# Assessing Teaching and Learning in the Responsible Conduct of Research

Kenneth D. Pimple, Ph.D.

October 2001



The author is Director of Teaching Research Ethics Programs at the Poynter Center for the Study of Ethics and American Institutions, Indiana University. He can be reached at 618 East Third Street, Bloomington IN 47405-3602; (812) 856-4986; FAX 855-3315; pimple@indiana.edu.

This is a background paper prepared for the Institute of Medicine (IOM) Committee on Assessing Integrity in Research Environments. The Committee made use of the information in this report, but no endorsement by IOM or the Committee is implied by that use. 

# 1. Introduction

My charge from the Committee on Assessing Integrity in Research Environments of the National Academies is to perform a literature search on empirical evaluations of pedagogical approaches to teaching the responsible conduct of research, or research ethics. (I use the similar terms "research ethics" and "the responsible conduct of research," or RCR, interchangeably in this report.)

In an earlier literature search for the Committee Pimple 2001, my graduate assistant and I spent nearly four months searching databases and other sources, finding very little of direct relevance to our precise topic. Making a virtue of necessity, I tried to use the publications indirectly relevant to our project to sketch a picture of what we know, what we do not know, and what we might profitably learn about the moral climate of research today. The duration of the current project has been about half as long; accordingly, we adopted a different search strategy (see Section 6 below). We once again found very few publications directly relevant to our topic (see especially Sections 3, 4, and 5) and many related publications that I felt worth including here. I believe that this report discusses the most salient publications, but we have no doubt missed some items. I have not attempted to deal with the many publications on training in biomedical or clinical ethics.

# 2. Background

It is important to include in this report reflections and arguments concerning appropriate and inappropriate ways to think about and approach the evaluation of educational efforts in the responsible conduct of research. For maximum clarity, it will be useful to make explicit a few of my assumptions and beliefs about the assessment of teaching before proceeding to a review of the literature.

## *2.1. Terminology*

In this report, I often use the word science as a shorthand for rigorous research. Many humanists and others who might not consider themselves scientists (or be considered scientists

by some standards) nevertheless undertake rigorous research that demands attention to ethical implications.

I make no distinction between training and education.

I use the word student broadly to include undergraduate and graduate students, but also junior and senior researchers, research technicians, research administrators, and anyone else, at any career level, who undertakes training or education in the responsible conduct of research.

I use the word assessment to refer to efforts to determine what knowledge or skills students acquire through educational efforts.

In contrast, I use evaluation to refer to efforts to determine the general effectiveness of an educational unit. Educational units come in many sizes; I refer to the smallest educational unit as a module, i.e., a single reading or exercise that can be described succinctly and incorporated easily into an existing curriculum (see Section 3). Larger units include two or more related modules in a course, an entire course (see Section 4), an entire minor or major, a core curriculum, and so on.

An incomplete list of venues for RCR training will make the breadth of possibilities clear. Education in research ethics can be offered

- As part of required undergraduate courses for students not majoring in science.
- As the focus of senior-level undergraduate courses for science majors (capstone courses).
- As part of introductory graduate courses, such as methods courses.
- As the focus of a full course for graduate students, with a disciplinary (e.g., biology) or topical but interdisciplinary (e.g., human subjects) focus.
- As an occasional part of laboratory meetings.
- As an occasional part of departmental seminars.
- As an informal series of meetings for graduate students and faculty members (brown bag lunches).
- As an occasional campus-wide lecture by an outside speaker.
- As a series of campus-wide presentations on topics cutting across disciplines.
- As an annual half-day, full-day, or multi-day meeting sponsored by the Vice President for Research (or the equivalent, e.g., Provost or Chancellor).
- As a session or forum at a professional meeting.
- As an Internet-based module, tutorial, or seminar.[1]

[1] This list and a few other parts of this paper are adapted from a presentation at the Planning Workshop for a Guide for Teaching Responsible Science, sponsored by the National Academy of Sciences, the National Science Foundation, and the National Institutes of Health, February 1997 and updated several times for use at the Teaching Research Ethics Workshop; see http://php.indiana.edu/~pimple/assessing.pdf.

Note that only the first four of the units listed are relevant only to students traditionally conceived.

I use program evaluation for evaluations of relatively large units, such as a whole course or a given university's IRB training efforts. A good example of program evaluation can be found in Bakken and Reichel 1996.

The line between student assessment and unit evaluation can be blurred. Assessments of particular students can be aggregated and used in a program evaluation. For example, assigning course grades requires some kind of individual assessment, while judging the effectiveness of high schools often uses changes in SAT scores of groups of students. It is important to note that an aggregated evaluation need not depend on particular assessments – an anonymous test or survey, while useless for determining what any given student learned or for assigning grades, can nevertheless be very useful for evaluating and improving a unit. In some venues the use of an anonymous tool is the only alternative to no evaluation at all.

At a certain level of abstraction, student assessment and program evaluation can be considered as if they were one because only a few steps of planning and execution separate them in practice. Indeed, in educational programs designed for adult learners, including researchers and other professionals, program evaluation is often the best measure of learning because professionals are often the best judge of whether a training program meets their needs.

### *2.2. Teaching methods*

People who teach research ethics often encounter resistance. Referring to the NRSA RCR training mandate (see Section 4), Greg A. Sachs and Mark Siegler observe that "faculty and trainee sentiments toward the new regulation were not unanimously favorable. Some saw the requirement for ethics and integrity instruction as yet another infringement on scientist's autonomy. . . . [Others] viewed the scientific integrity program as it was being planned as likely to be a waste of time" (Sachs and Siegler 1993:874).

Chemist Jeffery Kovac succinctly describes another common experience:

> When I talk to colleagues about teaching scientific ethics I hear two primary objections: (i) you can't teach ethics; people either are moral or they are not, and (ii) ethics is best learned in the research group; as situations arise the research adviser will either demonstrate the correct behavior or discuss the issues with the group over a cup of coffee. [Kovac 1996:927]

But "scientific ethics," as Kovac calls the field, "refers to standards of behavior that are specific to science" (Kovac 1996:927), the professional ethics of science. Good moral character does not provide an instinctive understanding of how to deal with outliers, and waiting until "situations arise" to discuss the responsible conduct of research does not help students fortunate enough to work in settings where behavior adheres to high standards – where "situations" rarely or never arise. They might not be so fortunate in their next research setting. (See also Sachs and Siegler 1993, especially pp. 874ff.)

In fact, methods available for teaching the responsible conduct of research are manifold. Such teaching can be woven into the texture of every research experience – for example, instruction in the ethical and technical aspects of keeping laboratory notebooks can easily be combined (Davidson, Cate et al. 2000). RCR training can be accomplished by reading and critiquing rules

and regulations, or holding mock IRB or IACUC reviews (Sweet 1999). It is widely believed that discussing fictional or historical case studies is the most effective method (Kovac 1996; Macrina and Munro 1995). Lectures, research assignments, small group work, and any other method of teaching and learning can be adapted for the teaching of research ethics.

Deni Elliott and Judith Stern make an interesting and useful distinction between "pedagogical hopes" and "instructional objectives." As they put it,

> One might have pedagogical hope that one's students become highly ethical practitioners in their careers and become highly ethical people in their private and public lives as well. But, that is not an objective that can guide the teaching plan for a specific class period. [Elliott and Stern 1996:346]

Another way to make the distinction is to think about proximate and ultimate goals. One proximate goal may be to improve our students' moral reasoning ability. An ultimate goal may be to reduce the incidence of misconduct in research, even when we admit that "no one seriously expects a single course to be able to reform a dishonest person who is likely to commit outright fraud" (Sachs and Siegler 1993:874). But progress toward ultimate goals can only be evaluated by social scientists and future historians; no given educator can be expected to answer this kind of question about her or his own classes.

Other realistic goals can also be articulated. Sachs and Siegler offer the following:

> Ethics courses can identify ethical issues that trainees are likely to encounter in their daily experience; can introduce the trainees to analysis of such issues to prepare them to make decisions; and can provide a background set of considerations and an approach to these problems so that each case does not require the trainees to "reinvent the wheel." [Sachs and Siegler 1993:874]

### *2.3. Assessing learning*

As with all effective teaching, a key element in any unit designed to teach research ethics is assessment of student learning. The key questions in assessment are these:

- How can we find out what our students learned?
- How can we find out whether they learned what we intended them to learn?
- How can we find out whether our instructional goals were met?

A great deal is known about assessing learning in general,[2] and many techniques that can be used to assess learning in any field can be used with regard to education in the responsible

[2] I am grateful to Duane Roen, Director of the Center for Learning and Teaching Excellence at Arizona State University for sharing a bibliography of the most important works in the field; see Apple, M. W., Ed. (1995). Review of Research in Education. Washington D.C., The American Educational Research Association.; Centra, J. A. (1993). Reflective Faculty Evaluation: Enhancing Teaching and Learning Effectiveness. San Francisco, Jossey-Bass Publishers.; Cross, K. P. and M. H. Steadman (1996). Classroom Research: Implementing the Scholarship of Teaching. San Francisco, Jossey-Bass Publishers.; Darling-Hammond, L., Ed. (1994). Review of Research in Education. Washington D.C., The American Educational Research Association.; Fisch, L., Ed. (1996). Ethical Dimensions of College and University Teaching. New Directions for Teaching and Learning. San Francisco, Jossey-Bass Publishers.; Katz, D., R. L. Kahn, et al., Eds. (1982). The Study of Organizations: Findings from Field and Laboratory. San Francisco, Jossey-Bass Publishers.; Kezar, A. and P. Eckel, Eds. (2000). Moving Beyond the Gap Between Research and Practice in Higher Education. New Directions for Higher Education. San Francisco, Jossey-

conduct of research as well, when suitably adapted. For example, one reasonable goal of RCR education is to ensure that students understand regulations and policies relevant to their research. Any number of methods can be used to assess learning in this area, including multiple-choice, fill-in-the-blank, or essay exams. Regulations and policies are objective information, and good teachers are already adept at discovering whether students have mastered objective knowledge.

Self-assessment and peer assessment are frequently overlooked approaches. At first glance these methods might seem easy and irresponsible, which they would be if they consisted merely of letting students grade themselves or each other. However, it has been well demonstrated in many settings[3] that well-designed self-assessment, peer assessment, or co-assessment is a highly effective teaching and learning tool Dochy, Segers et al. 1999. I have summarized some of these findings in Section 11.

Indeed, several methods of assessing student learning in research ethics are known to be in use. In their survey of research ethics education offered in four-year medical colleges in the United States, James M. DuBois and colleagues identified a total of six such methods in the 58 schools that provided ethics course syllabi. (There were 87 responses to the 121 surveys circulated.) The mean for individual schools was two methods of assessment (DuBois, Ciesla et al. forthcoming; DuBois, Ciesla et al. forthcoming;). The methods are as follows, listed from the most commonly to the least commonly used: class participation; examination; papers; case analysis; disposition and reactions to others; and journals. No details on the assessment methods are given.

### *2.4. Special challenges in assessing learning in ethics*

In spite of the many points of similarity it shares with other kinds of teaching, educational efforts in the responsible conduct of research undeniably have some unique – or at least distinct – aspects.

Persons who have not taught ethics often balk at the thought of assessing or grading students in an ethics course or students' performance in an ethics module. There are several possible reasons for this. First is the long reign of positivism and the virtual banishment of discussion of morality and ethics from higher education in the United States for most of the 20th century. In the positivist paradigm, values could not be measured or weighed, and therefore values were seen as merely a matter of opinion, not a proper subject for instruction or assessment.

Second is the influence of pluralism in the United States. We are proud of the fact that no one can force religious observance or belief on anyone else. But a belief in religious tolerance can sometimes be overextended and lead to doubts about whether we share any values. While there

---

Bass Publishers.; Pascarella, E. T. and P. T. Terenzini (1991). How College Affects Students: Findings and Insights from Twenty Years of Research. San Francisco, Jossey-Bass Publishers.; Smart, J. C., Ed. (1985). Higher Education: Handbook of Theory and Research. New York, Agathon Press Inc.; Theall, M., Ed. (1999). Motivation from Within: Approaches for Encouraging Faculty and Students to Excel. New Directions for Teaching and Learning. San Francisco, Jossey-Bass Publishers.; and Uhl, N. P., Ed. (1983). Using Research for Strategic Planning. New Directions for Institutional Research. San Francisco, Jossey-Bass Publishers..

[3] I do not know of a case in which self-, peer-, or co-assessment has been evaluated specifically with regard to research ethics.

are certainly areas of disagreement – and some of them frequently capture national headlines – I think it is also clear that there are many values that most Americans share.

