POLICY FORUM: ANIMAL RESEARCH

Reliability of Protocol Reviews for Animal Research

Scott Plous* and Harold Herzog

Over the past 20 years, the reliability of scientific peer-review judgments has been a topic of frequent debate and scrutiny. However, one area of peer review that has not received much empirical investigation is the system that protects animal subjects from research risks. At most research institutions, studies involving animal subjects must be approved by an Institutional Animal Care and Use Committee (IACUC).

Low levels of intercommittee agreement were found in an early study in which 32 IACUCs evaluated four mock animal research protocols (*1*). Although these findings are useful, they were based on a nonrandom sample of committees, and the protocols were not representative of actual animal research proposals (e.g., all protocols were modified to contain problems). In addition, protocols were reviewed at only the group level, leaving open the possibility that interrater agreement was high among individual members of the same committee.

To overcome these limitations, we conducted a study of randomly selected IACUCs from U.S. universities and colleges. Seventy committees were drawn from a master list of 916 IACUCs maintained by the U.S. Office for Protection from Research Risks. Of these 70, 50 agreed to participate in the study. Thirty-four IACUCs came from research or doctoral universities, seven came from master's colleges or universities, six came from specialized institutions (e.g., medical colleges), and three came from liberal arts colleges (*2*). In all, 494 of 566 voting members (151 females and 343 males), or 87% of those approached, took part in the study.

Each IACUC was asked to submit its three most recently reviewed protocols involving animal behavior, including the committee's decision on whether to approve the research in question (*3*). All information identifying the investigator or institution was then removed from the protocols, and each protocol was randomly assigned to be reviewed a second time by another participating IACUC. Voting members of the second committee were sent packets containing three masked protocols with a request to review the protocols and to send us a completed evaluation anonymously in a prepaid envelope.

For each protocol, IACUC members were asked to provide recommendations according to four categories used routinely by most IACUCs (see the table on this page). Members were also asked to rate each protocol on several four- or five-point dimensions (see the table on page 609). These dimensions were chosen because they represent the most common criteria IACUCs use to judge protocols and because several federal and professional guidelines explicitly require IACUCs to consider such questions (*4*–*6*). As recommended (*6*), IACUC members were furnished with a scale in order to rate the degree of pain animals were expected to experience (*7*).

Once we received reviews from individual committee members, the IACUCs were asked to meet as a group and render a final evaluation for each of the three protocols. Committees were asked to follow their standard operating procedures and to discuss the protocols as they would any other research proposal.

FREQUENCY DISTRIBUTION OF ORIGINAL AND SECOND PROTOCOL RECOMMENDATIONS*

	Original recommendation			
Second recommendation	Approve as written	Contingent approval	Defer decision	Disapprove protocol
Approve as written	6	11	2	0
Contingent approval	26	24	1	2
Defer decision	31	27	2	1
Disapprove protocol	9	7	1	0
Total	72	69	6	3

*"Contingent approval" means approval with clarification or minor modification. "Deferred decisions" require further information.

To assess agreement between committees we computed kappa (κ), a measure of chance-corrected agreement used with categorical judgments (*8*). To determine agreement within committees (i.e., intracommittee reliability) we used a specialized version of the intraclass correlation coefficient suitable for cases when the number of raters per object is unequal (*9*).

Protocol evaluations from the originating committee and from the second committee were not significantly related to one another ($\kappa = -0.04$, $P = .32$) (see the table on this page). This absence of a relation was found not only across the full set of 150 protocols, but for relatively invasive research involving procedures such as electric shock, food or water deprivation, surgery, and drug or alcohol research ($n = 111$, $\kappa = -0.05$, $P = .24$); for protocols involving euthanasia ($n = 108$, $\kappa = -0.04$, $P = .31$); and for protocols in which the reviewing IACUC expected animals to experience a significant amount of pain ($n = 70$, $\kappa = -0.05$, $P = .31$). Thus, regardless of whether the research involved terminal or painful procedures, IACUC protocol reviews did not exceed chance levels of intercommittee agreement.

