served on the faculty at the University of Missouri, and has been at the University of Calgary since 1990, where he is a professor in the Division of Applied Psychology. His work has been in the areas of human learning, social cognition, and educational technology, with additional interests in the history of psychology and issues involving academic freedom. You can reach him via e-mail: mueller@ucalgary.ca.



JOHN J. FUREDY, an APS Fellow, received his PhD in Psychology from the University of Sydney in 1967. His work has been in psychophysiology, philosophy of science, and sex differences in human and animal cognition. He also is engaged in advocacy, debate and scholarship concerning issues such as individual merit versus group-determined identity in the evaluation of faculty and students, and university autonomy, and on conceptual and empirical research regarding education vs. indoctrination in higher education. You can reach him via e-mail: furedy@psych.utoronto.ca.