Perhaps more subtly, the very importance of morality in everyday life leads many people to doubt whether it can be taught and whether the teaching of ethics can be assessed. We rightly recognize that ethics is an especially important domain of human belief, experience, and behavior, and we tend to think of morality as a central attribute of the self. Moreover, it is hard to think of ethics as being taught or learned because we view morality as a characteristic that one has (like creativity or intelligence) rather than as a skill that one develops (like doing quadratic equations or operating an electron microscope) or as knowledge that one learns (like the Pythagorean theorem or the atomic weight of helium).

Typically, we do not hesitate to grade our students based on their skills or knowledge, but we do not really want to – and we probably should not – grade them on their character. Teaching ethics is not the same as preaching a sermon.

Another way to talk about character is to talk about behavior. Do people act better after RCR training? Some people seem to assume that they should. I was once interviewed by a journalist about research ethics education. Midway through the conversation she brought up the high-profile case of a young researcher who had fabricated data for several publications that were subsequently retracted. She added, in a weighty tone, that he had taken a course in research ethics. I pointed out that he had taken many courses in science. Which had failed, his training in science or in research ethics?

Even researchers who are strongly committed to providing RCR training have to struggle to overcome the tendency to treat eliminating research misconduct as the major goal of the training, as is implied by this statement by Sachs and Siegler: "Although it may be difficult to prove that these teaching efforts will be successful in eliminating or reducing fraud or misconduct, we argue that it is still important for the scientific community to demonstrate its commitment to the responsible conduct of research in the training process" (Sachs and Siegler 1993:871-872). (See Section 4.1 for a description of the program developed by Sachs and Siegler.)

The aforementioned factors can create unreasonably high expectations for instruction in ethics. We do not expect students who are bad writers and have no imagination to become great novelists after one course in creative writing, nor innumerate students to become talented mathematicians after one course in algebra. But this does not imply that creative writing and algebra courses are of no use.

Students do not become novelists, or mathematicians, or ethical researchers, by taking one course. But just one course can help them make progress toward these goals if they have the necessary skills and abilities, and if they are willing to work at it.

Obviously, no teacher can make a novelist out of a student who doesn't want to be a novelist and isn't willing to work at being a novelist. But a good teacher can help even reticent students become better at writing. Likewise, no one who is basically dishonest is likely to be changed by education in the responsible conduct of research. But a good teacher can help students become better at recognizing ethical problems and thinking through solutions – even those students who initially believe that a course on research ethics is a waste of time.

The scientific penchant for quantification is often a stumbling block when researchers think about assessing student learning in research ethics. But humanists have been assessing student learning using qualitative methods for centuries. As the Hastings Center Project on the Teaching of Ethics concluded, "the most appropriate methods for evaluating ethics instruction are the traditional methods used in the humanities, including assessment of the students' ability to (1) understand central concepts, (2) construct coherent moral arguments orally and in writing, and (3) recognize moral problems and examine them in a rational way" (Bulger and Reiser 1993:S8, quoting Callahan and Bok 1980:81).

#### 2.4.1. A qualitative example: Portfolios

Penny J. Gilmer,[4] a chemist at Florida State University, has successfully adapted a qualitative approach often used in education and creative arts programs to teaching research ethics: The use of portfolios, in which students have a number of assignments through the semester, such as writing essays, responding to case studies, collecting relevant clippings from the popular press, or keeping a journal. At the end of the semester (or perhaps more often), they assemble their best efforts in a portfolio, which they submit for grading.

One obvious strength of this method is that it allows students to compensate for their weaknesses by showcasing their true talents. For example, a student who has a hard time writing an analytical response to a case study might excel at spotting ethical issues in the press, or in some other area. Another strength is that students are assessed based on a body of evidence of their learning, rather than on a piecemeal, quiz-by-quiz approach.

### *2.5. The learning environment*

Education in the responsible conduct of research takes place in a teaching and learning environment that sometimes encourages unethical behavior. In an abstract of a study surveying the level of data manipulation to which students taking seven undergraduate biology and chemistry laboratory courses would admit, researchers found that "from 84% to 91% of undergraduate students openly admitted to manipulating data 'almost always' or 'often.' . . . Students reported observing manipulation by others at the same or higher frequencies. Most attributed motivation to the desire for a better grade." The researchers conclude that students cheat so often because they believe "that a 'right' answer exists and that the 'wrong' answer will lead to a lower grade" and suggest that a "redesign of laboratory exercises to stress the scientific method rather than 'cook book' procedures in which students are expected to verify known biological, chemical, or physical laws can eliminate much of this manipulation." I strongly endorse their conclusion: "This study should raise major concerns about the impact of the techniques used in designing and evaluating undergraduate laboratory exercises on the ethical standards of future scientists and physicians" (Davidson, Cate et al. 2000).

In a special supplement to The Hastings Center Report, Elizabeth Heitman reports similar problems in medical education. It is clear "that the broader educational system in which biomedical researchers are trained contributes to the problems that ethics courses are meant to

[4] I could not find a publication by Dr. Gilmer describing her approach. She can be reached at the Department of Chemistry, Florida State University, Tallahassee FL 32306; 850-644-4026; gilmer@sb.fsu.edu.

address" (Heitman 2000:S43). The competitive nature of medical school leads many students to cut corners in their pursuit of success.

> For over a decade, individual reports of fraudulent research and formal studies on misconduct have suggested that trainees' desire to be "successful" often translates into a competitive drive to be "the best." In what has been called "the pre-med syndrome," relatively young students may develop the habit of cheating in order to satisfy the perceived demands of the competitive educational system in which they prepare to become physicians. [Heitman 2000:S43]

Observers of medical education have also debunked the idea that only people with inferior intellects – losers – commit research misconduct. "Ironically, those most at risk of engaging in misconduct appear to be the best and the brightest by standard measures of productivity: students and professionals with reputations for outstanding accomplishment" (Heitman 2000:S44).

While these observations pertain specifically to medical researchers, it seems reasonable to assume that they also hold for other highly-competitive research fields.

### *2.6. Inadequate attention to evaluation and assessment*

A significant impetus for the development of courses and programs in research ethics was the 1990 Federal requirement that "instruction about the responsible conduct of research" be provided to trainees supported by NIH National Research Service Award (NRSA) training grants. Anna C. Mastroianni and Jeffery P. Kahn helpfully provide an analysis of training materials – "syllabi, course outlines, case studies, reading lists, institutional research policies, and other information" – of 45 of approximately 200 NRSA RCR training programs, collected by DHHS in summer 1996 (Mastroianni and Kahn 1998:1250). The analysis was done in 1997, but in spite of being a bit dated, this article provides a useful snapshot of the state of RCR training at that time. (Summaries of articles describing two of these programs can be found in Sections 4.1 and 4.2.)

Among other findings, Mastroianni and Kahn show that evaluation and assessment of NRSA RCR training programs were spotty at best, making it difficult to determine how successful the educational efforts were.

> Even at the few institutions that appear to have had well-integrated education and training programs in RCR without evaluation efforts, it is unclear which teaching and training approaches are most effective. This is borne out by our own experiences, and has been reported by others. The lack of efforts at evaluation may be a function of the fact that the programs were often initiated to fulfill a federal requirement rather than to create comprehensive programs focused on the achievement of core competencies. Without an attempt at evaluation and the generation of baseline data it could create, there is little prospect of knowing what works and what does not or of identifying ways to improve education and training. [Mastroianni and Kahn 1998:1253]

The primary teaching material for "just over half of the programs [Mastroianni and Kahn] reviewed used one or more" of four resources, "which are each marketed as offering complete RCR training without the need to supplement" (Mastroianni and Kahn 1998:1253). Two of the four (Bulger, Heitman et al. 1993 and AAAS 1996) say virtually nothing about assessing student learning; the other two are far from comprehensive.

Scientific integrity: An introductory text with cases provides a good description of how to lead a discussion using case studies, but does not say much about assessing student learning. It does note that the authors have required students "to select four cases and write a response of one

to two type-written single-spaced pages per case. In effect this becomes a 'term paper' upon which part of the course grade can be based" (Macrina 1995:xx). It also offers three surveys on student experiences and opinions (Macrina 1995:241-250).

Teaching the responsible conduct of research through a case study approach: A handbook for instructors offers a 3½ page instrument for evaluating the effectiveness of individual case (Korenman and Shipp 1994:223-226) and a cursory (1½ page) treatment of evaluation, suggesting that "it is useful to assess the course itself" and that

> specific research integrity objectives may be developed for students, laboratories, and the institution as a whole. The role of the educational effort in furthering those objectives can then be evaluated to assess how well the exercise accomplished these goals. Some useful objectives might include:
>
> - For the students
>   - ability to recognize an ethical choice in research
>   - ability to make a principled decision when faced with an ethical choice
>   - sensitization to research integrity issues generally
>   - familiarity with institutional and governmental policies
>   - knowledge of institutional sources of support when faced with an ethical dilemma
>   - awareness of professionalism in science and its implications
> - For departments and research laboratories
>   - development of communication lines with trainees
>   - development of policies regarding data, authorship, mentoring, and supervision
>   - development of regular laboratory-sponsored workshops or seminars in research integrity
>   - involvement of investigators at all levels in ethics discussions and training activities
> - For institutions
>   - development and adoption of policies and guidelines in research
>   - fostering a climate that encourages responsible research conduct
>   - expansion and support of educational programs in the responsible conduct of research
>   - attempts to alter institutional goals or practices to reduce pressures to behave unethically
>
> It would be overly optimistic to expect major institutional changes on the basis of one course alone. However, sensitization of large numbers of faculty and students over time, in combination with increasing professional and societal attention, is likely to affect institutions positively. Courses in the responsible conduct of research can play a significant role in this process.
>
> The most important goal of an educational exercise is not the transmission of facts but enhancement of the awareness and problem-solving capability of the student. Therefore, instructors may wish to test the participants, both before the course and at a suitable time interval afterward, to see whether the experience has had any lasting impact. For comparison, it may be appropriate to test those who had no educational intervention at the same time. [Korenman and Shipp 1994:9-10]

Please note that my charge requires me to draw attention to the scanty consideration of assessment and evaluation in these works, and a review of their many strengths is beyond the scope of this report.

### *2.7. Designing effective training, evaluation, and assessment*

In order to succeed, training, program evaluation, and learning assessment strategies must be designed together. Even experienced research ethics educators have found that using an assessment tool that does not fit well with their curriculum provides disappointing – indeed, probably meaningless – results (see Sections 4.5.1 and 4.7).

Sarah Brown and Michael W. Kalichman designed a survey to expand and improve upon two earlier studies (Kalichman and Friedman 1992 and Eastwood, Derish et al. 1996) that "asked trainees about their history of ethics training and about their willingness to engage in questionable research practices. Neither survey found a positive effect of training on willingness to engage in questionable practices." Even though Kalichman was an author of one of these surveys, Brown and Kalichman conclude that the lack of a finding "may be due to the design of the surveys, not the failure of ethics training" (Brown and Kalichman 1998:488).

It is not clear that the survey by Brown and Kalichman was an improvement over the earlier two, however. It had a response rate of only 56% (283 replies out of 505 anonymous questionnaires distributed) and, like the two previous studies, found that "perceptions of standards were not significantly affected by hours spent in informal discussions about research ethics, in attending courses on research ethics, or in discussions of case studies." There was one new finding: "Self-reported knowledge of options for facing research ethics problems was significantly increased in association with increased hours of discussion, class time, or case study discussion." Brown and Kalichman emphasize the importance of good design in assessing RCR education: "This study emphasizes the need for increased attention to the definition and assessment of the goals of research ethics training" (Brown and Kalichman 1998:487).

Conversely, a well-designed and carefully executed program evaluation can show that the program itself is lacking, as reported in an abstract of a survey of attendees at a three-day workshop on research ethics for graduate students. The hypothesized outcomes – "greater emphasis placed on the importance of integrity in research methodologies and enhanced awareness of proper ethical procedures and standards when conducting research on human subjects" – were not supported. The author concludes that "the structure and presentation of the workshops needed to be reviewed due to the overwhelming negative responses of the graduate students attending" (Ayscue 2000).

Although a number of tools for assessing learning in research ethics are available, more are needed. In an abstract of a literature review and informal survey to identify "aspects of the teaching and learning of research integrity calling for further assessment research," researchers from the Colorado School of Mines emphasize the importance of "the development of multiple instruments for the assessment of teaching and learning about research integrity" and mention "early draft versions of two new instruments: one utilizing a naive cynicism-idealism attitude

scale, the other focusing on a general knowledge base to integrate science, engineering, ethics, and public policy" (Mitcham, Olds et al. 2000).[5]

### *2.8. Current RCR training efforts*

It is clear that efforts to provide education in the responsible conduct of research are widespread and varied.