Of the 118 instances in which the two committees differed in their protocol reviews (79% of all reviews), the second committee was more negative than the first 101 times. Indeed, the second committee rarely rated all dimensions of a protocol favorably (see the table on page 609). For example, only 43% of protocols were seen as having a fairly or completely convincing justification for the type and number of animals used [a requirement of the Animal Welfare Act (*10*)], and only 45% were rated as having good or excellent research designs and procedures. All told, 61% of protocols were judged as either "not very understandable" or "not understandable at all," as having "poor" research designs and procedures, or as justifying the type and number of animals in a way that was deemed "not very convincing" or "not convincing at all." Moreover, these ratings were directly related to protocol recommendations. Regression analyses using the dimensions in the table on the next page found that these factors accounted for nearly half the variance in protocol recommendations made by the second committee (adjusted $R^2 = 0.461$, $P < .001$).

Given the greater negativity of judgments in the second protocol review, it is possible that low intercommittee reliability arose from procedural differences between the first and second reviews. For instance,

S. Plous is in the Department of Psychology, Wesleyan University, Middletown, CT 06459–0408, USA. H. Herzog is in the Department of Psychology, Western Carolina University, Cullowhee, NC 28723, USA.

*To whom correspondence should be addressed. E-mail: splous@wesleyan.edu

low reliability might have resulted from protocols receiving greater scrutiny during the second review than the first. Or low reliability might have been due to the originating committee's relying on its knowledge of who the investigators were (something the second committee was unable to do during its masked review). On the other hand, these explanations for unreliability are less plausible if low interrater agreement exists among members of the same IACUC, because members of the same committee reviewed the protocols under identical conditions.

To explore this issue, we calculated the intraclass correlation coefficient for IACUC members' protocol recommendations made during the second review. The resulting coefficient was 0.28 ($P < .001$), a figure generally considered to be in the "poor" range of interrater agreement (*11*). This level of interrater agreement is comparable to levels found in research on manuscript and grant reviewing (*8*), and it suggests that low intercommittee agreement among IACUCs is not simply the result of procedural differences between the original and second reviews. Rather, the observed lack of agreement appears to be taking place at the individual level (*12*).

We also calculated the intraclass correlation coefficient for each dimension listed in see the table on this page. Here, too, the reliability of judgments fell into the "poor" category, with one notable exception: ratings of the pain or stress animals were expected to experience. The intraclass correlation coefficient for this rating was 0.59, compared with 0.23 to 0.28 for all other ratings (in all cases, $P < .001$). These findings demonstrate that when IACUC members are given detailed classification criteria (in this case, a pain scale), they can achieve a relatively high degree of interrater agreement. At the same time, the results indicate that in the absence of such criteria, interrater agreement among IACUC members will be low even when the same rating dimensions are used to judge identical protocols.

As others have noted (*13, 14*), the regulatory structure of human and animal research depends upon the ability of IACUCs and Institutional Review Boards (IRBs) to make reliable judgments about which research to approve and which to disapprove. Our findings suggest, however, that IACUC protocol recommendations exhibit low interrater agreement. While it is possible that these results are a function of differences between normal IACUC reviewing practices and the reviewing that took place in our study, this explanation cannot fully account for the results. Even when members of the same IACUC rated protocols under identical conditions, their judgments differed from one another. Furthermore, the rating dimensions we used represent key aspects of the protocol review process (e.g., justification for the number and type of animals in the study). Thus, to the extent that unreliability arose from a failure to consider these dimensions during the original protocol review, these results become even more serious. Only 2% of the animal research protocols submitted to us had been disapproved by the original IACUC; in the context of low interrater agreement, this base rate implies that IACUCs will rarely disapprove of protocols that other committees feel should be rejected.

Several authors have proposed techniques to improve reliability in the peer-review process, and recent studies have found that reliability can be significantly increased with procedures such as enhanced reviewer training, standardization of the review process, development of specific evaluative criteria, decomposition of global ratings into smaller categories, and averaging across multiple judgments (*15, 16*). If the IACUC protocol review process is to remain a credible and effective component in the regulation of animal research, the adoption of such techniques may be of considerable value.