- In an abstract of an analysis of "three [Internet-based] tools designed to support a comprehensive RCR curriculum" (a database of instructional materials; a tutorial on informed consent; and a module on intellectual property), Peggy Sundermeyer describes advantages of using the Internet to supplement RCR training as providing "more time for conceptual discussions; instant access to relevant materials; confidentiality of self-assessment; pacing and sequencing of material as well as time and place controlled by the learner; variation in formats to lighten straight passages of text; ease of updating when policies or federal regulations change; access to additional, more in-depth reference materials for further study; and exchanging and sharing information within and between institutions" (Sundermeyer 2000).
- Risa P. Hayes and colleagues describe "the usefulness of an open-ended case analysis test instrument for evaluating the effects of a one-year ethics course on medical students' decision-making skills." The course is a case-oriented seminar for third-year medical students. Before the course, students are given a test case "and asked to provide a line of reasoning for their clinical decisions." The students respond to the same case after the course. "Content analysis of pre- and postcourse responses of a random student sample revealed increases in student awareness in the following areas: 1) consideration of informed consent, 2) professional liability, 3) physician-assisted suicide, and 4) resource utilization. With some modifications, open-ended case analysis holds promise for evaluating medical ethics courses" (Hayes, Stoudemire et al. 1999:284). The instrument, with its emphasis on clinical decision-making, would have to be adapted for use in research ethics.

---

[5] This abstract of a conference presentation seems rather distinct from the paper that will appear in the conference proceedings (Mitcham, C., B. M. Olds, et al. (forthcoming). A plea for pursuing new dimensions of assessment in the teaching and learning of research integrity. Investigating Research Integrity: Proceedings of the First ORI Research Conference on Research Integrity. N. H. Steneck and M. D. Scheetz. Washington, D.C., Office of Research Integrity, National Institutes of Health.); for example, the latter does not make it clear that the authors are developing assessment instruments themselves. It is highly critical of the current state of affairs in RCR training and assessment, including (a) the dominance of biomedical research ethics; (b) the emphasis on "internalist" rather than "externalist" concerns – "issues concerned with doing things right crowd out all discussions about what might be the right things to do; process overshadows substance;" (c) the lack of evidence that RCR instruction reduces misconduct and the "scant agreement even on the immediate goals of RCR teaching and learning;" and (d) the focus moral reasoning (see Section 5).

Although the authors make a decent case for some of their criticisms, their suggestions are not as well supported. For example, they ask whether "research ethics need[s] to be conceptualized as distinct from engineering ethics, as it has been so far," and suggest that "scientists have something to learn from engineers regarding ethics." They point out that "long before scientists, engineers formulated ethics codes at the beginning of the 20th century," but they say nothing about what scientists stand to learn from engineers.

- Gerlinde Sponholz describes initial efforts to develop a teaching program in research ethics at the Universities of Ulm and Marburg in Germany using a case-based approach, including a brief description of their plans to develop instruments for evaluation (Sponholz 2000:513).
- Laura M. Barden and colleagues briefly describe a project in which case studies were used to teach research ethics to high school students, concluding that "the case method is an effective technique for discussing scientific ethics with high school science students" (Barden, Frase et al. 1997:14).

The next two sections of this report summarize published descriptions of five modules (Section 3) and eight courses or programs (Section 4) in the responsible conduct of research. The depth of detail on assessment methods is not uniform across these publications; some are sketchy on assessment while others describe controlled studies designed to determine empirically the effect of a given method of RCR training.

# 3. RCR modules

As described above (Section 2.1), I use module to indicate a single reading or exercise that can be described succinctly and incorporated easily into an existing curriculum.

## *3.1. Trimming, forging, and cooking in chemistry*

Chemist Paul Treichel provides detailed description of a case study exercise that he successfully used to encourage students to determine which of three examples of data manipulation is worst. The three one-paragraph cases, which are provided in the text and are specific to chemistry, describe the adventures of two fictional undergraduate chemistry students who engage in trimming (dropping outlying data points), forging (fabricating data), and cooking (altering data, in this case to adjust for a probable discrepancy between two experimental instruments). Treichel provides criteria for grading students' written responses to the cases, as well as the text of two responses written in iambic pentameter, demonstrating that teaching and learning research ethics can be fun (Treichel 1999).

## *3.2. Deception in psychology research*

Psychologist Bernard C. Beins describes a module intended to demonstrate the effects and ethical issues involved in psychology research in which subjects are deceived. He provides the example of a psychology methods class in which 39 undergraduate students "completed a 20-item bogus personality inventory, the Quacksalber Personality Inventory for Normal Populations. They subsequently received interpretations that were identical for all students. All feedback statements were intended to be neutral or mildly positive" (Beins 1993:33). Students were asked to rate how well the test described their personality and "how useful the test would be in five situations: personal adjustment, employment screening, assessment of honesty, identification of a person's minor problems, and identification of a person's major problems" (Beins 1993:34).

Afterward, the professor revealed the deception and asked the students how they felt about it; most students felt gullible and embarrassed because they were fooled by the bogus inventory. In

the next class meeting, students "wrote answers to questions about the suitability of this exercise to illustrate relevant points about deception in research and whether this demonstration should be repeated in future classes. . . . Of the 31[6] students who commented anonymously about whether this demonstration was effective in teaching about both the Barnum effect and deception, 30 students responded affirmatively" and only one objected to the exercise (Beins 1993:34). All but the one dissenting student thought that the pedagogical value of experiencing deception first-hand and thereby gaining a unique insight into the negative effects outweighed the slight, temporary harm of embarrassment most of them experienced.

### *3.3. Mock IRB review for sociology students*

Sociologist Stephen Sweet describes an exercise used to "teach undergraduate [sociology] students professional ethics standards, reasoning skills, and the role that institutional review boards (IRBs) play in social science research" (Sweet 1999:55). Before class, students are asked to read the American Sociological Association's Code of Ethics and two chapters on ethics and research. In the class, desks or seats are arranged in two groups with space between; one group of chairs is for students who support a study, the other for those who oppose it. The instructor provides an introduction to the role of the IRB and explains to students that they will be discussing and judging three cases. When students have read a case, they seat themselves appropriately to indicate their initial support for, or opposition to, the case; they are free to change places as they change their minds.

The three cases, provided in the article, are very loosely based on actual sociology studies. The first involves deceiving students as part of an experiment in educational psychology; the second is about maintaining confidentiality of subjects involved in illegal activities; and the third concerns analyzing data derived from Nazi experiments. Student discussion of the first two cases tends to take a consequentialist approach, whereas at least some students take a deontological stand when discussing the third case, arguing that the data from experiments involving torture and murder should not be used, no matter what good may come of it.

"Students find the classroom encounter stimulating and enjoyable. This is evident in the passionate debates that erupt during the exercise and in the comments individual students offer after class. The majority of students demonstrate the ability to apply the ethical standards of the American Sociological Association and describe the functions of an institutional review board on both essay and multiple-choice questions" (Sweet 1999:58).

### *3.4. Defending the ethics of questionable psychology studies*

Psychologists David B. Strohmez and Anne A. Skleder describe a controlled study modeled after Rosnow 1990 in which upper-division undergraduate students majoring in psychology and enrolled in seven sections of a required methods course were asked to find "a recently published study that they consider to be unethical." The students were directed to "read the study carefully and to be prepared to present it during the next class" (Strohmez and Skleder 1992:106).

[6] Apparently not all of the 39 students who completed the bogus inventory made comments on the effectiveness and appropriateness of the exercise (KDP note).

In class, students "rated the cost and utility of the selected studies on a scale ranging from *no cost or utility* (0) to *highest cost or utility* (100)." In the experimental sections, a student was then asked to "imagine himself or herself as the article's primary author or researcher" and defend the study before a peer review board, played by the rest of the class. After the role-play, the students rated the studies again (Strohmez and Skleder 1992:107).

The results supported the researchers' hypotheses:

> We predicted that participating in the role-play exercise would increase students' perceptions of the utility of their "unethical" studies. Similarly, we hypothesized that ratings of the perceived ethical costs of these studies would be lower after the role-play. Finally, we predicted that the magnitude of any obtained effects would be larger in these six classes than in the comparison class, which did not use the role-play exercise. [Strohmez and Skleder 1992:106]

Of the five modules described here, this is the only one I find problematic as described in the publication. The abstract to this article states, "Results [of the study] indicate that the exercise can be a valuable tool for sensitizing students to the factors involved in judging the ethics of research" (Strohmez and Skleder 1992:106). It appears to me, however, that the researchers discovered that students can convince themselves and each other that studies they had previously thought unethical were actually ethical. I suspect that a parallel study could show that students can convince themselves and each other that studies they had previously thought ethical were actually unethical. The authors offer no comment on the actual merits of the studies; the only reason readers are given to believe that the studies selected were ethical is that they were published. Publication is, of course, no guarantee of high ethical standards; Milgram's studies and reports of the Tuskegee Syphilis Study were published.

### *3.5. Case studies to improve and assess moral reasoning*

Moral Reasoning in Scientific Research: Cases and Materials (Bebeau, Pimple et al. 1995) is an 80-page booklet featuring six one- to two-page case studies as well as extensive information on how to use the cases and a discussion of the theoretical underpinnings of the approach. It is somewhat more than a module but less than a course or program, but I think it fits better here than in Section 4. I hope my judgment can be trusted, since I directed the project for which the booklet was developed and worked closely with Muriel Bebeau in its design.[7]

As its title implies, Moral Reasoning in Scientific Research is designed to facilitate improvement in moral reasoning skills (see Section 5), as well as assessment of such improvement. I believe that it is an effective tool (for evidence, see Section 10); it has certainly been popular. We distributed approximately 300 copies of the booklet in 1995 and 1996. We did not keep careful track of the number of copies distributed between 1996 and 1999, but a reasonable estimate would be another 300. In May 1999 I converted the booklet to PDF format and made it available via the Poynter Center's World Wide Web site. I made no particular effort to advertise its availability there, but in May the file was accessed 63 times, and by October it had over 500 hits. As of September 2001, the file has been accessed over 13,000 times (see table).

[7] I readily acknowledge that my intimate relationship with the Moral Reasoning booklet probably compromises my objectivity on the subject. However, it does seem to be a significant tool, and the Committee's charge to include the Poynter Center's experiences in this report made it clear that I should at least mention the booklet here.

| Moral Reasoning hits | |
|---|---|
| May-Dec 1999 | 2,288 |
| Jan-Dec 2000 | 4,578 |
| Jan-Sep 2001 | 6,789 |
| *Total* | 13,655 |

I have used several of the case studies with a total of 38 students in an NSF-funded Research Experience for Undergraduates (REU) program at the Center for the Integrative Study of Animal Behavior at Indiana University-Bloomington over the course of four summers (1996-1998 and 2001). The research ethics component of the REU program consisted of roughly equal measures of case study discussions and other presentations. Given the short duration of the program, we discussed only three or four cases each year. Even so, students consistently showed marked improvement in their analysis of Case 2, and generally showed modest improvement in Case 3 (see table). We did not discuss the same case every year for Case 3, and one year I did not assess the discussion of Case 3. The highest possible score is 20.

| Student scores on case study evaluations | | | |
|---|---|---|---|
| | **Case 1** | **Case 2** | **Case 3*** |
| N | 38 | 37 | 24 |
| Minimum | 7 | 13 | 5 |
| Median | 13 | 19 | 18 |
| Average | 12.47 | 18.27 | 17.83 |
| Maximum | 17 | 20 | 20 |

Information on Moral Reasoning, and the booklet itself in PDF format, can be found on the Poynter Center's World Wide Web site at http://poynter.indiana.edu/mr-main.html.

# 4. RCR courses and programs

As mentioned in Section 2.6, the number of courses and programs devoted to RCR education expanded dramatically following the mandate by NIH that all trainees funded under the National Research Service Award (NRSA) training grant program receive training in the responsible conduct of research. This section begins with summaries of reports on two such programs and continues with several other courses.

## *4.1. NRSA training grant RCR program, University of Chicago*

Greg A. Sachs and Mark Siegler describe the approach of the University of Chicago to fulfilling the NRSA RCR mandate. The University developed a two-year program featuring "lectures the first year and seminars the next." Instead of providing a short, one- or two-day workshop, they opted to develop an ongoing course. The program "is intended not as an effort to reform dishonest persons who are likely to commit outright fraud but to serve the large group of honest trainees by helping them learn the important ethical issues and norms in the practice of good science, recognize areas of ethical conflict in research and scientific training, and understand their own values better" (Sachs and Siegler 1993:871).

The lecture portion consisted of "seven monthly lectures, each followed by a panel response to the presentations." A syllabus and readings were distributed to participants at the beginning of the course. Lecture topics were

- government concerns with integrity and misconduct in science;
- whether responsible conduct in science can be taught;
- the evolution of policies for protecting human research subjects;
- the procedures of the university's Committee on Academic Fraud;
- ethical issues in scientific publishing;
- how several laboratory chiefs deal with issues of scientific responsibility and the handling of scientific data; and
- ethical and policy concerns in using animals in research. [Sachs and Siegler 1993:872]

The seminar portion consisted of "four two-hour seminars held every other month during the academic year. Each seminar was designed to involve approximately 25 scientists-in-training and two faculty facilitators," one bench or clinical scientist and one ethicist. Seminar topics were

- scientific fraud and misconduct,
- laboratory supervision and control of data,
- publication and reviewing practices, and
- societal concerns about research materials (genes, animals, and human subjects), [Sachs and Siegler 1993:873]

The article includes extensive discussion of logistical concerns, including the importance of securing support from deans, department heads, and other influential figures.

Even though Sachs and Siegler embraced the NRSA mandate as "an opportunity rather than as a burden" – and encouraged the scientific community at large to see it in the same light (Sachs and Siegler 1993:871) – they did not find the problems of evaluation and assessment easily solved. "One of the biggest questions we currently are struggling with is one that confronts any new teaching or training program: how can we evaluate our program? Cases of scientific fraud leading to formal or informal investigations are fortunately rare, so it is unlikely that we will detect any behavioral change from having students take our course." They found testing in this area to be difficult, but they did undertake some program evaluation: "Students' satisfaction and student and faculty feedback regarding various elements of the program were elicited" (Sachs and Siegler 1993:873).

### *4.2. NRSA training grant RCR program, University of Texas*

Researchers at the Department of Surgical Oncology at the University of Texas M. D. Anderson Cancer Center opted for a significantly different approach, and combined it with an interesting assessment/evaluation effort. Raphael Pollock and colleagues describe a seminar designed for NRSA trainees consisting of "a four-session seminar (6 hr total time) structured around assigned readings, didactic presentations, and group discussions" (Pollock, Curley et al. 1995:247; see also Pollock, Curley et al. 1994 for an earlier, shorter description of the same program).

The authors describe the aim of the course as "develop[ing] the skills of critical ethical judgment as the best protection against the ethical (or unethical) values of others. Such skills are

best developed in a curriculum that actively encourages trainees to scrutinize critically and to justify their own ethical assumptions, as well as those of their peers and mentors" (Pollock, Curley et al. 1995:250). To accomplish this goal, they endorse what they call a "Quandary Ethics," or case study discussion, approach.

Twelve trainees who had completed the seminar and eight who had not (the control group) "answered a 72-item questionnaire of our own design that examined a variety of issues in research ethics." The experimental and control groups showed some similarities in their responses to the questionnaire. For example, "both groups of trainees perceived that too much emphasis was placed on quantity rather than quality of publications," and "both groups felt that this pressure emanated from department chairmen rather than laboratory mentors ($P < 0.0001$)" (Pollock, Curley et al. 1995:247).

There were also differences:

- 66% of the experimental group "considered punishments for research error to be fair," but 66% of the controls "considered punishments for research error too lenient" (Pollock, Curley et al. 1995:248).
- 73% of the experimental group "believed they could define NIH-sanctioned research standards" compared to 17% of controls (Pollock, Curley et al. 1995:249).
- "It was very interesting that although both groups were satisfied with their understanding of how many times to repeat an experiment prior to writing a manuscript, the actual number of repetitions perceived as necessary varied between groups ($P = 0.071$)," with controls thinking that fewer repetitions were necessary (Pollock, Curley et al. 1995:249).
- 87% of the experimental group thought that "sloppy data analysis" would be detected, compared to 27% of controls (Pollock, Curley et al. 1995:249).
- 100% of the experimental group "were satisfied with their level of understanding of requirements for authorship" compared to 62% of controls (Pollock, Curley et al. 1995:249).
- 75% of the experimental group "knew how to proceed if they lacked a sufficient quantity of a reagent critical for experimental data replication" compared to 37% of controls (Pollock, Curley et al. 1995:250).
- 92% of the experimental group "knew how to address problems with discordant or outlier experimental data points" compared to 50% of controls (Pollock, Curley et al. 1995:250).
- 100% of the experimental group "were prepared to seek third party input into an ethical dilemma involving their own work" compared to 37% of controls (Pollock, Curley et al. 1995:250).

The authors conclude that "a short course in ethics of research can successfully provide an 'ethical compass' for surgical-investigator research trainees" (Pollock, Curley et al. 1995:247). Given the responses to their questionnaire, I would say this conclusion is justified, even though it is based on a small sample.

It appears from this article that the 72-item instrument would be a valuable tool for evaluation and assessment. Unfortunately, "the questionnaire has not been retained" (Pollock 2001).

### *4.3. Honors section for chemistry students, University of Michigan*

Chemist Brian Coppola describes his special honors section (160 students) of a normal introductory chemistry course (1000 students). The section features heavy emphasis on discussing case studies and writing case studies with reflection and peer review. Examples of case studies provided by the instructor and written by the students are provided.

Coppola reflects on his own education in research ethics as well. "After participating in the Teaching Research Ethics (TRE) program at the Poynter Center in 1994, I became sensitized to the idea of formal, explicit reasoning as a way to think about moral dilemmas. After 6 years of involving students and student leaders, I have seen the same thing in them that happened to me: the language of ethical decision-making permeates one's normal discourse and affects one's subsequent actions" (Coppola 2000:1507).

### *4.4. "Professional Ethics for Scientists," Towson University*

Chemist Linda Sweeting describes her undergraduate course, "Professional Ethics for Scientists," as having a focus on "healthy science rather than pathology, with ethics presented as the basis of excellence in science." Course prerequisites are "three courses in science, at least two with laboratory, plus college writing." Students majoring in "chemistry, biology, computer science, geology, mathematics, geography, and philosophy" have taken the course (Sweeting 1999:369). Emphasis is on discussion rather than lecture, with the teacher acting as the "guide on the side" rather than "the sage on the stage." Assignments include a weekly journal, at least one take-home essay exam, and two 5-page papers. "The papers are developed by editing and rewriting over the course of several weeks," with feedback from the professor and other students (see Section 11 for more on peer review).

> The two papers are chosen from four types, for which there are ample examples and references in the Web syllabus (and students always find others):
>
> 1. Review of a science-oriented novel, biography, or autobiography, with a focus on the ethical issues. . . .
> 2. Review and summary of a book on some aspect of ethics in science. . . .
> 3. Summary and comment on a recent case of fraud or other misconduct using multiple sources. . . .
> 4. An analysis of the ethical issues in one aspect of scientific endeavor beyond the scope of the course. Students have chosen conflict of interest, thalidomide, euthanasia and the Hippocratic oath, environmental research ethics, animal rights, gender discrimination in science. [Sweeting 1999:371]

The drafts and the papers are graded.

Sweeting notes that "students do not necessarily develop perfect assessments of and responses to the temptations and challenges they meet, any more than they can solve every

chemical equilibrium problem perfectly after their first introductory chemistry course. You should not be discouraged by their, or your, imperfections as ethical scientists" (Sweeting 1999:371).

### *4.5. "The Ethical Dimensions of the Biological Sciences," University of Texas*

Two articles (Bulger and Reiser 1993 and Reiser and Heitman 1993) describe the authors' 10 years of experience in developing, introducing, and providing a course on the responsible conduct of research at the University of Texas-Houston Health Sciences Center. Work on the course included development of a textbook (Bulger, Heitman et al. 1993). The course uses case study discussion to develop "students' analytical skills and moral reasoning" (Reiser and Heitman 1993:877), observing that "cases have been used successfully to analyze ethical issues in the fields of medicine, law, journalism, engineering, and business, and we have found that using them to examine the ethics of science has been equally valuable" (Reiser and Heitman 1993:878).

Assessment included use of a term paper and final exam. The authors provide a helpful discussion of difficulties, both political and logistical, in using a final exam, as well as their strategies for overcoming the problems.

> The introduction of a final examination proved controversial four years ago and remains a source of anxiety for some students. The examination, which covers both the reading and class discussion, is graded on a pass-fail basis and is intended more to make students reflect actively on conceptual issues than to test them on particular facts in the readings. However, most students in the sciences are not accustomed to writing essay examinations, the traditional testing format in ethics, and are insecure about their ability to express themselves in writing. Others anticipate (wrongly) that, as in their basic science courses, they will be graded on their ability to memorize and restate the "facts" of ethics, and are frustrated in their efforts to identify such material. The test focuses on the analysis of scenarios that students are likely to encounter, and asks for students' own assessments of various issues. We provide a study guide, and stress that the examination is comprehensive but not detailed. While a few students have failed the written test, they have passed the course by taking a repeat, oral exam. [Reiser and Heitman 1993:878]

Program evaluation appears to have been an integral part of the course; one article emphasizes the importance of making necessary modifications to the course based on student evaluations and feedback (Reiser and Heitman 1993:878) and the other observes

> After ten years of experience with this course and after receiving both oral and written student evaluations over that time, we believe that a systematic effort to explore the process of scientific discovery and the life of a scientist in the context of ethics has been well accepted by the students and has been useful to their development. [Bulger and Reiser 1993:S6]

Given these comments and my own review of the course's textbook (Pimple 1994), I feel certain that this course is successful, which makes a study, discussed in the next section, of its effectiveness in developing moral reasoning an important example of the challenges involved in course evaluation and student assessment.

#### 4.5.1. DIT fails to detect improvement in moral reasoning

Before they undertook a controlled study on moral development, the course faculty felt that the efforts were likely to improve their students' moral reasoning skills: "Although we made no

attempt to measure changes in the levels of moral reasoning in our students, the way our course was structured, with much of the time devoted to student interchange, was consonant with the development of moral reasoning skills in the students" (Bulger and Reiser 1993:S6).

The controlled study, however, failed to support this assumption. In an abstract of the study, the course was evaluated using the Defining Issues Test (see Section 5.2) as a pre- and post-test measure. The study failed to measure any change in moral reasoning skills, suggesting "a need for more careful definition of specific goals, content, and methods" (Heitman, Salis et al. 2000). Again, having every other reason to believe the course is well-designed and well-delivered, I take the results of this study to indicate that the instrument was not well suited to the course. The faculty could decide with equal validity either (a) to find or develop an instrument with a better fit to their actual efforts, or (b) to adjust their course to be more effective in developing students' moral reasoning skills.

### *4.6. Short course in research ethics, Florida International University*

Physicist Bernard Gerstman describes his course for graduate and undergraduate science students, a maximum of 15 students. The course meets for one hour each week; the article does not stipulate the duration of the course, but does mention that it could be compressed into two weeks.

The course is divided into three sections. In the first section, the faculty member provides "a short (one hour) summary of the general principles of ethics in Western society, which can then be used as the basis for the principles of research integrity and ethics." The second section concentrates on "faculty led discussions of selected reading material on recent cases concerning violations of research integrity."

The first two sections lay the groundwork for the "crucially important" third section, in which each student is required "to make a half-hour presentation to the class about a case of suspected unethical behavior in research that they have investigated through a literature search."

> The students are enthusiastic about making their presentation and peer pressure motivates them to do a thorough job. The presentation forces the students to "step into the mind" of a scientist who is behaving unethically. This obliges them to confront the temptations to behave unethically and solidifies the need for self-vigilance. [Gerstman forthcoming]

Gerstman has not done a formal study of the course's effectiveness, but his experience shows that it is successful.

> The effectiveness of the course is based upon my observations of the increase displayed by the students as the course progresses in terms of their awareness and thoughtfulness of ethical questions. At the beginning of the course, many students consider unethical behavior to consist only of fabricating data; by the end of the course they are aware of the myriad aspects of unethical behavior and the danger of a slippery slope developing from what they originally considered merely cutting corners. [Gerstman 2001]

### *4.7. Pilot seminar for graduate students, Dartmouth College*

A particularly useful report by Deni Elliott and Judith Stern describes another pretest-posttest assessment method, including an unsuccessful first attempt and a more successful second effort

at assessing student learning in a graduate-level seminar in academic research ethics at Dartmouth College. The course was developed by a team of faculty members and supported by grants from the Department of Education's Fund for the Improvement of Postsecondary Education (FIPSE) and by the National Science Foundation.