PROTOCOL ATTRIBUTE RATINGS AND THEIR RELATIONSHIP TO APPROVAL RECOMMENDATIONS

	Frequency[1]	β[2]
Quality of research design and procedures		0.425**
Excellent	13 (8.7)	
Good	54 (36.0)	
Fair	39 (26.0)	
Poor	27 (18.0)	
Can't say/not sure	17 (11.3)	
Clarity of research proposal		0.297**
Completely understandable	22 (14.7)	
Generally understandable	75 (50.0)	
Not very understandable	43 (28.7)	
Not understandable at all	10 (6.7)	
Can't say/not sure	0 (0.0)	
Justification for type and number of animals		0.167*
Completely convincing	23 (15.3)	
Fairly convincing	42 (28.0)	
Not very convincing	43 (28.7)	
Not convincing at all	33 (22.0)	
Can't say/Not sure/Isn't addressed	9 (6.0)	
Rating of scientific (basic research) value		–0.121
Extremely valuable	7 (4.7)	
Very valuable	37 (24.7)	
Somewhat valuable	59 (39.3)	
Not too valuable	32 (21.3)	
Not valuable at all	15 (10.0)	
Pain scale classification		0.068
I. Experiments involving either no living materials, live isolates, simple invertebrate species, or unobtrusive observations	11 (7.3)	
II. Experiments that involve complex invertebrates or vertebrates but cause little or no pain or stress	35 (23.3)	
III. Experiments that cause minor pain or stress to vertebrate species	34 (22.7)	
IV. Experiments that involve significant pain or stress to vertebrate species	64 (42.7)	
V. Experiments that involve intolerable pain or stress to vertebrate species	6 (4.0)	
Rating of clinical and applied value		–0.006
Extremely valuable	7 (4.7)	
Very valuable	28 (18.7)	
Somewhat valuable	55 (36.7)	
Not too valuable	33 (22.0)	
Not valuable at all	27 (18.0)	

[1]Numbers in parentheses indicate column percentages. [2]Standardized regression coefficients using all variables in column 1 to predict protocol approval recommendations (excluding 14 cases in which IACUCs answered "Can't say/Not sure" on one or more items). *$P = .068$. **$P \leq .001$.

References and Notes

1. R. Dresser, *JAVMA (J. Am. Vet. Med. Assoc.)* **194**, 1184 (1989).
2. Based on the Carnegie Foundation Classification of Institutions of Higher Education.
3. Research protocols covered areas such as behavioral neuroscience, psychopharmacology, animal cognition and perception, comparative psychology, and ethology (for details, see www.sciencemag.org/cgi/content/full/293/5530/608/DC1).
4. 9 Code of Federal Regulations, Subpart C (§2.31, Animals and Animal Products, 1 January 2000 ed.).
5. *Guidelines for Ethical Conduct in the Care and Use of Animals* (American Psychological Association, Washington, DC, 1992).
6. *Guide for the Care and Use of Laboratory Animals* (National Research Council, Washington, DC, 1996).
7. Based on the pain scale in S. Plous, *Am. Psychol.* **51**, 1167 (1996). A copy of the pain scale is posted on www.sciencemag.org/cgi/content/full/293/5530/608/DC1
8. D. V. Cicchetti, *Behav. Brain Sci.* **14**, 119 (1991).
9. J. J. Bartko, W. T. Carpenter, *J. Nerv. Ment. Dis.* **163**, 307 (1976).
10. U.S. Animal Welfare Act, 7 U.S.C. §§ 2131–2157 (1966; as amended in 1970, 1976, 1985, and 1990).
11. D. V. Cicchetti, D. Showalter, *Educ. Psychol. Meas.* **48**, 717 (1988).
12. Because each IACUC member evaluated three protocols, the full set of 150 protocol recommendations does not represent a sample of statistically independent ratings. We therefore ran a check on all reliability analyses to ensure that our conclusions were unchanged when ratings of only the first protocol of three were analyzed. The effect of this analysis was to decrease the observed level of interrater agreement even further. For example, the intraclass correlation coefficient for recommendations on the first protocol reviewed by each IACUC member (50 protocols in all) was 0.16. Analyses of second and third ratings showed no linear trends or practice effects.
13. J. Goldman, M. D. Katz, *JAMA* **248**, 197 (1982).
14. E. D. Prentice, D. A. Crouse, R. W. Rings, *Invest. Radiol.* **25**, 271 (1990).
15. J. Strayhorn Jr., J. F. McDermott Jr., P. Tanguay, *Am. J. Psychiatry* **150**, 947 (1993).
16. R. A. Jako, K. R. Murphy, *J. Appl. Psychol.* **75**, 500 (1990).

17. We thank R. Luyster, B. Wadler, J. De Rivera, C. Carter, R. Boutin, J. Bonds, and T. Dale for their research assistance, D. Cicchetti for statistical consultation, and G. Borkowski for advice on survey design. Supported by NSF grant SBR-9616801.