The instructors tried to assess both what the students learned (in terms of both skills and content) and the learning environment of the classroom. The latter is important because

> Learning applied ethics requires students to take intellectual risks; it requires students to give close examination to their beliefs, values and methods of thinking about adequate and inadequate professional behavior. Clearly, some environments encourage this kind of risk taking and other environments discourage it. [Elliott and Stern 1996:352-353]

The team's attempts to measure the learning environment included development of a fairly elaborate instrument, which was ultimately deemed unsuccessful; they concluded that the use of a standard student evaluation is adequate for this purpose.

Elliott and Stern point out that a prerequisite for evaluating the success of a course is a clear understanding of the instructional objectives of the course, and their goals are particularly well-stated. At the end of the course, students will

1. be able to clearly describe relevant scientific conventions including: laboratory practice, institutional responsibility, etc.;
2. be able to describe what leads to ethical problems including causes inherent in the social context of the practice of science;
3. be able to identify ideal scientific practice and consider how to bring scientific conventions more in line with the ideal;
4. be able to separate behaviors into four categories: morally prohibited, required, permitted, and encouraged, thus illustrating an understanding of the role of the scientist in society. [Elliott and Stern 1996:349]

In their initial assessment plan, the team of instructors made extensive efforts to assure the validity and reliability of the pretest-posttest assignment, including the use of three outside scorers and a fourth outsider who compiled and analyzed the results. The test required students to read an edited article on an actual case of research misconduct (included as an appendix to the article) taken from the journal Science and answer the following question:

> Identify the ethics problems in this case. Discuss what the individuals involved did right. Discuss what the individuals involved could have or should have done differently. [Elliott and Stern 1996:350]

The results were disappointing. There was no inter-rater reliability, and "there was no significant difference between how students approached the vignette at the beginning of the term and how they approached it at the end." The authors identify several reasons for this failure.

> We realized that even if students had learned the material they had not been encouraged to express what they had learned [because] the post-tests were not graded. . . . Few students made any attempt to integrate the three questions and to evaluate the responsibilities of individual moral agents as complex people. . . . We did not provide the students with any explicit instruction in conducting systematic moral analysis. We instead expected students to intuit the process by examining a series of cases. [Elliott and Stern 1996:350-351]

Given the mismatch between the course content and the assessment plan, it should not have been a surprise that the pretest-posttest model failed. Since the students were in no way prepared to write an explicit moral analysis, it is not surprising that they failed to rise to the occasion when they were asked to do so at the end of the course. Moral analysis is one skill; expressing moral analysis in writing is another.

After this initial failure, the instructors tried a different method for assessing student learning. Instead of replicating the pretest exactly, they asked their students at the end of the course to do a meta-analysis of their pretest.

> 1. We provided [three] short vignettes. . . . Two contained issues of ethical importance (from the faculty's perspective) and one did not.
> 2. More explicit instructions were given for the pre-test and students were told that a "high quality" response to the pre-test/post-test was necessary to receive a "Pass" in the class.
> 3. At the time of the post-test, students received back their pre-tests, with instructions to analyze how well they had responded to the pre-test. [Elliott and Stern 1996:351-352]

The vignettes and the instructions are included as appendices to the article. Unfortunately, the authors do not include item #2 (the more explicit instructions for the pre-test). They do provide the instructions for the post-test:

> The purpose of this final exam is to help assess what difference this class has made in the way that you think, dealing with ethical problems.
>
> The diagnostic test that you completed at the beginning of the term is attached.
>
> 1. Please review the case, the instructions you received at the beginning of term and your responses.
> 2. Analyze your initial response. Describe how your thinking has changed. Be sure to discuss understandings or information that you have now that you didn't have at the beginning of the term.
>
> This is your opportunity to consider how your thinking has changed. Please notice changes in HOW you think as well as any changes in WHAT you think. It may be that you reach the same conclusion now than you did in the beginning of the term, but that you think about the situation in a different way.
>
> 3. Please attach your diagnostic test to the final exam.
>
> Please keep in mind that you are NOT being asked to repeat the assignment from the beginning of the term. You are being asked to analyze how you initially responded to that assignment. [Elliott and Stern 1996:362, emphasis in original]

The authors describe the new pretest-posttest method as a success, but note, "As we did not involve external evaluators, the report of the results are anecdotal but compelling" (Elliott and Stern 1996:352). A similar approach is described in Section 5.4.

The pilot program also resulted in two books (Elliott and Stern 1997 and Stern and Elliott 1997).

### *4.8. "Ethics-enhanced" curricula in psychology, Fordham University and Loyola University-Chicago*

The most extensive effort at evaluating a research ethics curriculum is described by Celia B. Fisher and Tara Kuther. Their article describes a study in which six case studies in research ethics were integrated into introductory psychology courses by faculty members at two universities (Fordham and Loyola-Chicago). The case studies, designed by Fisher, drew on well-known empirical psychology studies. Each case highlighted at least one ethical issue in psychology research: (a) crises in public places staged by social psychologists; (b) animal experimentation; (c) "the use of aversive procedures with human participants;" (d) socially sensitive research; (e) deception research; and (f) "randomized clinical trial (RCT) research with participants with psychological disorders" (Fisher and Kuther 1997:172).

Each case comprised a brief abstract and detailed description of the study and "homework assignments composed of four sets of focus questions requiring students to critically evaluate ethical issues derived from the Belmont Report and the APA Ethics Code" (Fisher and Kuther 1997:173). The focus questions asked about "the scientific validity and social value of the study; . . . potential research risks within the context of the need for experimental control; . . . protections and threats to participant autonomy and privacy; . . . [and] the investigator's dual responsibility to conduct well-controlled experiments and protect participant welfare" (Fisher and Kuther 1997:173). In addition to the cases, which were distributed to students, faculty members received an instructor's guide.

The size of the study is impressive, consisting of "585 students enrolled in a total of 24 introductory psychology sections." The experimental group (half of the sections) "received the ethics-enhanced instruction" while the other half "received standard ethics instruction. . . . Both the enhanced and standard instructional groups received pretest and posttest questionnaires" consisting of "three test vignettes: . . . a deception study, . . . an animal aversive conditioning study, and an RCT study with a nursing home population" (Fisher and Kuther 1997:173). Students answered two questions on each vignette, one on "modifications they would use to protect the welfare and rights of the research participants" and the other on "ethical reasons why they would or would not conduct the study in its original form or with their modifiers" (Fisher and Kuther 1997:173).

The enhanced curriculum was evaluated based on "scores on student essays, student course evaluations, and instructor curriculum evaluations" (Fisher and Kuther 1997:173). Scores on student essays showed that "significant posttest improvement emerged only for students who received the ethics-enhanced instruction" (Fisher and Kuther 1997:173). Both students and faculty members "responded favorably toward the curriculum and judged the instructional and testing materials to be appropriate for introductory psychology students" (Fisher and Kuther 1997:174).

The study demonstrated that

> expanded instruction in the ethics of scientific psychology using the case study method can be easily incorporated into introductory psychology classes. Ethics-enhanced instruction increased student awareness of particular ethical procedures used to protect participant's rights and welfare and to a lesser extent increased student sensitivity to the importance of considering both scientific responsibility and participant welfare in ethical decision making. [Fisher and Kuther 1997:174]

More information on the cases, the instructor's manual, and the student workbook can be requested from Fisher.

# 5. Moral reasoning

The single most thoroughly studied aspect of moral psychology is moral reasoning, starting with the path-breaking work of Lawrence Kohlberg and elaborated into a Four Component Model of Morality by James Rest.[8] Efforts to improve and assess the development of moral reasoning in conjunction with RCR training is much more well-developed than any other single approach, warranting its own section in this report.

## *5.1. Survey of studies on training in professional and research ethics*

Muriel Bebeau, a leader in the field of assessing moral reasoning development and applying the findings of moral psychology research to professional and research ethics, has provided a helpful survey of work in the field. My summary draws on her synthesis of an enormous database on the effects of ethics curricula designed to "promote functional processes that give rise to morality: 1) ethical sensitivity; 2) moral reasoning; 3) moral motivation and commitment; and 4) ethical implementation (Bebeau forthcoming).[9]

Bebeau describes five instruments used to assess these processes. Some of the instruments (e.g., the Dental Ethical Sensitivity Test) are field-specific and would require adaptation to be used in research ethics, while others (e.g., the Defining Issues Test; see Section 5.2) can be used directly for research ethics.

A particularly intriguing instrument is the Professional Role Orientation Inventory (PROI), which "assesses commitment to privilege professional values over personal values." It appears that the lack of a clear concept of the professional's role is an excellent predictor of professional malfeasance.

> The most direct evidence of a relationship between role concept and professionalism comes from the study of performance of the 28 members of the practicing [dental] community, referred for courses in dental ethics because of violations of the dental practice act. Although the practitioners varied considerably on measures of ethical sensitivity, reasoning, and ethical implementation, 27 of 28 were unable to clearly articulate role expectations for a professional. [Bebeau forthcoming]

It is not clear from Bebeau's description whether the PROI is specific to dentistry, nor how difficult it might be to adapt to research ethics.

Bebeau's synthesis gives substantial support to the proposition that attention to these psychological processes in creating curricula and assessment tools is effective.

1. "Striking individual differences among students and practicing professionals on each of the measures" have been shown.

[8] Moral reasoning is only one of James Rest's Four Components; this section is named for the component most widely studied and, apparently, easiest to measure.

[9] I do not include page numbers for quotations because the Proceedings will not have the same pagination as the preprint from which I am working.

2. It is well established that the four processes are functionally independent and "competence on one of the processes does not predict competence on another" – e.g., a person can have a well-developed moral sensitivity but poorly developed moral reasoning skills, and vice versa.
3. All of the processes can be improved by educational interventions and "curricula of rather modest duration can influence performance in measurable ways."
4. Perhaps most heartening to educators, it has been shown "that strengths and weaknesses in each of the processes are linked to real-life ethical behavior." For example, one study showed "a relationship between the number of malpractice claims and moral judgment scores, noting that a high DIT score had a kind of protective effect, insulating one from claims" (Bebeau forthcoming).[10]

Bebeau concludes that "the findings not only support Rest's contention that moral failings can result from deficiencies in one or more of the processes, but support the importance of attending to each when designing curriculum. Further, whether a curriculum promotes ethical development depends on whether that curriculum incorporates the elements of effective instruction" (Bebeau 2000). I have mentioned this last point several times in this report, and I should perhaps note here that my understanding of the importance of carefully integrating teaching and assessment is rooted in my close work with Bebeau in 1994-1996.

### *5.2. The Defining Issues Test*

The Defining Issues Test (DIT) is perhaps the best established tool for assessing moral reasoning ability. Although increasing moral reasoning ability is only one of the many possible goals for RCR training, and should probably not be the only goal of such a curriculum, pedagogical techniques for improving moral reasoning are so well established that, coupled with the assessment power of the DIT and other tools, it seems a very attractive and practical instructional objective.

The DIT, developed by the late James Rest and his colleagues, is a multiple-choice, standardized test designed to measure moral reasoning ability and emulate the lengthy open-ended interview Lawrence Kohlberg used to develop his typology of levels of moral development. Innumerable studies indicate that the DIT succeeds in this quite well.

For all of its strengths, the DIT is not the alpha and omega of RCR training and assessment.

1. Improving moral reasoning skills (and measuring that improvement via the DIT) may not be an appropriate or meaningful goal for all educational efforts in research ethics. For example, I would guess that training for senior researchers would find only modest gains in moral reasoning ability, and training for IRB or IACUC administrators would be better focused on interpreting and applying regulations and policies.
2. To be useful as a measure of training effectiveness, the DIT would have to be used as a pretest-posttest, which may not be effective for a curriculum of a few days or weeks.

[10] The language here seems unfortunate to me, but I cannot be certain that paraphrasing it to something like "practitioners with high DIT scores were less likely to be sued for malpractice" would accurately capture the study's conclusions.

3. One of the strengths of the DIT – it is objective and quantitative – is also a potential pitfall. Since it may seem as if the DIT provides more robust, trustworthy, or "true" results than qualitative methods, it may seem to many researchers that it is the only tool they need, or the only valid tool available. There is no problem with this if a curriculum is called "Moral Reasoning Improvement." But since there is more to research ethics than moral reasoning, any course or program billed as RCR instruction should include assessment for other goals, even if the assessment tools are not as objective or well-established as the DIT.
4. The DIT is only an instrument; it is not a curriculum. It can be used for assessment, but not for teaching.

These points are not meant to detract from the usefulness of the DIT. My impression is that some researchers give the DIT too much weight at the expense of other approaches. The DIT is easily administered and extremely well-validated, though, and can be a useful tool in many RCR programs.[11]

### *5.3. Moral reasoning improvement through medical school*

A number of studies by Donnie J. Self and colleagues, some of them undertaken in response to earlier reports suggesting that the "structure of medical education may serve to inhibit or prevent expected and desirable development of moral reasoning among medical students" (Self and Baldwin 1998:S93), have shown that moral reasoning skills do improve through medical school (Self, Schrader et al. 1991; Baldwin, Daugherty et al. 1991; Self, Olivarez et al. 1998). Furthermore, teaching medical ethics to medical students using a case study approach (Self, Wolinsky et al. 1989) or discussions of short films (Self, Baldwin et al. 1993) increases the students' moral reasoning skills, and using case studies is more effective than using lectures (Self, Wolinsky et al. 1989). Twenty hours of discussion appears to be sufficient to increase moral reasoning ability measurably (Self, Olivarez et al. 1998).

Self is so convinced of the importance of moral reasoning skills generally that he and a colleague have recently suggested that moral reasoning scores could be used as one criterion in selecting medical students and residents (Self and Baldwin 2000).

Although Self's research is on medical ethics rather than research ethics, his work is so extensive and complements Bebeau's so well that I felt it appropriate to include it in this report. Another point of support can be found in a study of the "relationship between ethical dilemma discussion and moral development of ninety-six second-year students" in pharmacy that showed, using the DIT, that "moral reasoning skills are both teachable and measurable, and that ethical dilemma case discussions may enhance moral development" (Latif 2000:126).

[11] A recent description of the DIT and its theoretical foundation can be found in Rest, J. R., D. Narvaez, et al. (2000). "A neo-Kohlbergian approach to morality research." Journal of Moral Education **29**(4): 381-395.. For information on using the DIT: Center for the Study of Ethical Development, University of Minnesota, 206 Burton Hall, 178 Phillsbury Dr SE, Minneapolis MN 55455 or http://edpsy.coled.umn.edu/psychf/csed.

### *5.4. The Ethical Reasoning Tool*

The Ethical Reasoning Tool (ERT) was developed to test improvement in moral reasoning for nursing students following a required clinical ethics unit. Unlike the DIT, which is a multiple-choice test, in the ERT "responses are generated by respondents themselves, without the use of external prompts" (McAlpine, Kristjanson et al. 1997[12]).

Using a pretest-posttest method, the instructor-researchers had students read a case study detailing

> the dilemma of an anxious, upset patient asking nurses for diagnostic information in a ward setting where "common practice" was for nurses to say nothing (since some doctors preferred that patients not know of a malignant diagnosis), and where the doctor could not be contacted for several hours. [McAlpine, Kristjanson et al. 1997]

Students were "given 40 minutes of class time" to "identify significant ethical/moral issues raised by the case study, state what they thought should be done and support their answers from an ethical/moral perspective (McAlpine, Kristjanson et al. 1997).

The ERT allows the instructor to locate respondents in one of three levels of ethical reasoning – the egocentric, conventional, or reflective (parallel to the levels used in the DIT) – through analysis of students' pretest-posttest essays. It provides "exemplars" to aid in analysis of the essays on eight ethical reasoning components. For example, the exemplars for Component 3, "Use of personal values," are as follows:

Level 1: Traditional – "Non-reflective use of own values as determinants of right/wrong. Personal opinions the focus (e.g., I'd want to know if I were in her position)."

Level 2: Traditional/Reflective – "[Use of own values is] evident, but not total focus. Some acknowledgment of contextual factors (e.g., patient values)."

Level 3: Reflective – "[Own values] may be acknowledged, but do not drive decision making. Focus on patient values vs. those of health care providers."

Statistically significant, even dramatic, changes were evident in three of the eight ethical reasoning components: "recognition of ethical issues . . . use of an ethical framework; and . . . use of personal values to direct decision making. In these three areas there was an identifiable shift from Level 1 [egocentric] responses in the pre-tests to Level 2 [conventional] responses in the post-tests."

In a strategy similar to that described in Section 4.7, students were also asked to compare their pretest essays to their posttest essays. Excerpts from these self-assessments show that students recognized striking differences; for example, "I was surprised to re-read my first answer and realize that I wasn't on the patient's side at all."

The ERT is clearly harder to score than the DIT, which is machine-read, but the study showed that the ERT demonstrated a high level of clarity and content validity, inter-rater reliability, and construct validity. The DIT is also a general-use tool, whereas the ERT would

[12] I do not include page numbers for quotations because I am working from a copy of the article printed from Ovid (http://gateway1.ovid.com).

require adaptation from one context to another – it is not a "plug-and-play" tool. However, the results of this one study are impressive and the general approach is promising.

# 6. Summary

I begin this report (Section 2) by offering a background of terminology (including the distinction between assessment and evaluation); available methods for teaching and assessment; some of the distinctive challenges inherent in assessing learning in ethics; and the generally underdeveloped state of assessment and evaluation of teaching in research ethics.

In Section 3 I describe the five modules in the responsible conduct of research uncovered by my literature search. All five include some description of assessment or evaluation, and I present them roughly in order of the sophistication of the assessment or the amount of detail provided. All five modules use either case study discussion or role playing, requiring students' active involvement rather than passive reception of information.

In Section 4 I summarize publications about eight courses or training programs in the responsible conduct of research. Again, all of the publications include some description of assessment and/or evaluation, and the level of rigor varies.

Section 5 is devoted to moral reasoning, a particularly well-studied and well-developed area.

# 7. Conclusion

This report demonstrates that methods for teaching, assessing, and evaluating research ethics exist, but more would be welcome. It is clear that (a) research ethics can be taught, (b) the quality of student learning in research ethics can be assessed, and (c) the effectiveness of programs in research ethics can be evaluated.

All three points have to be qualified, however; these objectives can be reached when "research ethics" is properly understood. I suggest that it is unsuitable to interpret "research ethics" in this context as "proper behavior in research;" a better interpretation would be "a proper understanding of responsible research practices."[13]

If this distinction is accepted, we can banish the bugaboo of reducing or eliminating misconduct in research as the only relevant goal of RCR instruction.

I will conclude by suggesting a few factors that have contributed to the difficulty the research community has had in coming to terms with appropriate goals for teaching research ethics and suitable methods of assessing and evaluating success.

1. Academic freedom. Scientists do not want to be told how to do their work (Sachs and Siegler 1993:874). Academic freedom involves both the liberty to investigate controversial or seemingly trivial topics, and the autonomy to decide how such

[13] "Premature" may actually be a better word than "inapt." We may be able to teach, assess, and evaluate behavior modification in research ethics one day, but evidence that we can do so now is scant (though not non-existent; see Section 5.1).

investigation will proceed. Discussions of research ethics can understandably feel like encroachments on this fundamental value.

2. The typical style of science teaching and assessment. A great deal of science instruction and assessment is focused, rightly, on objective, quantifiable knowledge. Scientists and science students are used to dealing with right and wrong answers – very different from the muddy and highly contextual (not to say subjective) world of research ethics.
3. Apprehension of a regulatory chilling effect. Scientists went ballistic when the Federal government suggested that a definition of "research misconduct" was needed, along with policies and procedures for dealing with allegations of research misconduct. The reasons for this response are complex, but at least one factor was the fear that too much or inappropriate oversight of the research process would stifle creativity.

I believe that since the late 1980's and the early 1990's, when research misconduct was defined, the first high-profile investigations into research misconduct took place, and pressure to provide training in the responsible conduct of research began to build, many researchers – and the research community in general – have learned how to balance these factors against the increasingly evident need for widespread education in research ethics. There is still a great deal of work to be done, but a good foundation has been laid.

# 8. Methodology

In my previous literature search for the Committee, my graduate assistant and I discovered that finding citations on these topics using general search terms, such as "*Ethics/ or *Scientific Misconduct," returned hundreds of citations, only a few of which were actually relevant. To cut down time wasted sifting through citations and abstracts, we adopted a three-pronged strategy for this literature review.

## *8.1. Poll*

On August 17, 2001, I sent an e-mail message to more than 1,800 researchers with some interest in the responsible conduct of research asking them to share citations of their publications, or the publications of other scholars, on "empirical evaluations of pedagogical approaches to teaching research ethics." I received more than one hundred responses (see Section 9), some expressing interest in the project, others referring me to other researchers, and some providing one or more citations. I pursued the likely-looking citations and incorporated many of them into this report.

### *8.2. Search*

On August 30 and September 3, 2001, I searched the three databases in ISI Web of Science (http://webofscience.com/) with the following results:

| ID | Search criteria | Hits | Action taken |
|---|---|---|---|
| a | (ethic* or moral*) same (teach* or train* or educat*) same (assess* or evaluat*) same (scienc* or research*) | 21 | All imported for study |
| b | (ethic* or moral* or responsibl*) same (teach* or train* or educat*) same (assess* or evaluat*) same (scienc* or research*) | 27 | Six that were not duplicates from search (a) imported for study |
| c | (ethic* or moral* or responsibl*) same (teach* or train* or educat*) same (assess* or evaluat*) | 174 | Review of the titles of the first 30 showed none relevant |
| d | (ethic* or moral* or responsibl*) and (teach* or train* or educat*) and (assess* or evaluat*) and (scienc* or research*) | 482 | Review of the titles of the first 30 showed none relevant |
| e | related to REISER SJ. CREATING A COURSE ON ETHICS IN THE BIOLOGICAL SCIENCES | 11 | All imported for study |
| f | related to BROWN S. EFFECTS OF TRAINING IN THE RESPONSIBLE CONDUCT OF RESEARCH | 25 | Two relevant citations imported for study |
| g | (ethic* or moral* or responsibl*) and (teach* or train* or educat*) and (assess* or evaluat*) and (scienc* or research*) | 482 | Review of the titles of the first 30 showed none relevant |

After performing these seven searches and scrutinizing the results, I concluded that I had reached the point of diminishing returns.

### *8.3. Snowball*

For all relevant publications, I checked the works cited for further sources.

## 9. Acknowledgements

I wish to express my thanks to my graduate assistant, Jeannie Bermudez, who once again provided substantial logistical and moral support.

I am particularly grateful to Jennifer, Gwendolyn, and Vivian Livesay, who generously freed up precious family time to allow me to finish this report on deadline. Jennifer also provided valuable comments on the penultimate draft.

I wish to acknowledge the following scholars who responded to my call for assistance by sending articles, citations, or contact information for other scholars with interest and expertise in this area. I have not used all of the sources they offered, deeming some of them off-target for my purposes.

I am grateful to those who provided one or more citations: Helmut Baitsch and Gerlinde Sponholz, University of Ulm (Germany); Floyd E. Bloom, Neurome, Inc.; Richard Cash, Harvard University; Celia Fisher, Fordham University; Paul J. Friedman, University of California-San Diego; Della Hann, National Institutes of Health; Peter G. Hartel, University of Georgia; C. Ronald Kahn, Joslin Diabetes Center; Ron Kline, Cornell University; Anne Koerber, University of Illinois-Chicago; Michael C. Loui, University of Illinois-Urbana/Champaign; and Bruce Matis, Indiana University/Purdue University-Indianapolis.

I am also indebted to several scholars who provided other assistance, usually a reference to another researcher: George Abraham, University of Rochester; Audrey Bernfield, Harvard University; Steven J. Fliesler, Saint Louis University; Russell Garth, Council of Independent Colleges; John Hardwig, University of Tennessee-Knoxville; Robert J. Huskey, University of Virginia; Nita J. Maihle, Mayo Clinic; Mara McDonald, University of Wisconsin; Susan McHale, Penn State University; J. Gail Neely, Washington University School of Medicine; G. Michael Poteat, East Carolina University; Alan Price, DHHS; David B. Resnik, East Carolina University; Jerry L. Rosenberg, University of Pittsburgh; and Susan Scrimshaw, University of Illinois-Chicago.

Three worthies provided actual bibliographies: Julie Reyes, Michigan State University, provided six citations; Duane Roen, Arizona State University, shared an annotated bibliography of six items (see footnote 2); and Martha Glock, National Library of Medicine, performed a literature search of MEDLINE, BIOETHICSLINE, ERIC, Education Abstracts, Magazine Index, Sociological Abstracts, Social SciSearch, and CINAHL for the years 1992-2001 and provided 246 citations, most with abstracts. Thanks also to Betsy L. Humphreys, Associate Director for Library Operations at NLM for her support.

Finally, five gentlemen provided copies of their own publications: Brian P. Coppola, University of Michigan; James M. DuBois, St. Louis University; Lawrence K. Duffy, University of Alaska-Fairbanks; Bernard S. Gerstman, Florida International University; and Lewis Pyenson, University of Louisiana-Lafayette.

# 10. Appendix: Evaluation of Moral Reasoning in Scientific Research

The report abstracted here was submitted to the Department of Education's Fund for the Improvement of Postsecondary Education (FIPSE) in November of 1996 as part of our final report. FISPE funded the first three years of the Teaching Research Ethics project. This version of the report abstract is adapted from http://poynter.indiana.edu/mr-abs.html. The full report, which includes the text of all written responses, can be found at http://poynter.indiana.edu/mr-rpt.html.

### *10.1. Report abstract*

A survey was sent in the summer of 1996 to about 300 persons who had received one or more copies of *Moral Reasoning in Scientific Research: Cases for Teaching and Assessment*.

- 84.4% of the respondents have used or intend to use the booklet to teach research ethics; only 13.5% responded "No" to this item. Of the 13 who selected "No," 10 included an answer to our question, "Why not?" None indicated they were not using the booklet because it was inappropriate or flawed.
- 62.2% of the respondents indicated that they had, or planned to, use the booklet to assess their students' moral reasoning ability.

Among our complimentary comments were the following:

- Very user friendly, a great resource, fine the way it is.
- I found the booklet very useful.
- One of the general areas covered in my course is research. Your booklet has been a valuable adjunct for reference when my class reaches that specific area.
- Thanks for a very helpful resource!
- A good workbook, useful in stimulating discussion after introducing basic moral principles in science.

| **Responses to poll (97 responses out of c. 300 sent).** | | | |
|---|---|---|---|
| Have you used, or do you intend to use, Moral Reasoning to teach research ethics? | Yes | 81 | 84.4% |
| | No | 13 | 13.5% |
| | Not certain | 2 | 2.1% |
| | TOTAL | 96 | 100% |
| If yes, at what level did you/will you use it? | Undergraduate | 31 | 30.1% |
| | Graduate | 52 | 50.4% |
| | Post-graduate | 18 | 17.5% |
| | Professional | 2 | 2% |
| | TOTAL | 103 | 100% |
| Did you/will you use the cases to assess your students' development in moral reasoning? | Yes | 51 | 62.2% |
| | No | 26 | 31.7% |
| | Maybe | 5 | 6.1% |
| | TOTAL | 82 | 100% |

### *10.2. The Survey*

August 5, 1996

Dear Colleague,

I am writing to persons who have received one or more copies of Moral Reasoning in Scientific Research: Cases for Teaching and Assessment.

As you may know, development of the booklet was made possible in part by a grant from the United States Department of Education's Fund for the Improvement of Postsecondary Education

(FIPSE).

The term of our grant is drawing to a close, and I am preparing to write the final report to FIPSE, in which I would like to be able to give some indication of how the booklet has been received. I would greatly appreciate it if you would take a few moments to fill out the form below and return it to me at the Poynter Center. (Feel free to add comments on the back or additional sheets.)

Thank you for your help,

Kenneth D. Pimple, Ph.D.
Research Associate

A) What is your name?

B) Have you used, or do you intend to use, Moral Reasoning to teach research ethics? (circle one)

Yes No

B.1) If no, why not?

B.2) If yes, at what level did you/will you use it?

Undergraduate Graduate Post-graduate

B.3) Did you/will you use the cases to assess your students' development in moral reasoning?

Yes No

B.4) If no, why not?

C) How could the booklet be *changed* to make it more useful to you?

D) What could be *added* to the booklet to make it more useful to you?

E) What other materials or resources for teaching research ethics would you find useful?

# 11. Appendix: Self-assessment and peer assessment

Many studies have been done on the use of self- and peer assessment in teaching; these studies are helpfully summarized by F. Dochy and colleagues (Dochy, Segers et al. 1999). I offer this abstract of the article for ease of reference.

"The era of testing can be characterized by a complete separation of instruction and testing activities, by a measurement that was passively undergone by the students, by measurement of knowledge of decontextualised subject matter that was unrelated to the student's experiences, and by measuring products solely in the form of a single total score. The assessment era promotes integration of assessment and instruction, seeing the student as an active person who shares responsibility, reflects, collaborates and conducts a continuous dialogue with the teacher. Assessment is then characterized by a pluralistic approach and by the use of interesting real-life (i.e. authentic) tasks" (Dochy, Segers et al. 1999:331).

"These new methods, such as case-based and problem-based learning, are directed towards producing highly knowledgeable individuals, but do also stress problem-solving skills, professional skills and authentic learning, i.e. learning in real-life contexts" (Dochy, Segers et al. 1999:332).

### *11.1. Self-assessment*

"*Definition:* Self-assessment refers to the involvement of learners in making judgments about their own learning, particularly about their achievements and the outcomes of their learning. Self-assessment is not a new technique. It is a way of increasing the role of students as active participants in their own learning, and is mostly used for formative assessment in order to foster reflection on one's own learning processes and results" (Dochy, Segers et al. 1999:334).

D. Boud and N. Falchikov "analyzed studies published between 1932 and 1988, which investigated student self-ratings compared to the ratings of students by teachers, and reported the overrating and the underrating of students. They related these findings to the different abilities of students. Their finding was that good students tended to underrate themselves and that weaker students overrated themselves. Students in higher-level classes could better predict their performance than students in lower-level classes" (Dochy, Segers et al. 1999:334).

"Overall, it can be concluded that research reports positive findings concerning the use of self-assessment in educational practice. Students who engage in self-assessment tend to score most highly on tests. Self-assessment, used in most cases to promote the learning of skills and abilities, leads to more reflection on one's own work, a higher standard of outcomes, responsibility for one's own learning and increasing understanding of problem-solving. The accuracy of the self-assessment improves over time. This accuracy is enhanced when teachers give feedback on students' self-assessment" (Dochy, Segers et al. 1999:337).

### *11.2. Peer assessment*

"*Definition:* . . . Peer assessment [is] the process through which groups of individuals rate their peers. This exercise may or may not entail previous discussion or agreement over criteria. It may involve the use of rating instruments or checklists which have been designed by others before the peer assessment exercise, or designed by the user group to meet its particular needs" (Dochy, Segers et al. 1999:337).

"Experience from peer assessment indicates that peer assessment can be valuable as a formative assessment method and hence as a part of the learning process. Students become more involved, both in the learning and in the assessment process. They find peer assessment sufficiently fair and accurate. However, the following can also be observed during peer assessment: friendship marking, resulting in overmarking; collusive marking, resulting in a lack of differentiation within groups; decibel marking, where individuals dominate groups and get the highest marks; and parasite marking, where students fail to contribute but benefit from group marks. These problems can be prevented by combining peer assessment with self-assessment of co-assessment. This may be why the majority of studies investigate these combinations of assessment forms" (Dochy, Segers et al. 1999:340).

### *11.3. Self- and peer assessment*

"*Definition:* Self- and peer assessment are combined when students are assessing peers but the self is also included as a member of the group and must be assessed. This combination fosters reflection on the student's own learning process and learning activities compared to those of the other members in the group or class" (Dochy, Segers et al. 1999:340).

"The development of criteria through active cooperation between teachers and students seems to be a critical success factor for self- and peer-assessment, as is the development of a series of instructions for students to set criteria for themselves. A third critical success factor is congruence between the mode of group activity and the evaluation of group work. [K. A. Oldfield and J. M. K. Macalpine] suggest a stepwise approach to group project assessment starting with peer assessment from the other groups' work, moving on to peer assessment of their own group's work, and ending with self-assessment" (Dochy, Segers et al. 1999:342).

### *11.4. Co-assessment*

"*Definition:* Co-assessment, the participation of students and staff in the assessment process, is a way of providing an opportunity for students to assess themselves whilst allowing the staff to maintain the necessary control over the final assessments" (Dochy, Segers et al. 1999:342).

"The findings indicate that the use of the combination of self-, peer, and co-assessments is effective. The results regarding accuracy indicate that self- and peer assessment can be used for summative purposes as a part of the co-assessment, by giving the tutor the power to express the final decision about a process or a product. In this way the traditional assessment, where the tutor makes an autonomous decision, is not comparable with co-assessment. The combination of self-, peer and co-assessment makes tutors and students work together in a constructive way and as a result they come to higher levels of understanding by negotiation. When the students becomes teacher, this role-change provides him or her with insights into the assessment process" (Dochy, Segers et al. 1999:344).

> We detected eight positive effects of self-, peer, and /or co-assessment which arise from our body of research.
>
> 1. Increased student confidence in the ability to perform.
> 2. The increased awareness of the quality of students' own work.
> 3. Increased student reflections on their own behavior and/or performance.
> 4. Increased student performance on assessments, increased quality of the learning output.
> 5. Effectiveness of approaches to learning.
> 6. Taking responsibility for learning; the independence of students.
> 7. Increased student satisfaction.
> 8. Ameliorated learning climate. [Dochy, Segers et al. 1999:345]
>
> The following guidelines have emerged from our study.
>
> 1. Training in the skill to self-assess or to peer assess has to be provided in order to obtain an optimal impact on the learning process, at least for beginning students. The first assessments which involve students as assessors should perhaps be implemented with groups of third and fourth year university students.
> 2. Self-assessment takes time, and sometimes support for students will be necessary during the self-assessment.
> 3. Self-assessment can be used fairly easily for formative purposes. Students should learn to see this as a tool for learning.

4. The habit of academics to do the teaching and all the marking is hard to change, and it seems likely that a staff development program will be needed if the approaches discussed in this article are to be implemented widely.
5. In peer assessment, criteria should be determined beforehand. Experiences show that it works well if these criteria are determined jointly by staff and students.
6. Peer assessment criteria should be presented in operational terms with which all students are familiar. Students can play a role in the process of operationalisation.
7. Peer assessment can be used as a tool for summative assessment, in combination with other assessment instruments. It can lead to a student profile or a peer assessment factor, i.e. a correction facto calculated from the peer assessment scores that adjusts a preceding group score for a collaborative product. Peer assessment measures should not be used as the sole indicator in a summative assessment. [Dochy, Segers et al. 1999:346-347]

# 12. Appendix: Thoughts on program evaluation

In August 2000, when universities and other research institutions were gearing up to implement the NIH mandate for training in the protection of human subjects (see http://grants.nih.gov/grants/guide/notice-files/NOT-OD-00-039.html), I attended a conference sponsored by the University of Michigan and attended by representatives from most of the university members of the Committee on Institutional Cooperation.[14] Among the many topics we discussed, one of our shared concerns was program evaluation. We have to provide this training; how can we tell whether we are having an effect?

I quickly drafted a set of questions that I thought would be helpful in program evaluation. I include a slightly edited version of that list here in the hope that it will be of some use.

In order to evaluate the success of this broad training program, here are some of the things we should know and do to:

- Can we easily identify researchers who perform (a) less than and minimal risk research? (b) low risk research? (c) medium-to-high risk research? If we can't, we should find a way to do so. This will allow us to train researchers involved in high-risk research first and to create training appropriate to the audience – for example, if we have hundreds of minimal risk researchers but only a handful of high-risk researchers, we need to distribute our resources and effort appropriately.
- What do we currently know about compliance and non-compliance over the last 5 years? What hard data do we have? What anecdotal knowledge can we record and systematize now for later comparison? Anecdotal evidence – the impressions of IRB members and administrators of problem areas – is not as good as hard data, but it is better than nothing. Furthermore, taking the time to record and systematize impressions may seem like a waste of time, but it could provide valuable information for evaluating training

[14] Indiana University-Bloomington, Indiana University/Purdue University-Indianapolis, Michigan State University, Northwestern University, The Ohio State University, The Pennsylvania State University, Purdue University, the University of Chicago, the University of Illinois-Chicago, the University of Illinois-Urbana/Champaign, the University of Iowa, the University of Michigan, the University of Minnesota, and the University of Wisconsin-Madison.

effectiveness a few years down the road. If impressions are not recorded before the programs are put in place, an already soft data source will become hopelessly squishy.

- How can we increase our knowledge on compliance and non-compliance immediately? By next year? Are we keeping statistics? Are we doing audits and surprise inspections?
- Are there disincentives for non-compliance?
- Should we institute remedial training and additional oversight (probation) for non-compliance?
- Are there rewards and incentives for compliance? Does it figure in tenure and promotion, or awarding of internal grants and other forms of support?
- How can we minimize current disincentives for compliance (for example, by creating shorter, simpler forms for IRB approval)?
- How can we increase researcher identification with the approval process? We might want to
    - increase appreciation of the stakes, including the possibility of an OHRP shutdown a la Duke, UIC, Johns Hopkins, etc.;
    - promote an atmosphere of high standards and dedication to compliance;
    - provide advanced, ongoing, and interesting training in research ethics (not just regulatory compliance);
    - encourage more researchers to serve on an IRB, if only for a few meetings, to enhance their understanding and support of the process; and
    - ensure that each unit doing human subjects research has at least one trained spokesperson (and alternate) who can act as a liaison between the unit and the IRB, helping researchers (especially new researchers) with developing good protocols and with interpreting responses from the IRB.

## 13. Works Cited

AAAS (1996). Integrity in Scientific Research: Five Video Vignettes. Washington, D.C., American Association for the Advancement of Science.

Apple, M. W., Ed. (1995). Review of Research in Education. Washington D.C., The American Educational Research Association.

Ayscue, G. (2000). Research integrity training: Ethics workshop experiences of first year graduates. Washington, D.C., Office of Research Integrity, National Institutes of Health.

Bakken, L. L. and P. A. Reichel (1996). "Designing educational workshops for clinical research study support staff." Journal of the Society of Research Administrators **28**(1/2): 11-26.

Baldwin, D. C., S. R. Daugherty, et al. (1991). "Changes in moral reasoning during medical school." Academic Medicine **66**(9): S1-S3.

Barden, L. M., P. A. Frase, et al. (1997). "Teaching scientific ethics: A case studies approach." American Biology Teacher **59**(1): 12-14.

Bebeau, M. J. (2000). Influencing the moral dimensions of professional practice: Implications for teaching and assessing for research integrity. Washington, D.C., Office of Research Integrity, National Institutes of Health.

Bebeau, M. J. (forthcoming). Influencing the moral dimensions of professional practice: Implications for teaching and assessing for research integrity. Investigating Research Integrity: Proceedings of the First ORI Research Conference on Research Integrity. N. H. Steneck and M. D. Scheetz. Washington, D.C., Office of Research Integrity, National Institutes of Health.

Bebeau, M. J., K. D. Pimple, et al. (1995). Moral Reasoning in Scientific Research: Cases for Teaching and Assessment. Bloomington, Poynter Center, Indiana University.

Beins, B. C. (1993). "Using the Barnum effect to teach about ethics and deception in research." Teaching of Psychology **20**(1): 33-35.

Brown, S. and M. W. Kalichman (1998). "Effects of training in the responsible conduct of research: A survey of graduate students in experimental sciences." Science and Engineering Ethics **4**(4): 487-498.

Bulger, R. E., E. Heitman, et al., Eds. (1993). The Ethical Dimensions of the Biological Sciences. Cambridge, England, Cambridge University Press.

Bulger, R. E. and S. J. Reiser (1993). "Studying science in the context of ethics." Academic Medicine **68**(9): S5-S9.

Callahan, D. and S. Bok (1980). The Teaching of Ethics in Higher Education: A Report by The Hastings Center. Hastings-on-Hudson, NY, The Hastings Center.

Centra, J. A. (1993). Reflective Faculty Evaluation: Enhancing Teaching and Learning Effectiveness. San Francisco, Jossey-Bass Publishers.

Coppola, B. P. (2000). "Targeting entry points for ethics in chemistry teaching and learning." Journal of Chemical Education **77**(11): 1506-1511.

Cross, K. P. and M. H. Steadman (1996). Classroom Research: Implementing the Scholarship of Teaching. San Francisco, Jossey-Bass Publishers.

Darling-Hammond, L., Ed. (1994). Review of Research in Education. Washington D.C., The American Educational Research Association.

Davidson, E. W., H. E. Cate, et al. (2000). Data manipulation in the undergraduate classroom: What are we teaching? Washington, D.C., Office of Research Integrity, National Institutes of Health.

Dochy, F., M. Segers, et al. (1999). "The use of self-, peer and co-assessment in higher education: a review." Studies in Higher Education **24**(3): 331-350.

DuBois, J. M., J. Ciesla, et al. (forthcoming). "Ethics education in U.S. medical schools: A study of syllabi." Academic Medicine.

DuBois, J. M., J. Ciesla, et al. (forthcoming). Research ethics in U.S. medical education: An analysis of ethics course syllabi. Investigating Research Integrity: Proceedings of the First ORI Research Conference on Research Integrity. N. H. Steneck and M. D. Scheetz. Washington, D.C., Office of Research Integrity, National Institutes of Health.

Eastwood, S., P. Derish, et al. (1996). "Ethical issues in biomedical research: Perceptions and practices of postdoctoral research fellows responding to a survey." Science and Engineering Ethics **2**(1): 89-114.

Elliott, D. and J. E. Stern (1996). "Evaluating teaching and students' learning of academic research ethics." Science and Engineering Ethics **2**(3): 345-366.

Elliott, D. and J. E. Stern, Eds. (1997). Research Ethics: A Reader. Hanover, NH, University Press of New England.

Fisch, L., Ed. (1996). Ethical Dimensions of College and University Teaching. New Directions for Teaching and Learning. San Francisco, Jossey-Bass Publishers.

Fisher, C. B. and T. L. Kuther (1997). "Integrating research ethics into the introductory psychology course curriculum." Teaching of Psychology **24**(3): 172-175.

Gerstman, B. S. (2001). E-mail message.

Gerstman, B. S. (forthcoming). An effective short course on research integrity. Investigating Research Integrity: Proceedings of the First ORI Research Conference on Research Integrity. N. H. Steneck and M. D. Scheetz. Washington, D.C., Office of Research Integrity, National Institutes of Health.

Hayes, R. P., A. Stoudemire, et al. (1999). "Qualitative outcome assessment of a medical ethics program for clinical clerkships: A pilot study." General Hospital Psychiatry **21**(4): 284-295.

Heitman, E. (2000). "Ethical values in the education of biomedical researchers." Hastings Center Report **30**(4): S40-S44.

Heitman, E., P. J. Salis, et al. (2000). Teaching ethics in biomedical sciences: Effects on moral reasoning skills. Washington, D.C., Office of Research Integrity, National Institutes of Health.

Kalichman, M. W. and P. J. Friedman (1992). "A pilot study of biomedical trainees' perceptions concerning research." Academic Medicine **67**(11): 769-775.

Katz, D., R. L. Kahn, et al., Eds. (1982). The Study of Organizations: Findings from Field and Laboratory. San Francisco, Jossey-Bass Publishers.

Kezar, A. and P. Eckel, Eds. (2000). Moving Beyond the Gap Between Research and Practice in Higher Education. New Directions for Higher Education. San Francisco, Jossey-Bass Publishers.

Korenman, S. G. and A. C. Shipp (1994). Teaching the Responsible Conduct of Research through a Case Study Approach: A Handbook for Instructors. Washington, D.C., Association of American Medical Colleges.

Kovac, J. (1996). "Scientific ethics in chemical education." Journal of Chemical Education **73**(10): 926-928.

Latif, D. A. (2000). "The relationship between ethical dilemma discussion and moral development." American Journal of Pharmaceutical Education **64**(2): 126-133.

Macrina, F. L. (1995). Scientific Integrity: An Introductory Text with Cases. Washington, D.C., ASM Press.

Macrina, F. L. and C. L. Munro (1995). "The case-study approach to teaching scientific integrity in nursing and the biomedical sciences." Journal of Professional Nursing **11**(1): 40-44.

Mastroianni, A. C. and J. P. Kahn (1998). "The importance of expanding current training in the responsible conduct of research." Academic Medicine **73**(12): 1249-1254.

McAlpine, H., L. Kristjanson, et al. (1997). "Development and testing of the ethical reasoning tool (ERT): An instrument to measure the ethical reasoning of nurses." Journal of Advanced Nursing **25**(6): 1151-1161.

Mitcham, C., B. M. Olds, et al. (2000). Assessing the teaching and learning of research integrity: Research opportunities. Washington, D.C., Office of Research Integrity, National Institutes of Health.

Mitcham, C., B. M. Olds, et al. (forthcoming). A plea for pursuing new dimensions of assessment in the teaching and learning of research integrity. Investigating Research Integrity: Proceedings of the First ORI Research Conference on Research Integrity. N. H. Steneck and M. D. Scheetz. Washington, D.C., Office of Research Integrity, National Institutes of Health.

Pascarella, E. T. and P. T. Terenzini (1991). How College Affects Students: Findings and Insights from Twenty Years of Research. San Francisco, Jossey-Bass Publishers.

Pimple, K. D. (1994). "Teaching ethics in the biological sciences." Medical Humanities Review **8**(2): 65-68.

Pimple, K. D. (2001). The moral climate of research in the United States today. Washington, D.C., National Academies.

Pollock, R. E. (2001). E-mail message.

Pollock, R. E., S. A. Curley, et al. (1994). "A short course in research ethics for trainees." Academic Medicine **69**(3): 213-214.

Pollock, R. E., S. A. Curley, et al. (1995). "Ethics of research training for NIH T32 surgical investigators." Journal of Surgical Research **58**(2): 247-251.

Reiser, S. J. and E. Heitman (1993). "Creating a course on ethics in the biological sciences." Academic Medicine **68**(12): 876-879.

Rest, J. R., D. Narvaez, et al. (2000). "A neo-Kohlbergian approach to morality research." Journal of Moral Education **29**(4): 381-395.

Rosnow, R. L. (1990). "Teaching research ethics through role-play and discussion." Teaching of Psychology **17**(3): 179-181.

Sachs, G. A. and M. Siegler (1993). "Teaching scientific integrity and the responsible conduct of research." Academic Medicine **68**(12): 871-875.

Self, D. J. and D. C. Baldwin (1998). "Does medical education inhibit the development of moral reasoning in medical students? A cross-sectional study." Academic Medicine **73**(10): S91-S93.

Self, D. J. and D. C. Baldwin (2000). "Should moral reasoning serve as a criterion for student and resident selection?" Clinical Orthopaedics and Related Research(378): 115-123.

Self, D. J., D. C. Baldwin, et al. (1993). "Teaching medical ethics to first-year students by using film discussion to develop their moral reasoning." Academic Medicine **68**(5): 383-385.

Self, D. J., M. Olivarez, et al. (1998). "The amount of small-group case-study discussion needed to improve moral reasoning skills of medical students." Academic Medicine **73**(5): 521-523.

Self, D. J., M. Olivarez, et al. (1998). "Clarifying the relationship of medical education and moral development." Academic Medicine **73**(5): 517-520.

Self, D. J., D. E. Schrader, et al. (1991). "A pilot study of the relationship of medical education and moral development." Academic Medicine **66**(10): 629-629.

Self, D. J., F. D. Wolinsky, et al. (1989). "The effect of teaching medical ethics on medical students' moral reasoning." Academic Medicine **64**(12): 755-759.

Smart, J. C., Ed. (1985). Higher Education: Handbook of Theory and Research. New York, Agathon Press Inc.

Sponholz, G. (2000). "Teaching scientific integrity and research ethics." Forensic Science International **113**(1-3): 511-514.

Stern, J. E. and D. Elliott (1997). The Ethics of Scientific Research: A Guidebook for Course Development. Hanover NH, University Press of New England.

Strohmez, D. B. and A. A. Skleder (1992). "The use of role-play in teaching research ethics: A validation study." Teaching of Psychology **19**(2): 106-108.

Sundermeyer, P. A. (2000). Integrating Internet based materials into a curriculum for responsible conduct of research. Washington, D.C., Office of Research Integrity, National Institutes of Health.

Sweet, S. (1999). "Using a mock Institutional Review Board to teach ethics in sociological research." Teaching Sociology **27**(1): 55-59.

Sweeting, L. M. (1999). "Ethics in science for undergraduate students." Journal of Chemical Education **76**(3): 369-372.

Theall, M., Ed. (1999). Motivation from Within: Approaches for Encouraging Faculty and Students to Excel. New Directions for Teaching and Learning. San Francisco, Jossey-Bass Publishers.

Treichel, P. M. (1999). "Ethical conduct in science: The joys of teaching and the joys of learning." Journal of Chemical Education **76**(10): 1327-1329.

Uhl, N. P., Ed. (1983). Using Research for Strategic Planning. New Directions for Institutional Research. San Francisco, Jossey-Bass Publishers.