2014

Aligning objectives and assessment in responsible conduct of research instruction

Alison Antes
Washington University School of Medicine in St. Louis

James M. DuBois
Washington University School of Medicine in St. Louis

JOURNAL OF MICROBIOLOGY & BIOLOGY EDUCATION, December 2014, p. 108-116
DOI: http://dx.doi.org/10.1128/jmbe.v15i2.852

Aligning Objectives and Assessment in Responsible Conduct of Research Instruction

Alison L. Antes* and James M. DuBois
Division of General Medical Sciences, Washington University School of Medicine, St. Louis, MO 63110

Efforts to advance research integrity in light of concerns about misbehavior in research rely heavily on education in the responsible conduct of research (RCR). However, there is limited evidence for the effectiveness of RCR instruction as a remedy. Assessment is essential in RCR education if the research community wishes to expend the effort of instructors, students, and trainees wisely. This article presents key considerations that instructors and course directors must consider in aligning learning objectives with instructional methods and assessment measures, and it provides illustrative examples. Above all, in order for RCR educators to assess outcomes more effectively, they must align assessment to their learning objectives and attend to the validity of the measures used.

Aligning objectives and assessment in responsible conduct of research instruction

Ethical practices in scientific research are essential to advancing the aims of science. Ethical standards promote the accuracy and objectivity of research, collaboration among scientists, public support for research, and respect for research subjects. To realize these goals, scientists must translate ethical standards into their research practices and behaviors.

Courses in the responsible conduct of research (RCR) are a primary strategy for educating scientists about ethical concerns and regulatory requirements. The National Institutes of Health (NIH) and the National Science Foundation (NSF) have mandated RCR instruction for all trainees (33). Despite the investment of millions of dollars and hours in RCR training,[a] we have limited evidence about whether RCR instruction is associated with any positive outcomes (19).

Our aim is to provide instructors and RCR program directors with guidance regarding the assessment of RCR instruction. Three critical points—points frequently ignored in RCR education—provide the basis for our discussion:

1. Educational objectives should guide instructional methods and assessment.
2. Assessing outcomes is essential to developing good educational programs.
3. Only reliable and valid measures should be used to assess outcomes.

Objectives drive everything

It is impossible to evaluate outcomes unless one knows what one is trying to accomplish. Several schemes classify learning outcomes (7, 9, 16, 22, 23), and one practical categorization includes knowledge, skills, and attitudes. Knowledge focuses on understanding, remembering, and recalling concepts, facts, and procedures. Skills require executing technical, mental, or interpersonal tasks. Attitudes are ingrained ways of thinking or feeling about something, and they are closely tied to people's beliefs and values. In short, knowledge represents knowing "what," skills "how," and attitudes "why."

What are reasonable objectives for RCR education?

DuBois and Dueker conducted a Delphi survey with 18 RCR experts to establish a consensus on the aims of RCR education (15). Eight learning objectives received strong support from 80% or more of the panelists. In the area of knowledge: identifying sources of RCR regulations and policies and resources for sound information; increasing knowledge of ethical and regulatory issues in research; and understanding the difference and relationship between ethics and compliance. In the area of skills: increasing ethical sensitivity; fostering ethical problem-solving skills; and developing strategies for avoiding ethical problems. In the area of attitudes: appreciating the importance of RCR and

*Corresponding author. Mailing address: Division of General Medical Sciences, Washington University School of Medicine, 660 South Euclid Avenue, Campus Box 8005, St. Louis, MO 63110. Phone: 314-362-6006. Fax: 314-454-5113. E-mail: aantes@dom.wustl.edu.

[a] The CITI training program alone states that over 6.4 million courses have been completed, with an average of 4.5 hours invested in each of their basic courses. See https://www.citiprogram.org/ (accessed on 5 September 2014). The U.S. Office of Research Integrity's RCR Resource Development program awarded $1.5 million to various institutions to develop programs, and has spent over $1 million developing their own training programs, such as The Lab. See http://ori.hhs.gov/ (accessed on 5 September 2014).

fostering research integrity and professional character, which includes motivating moral action.

The ultimate objective in RCR education is to foster ethical behavior. Here we see the artificiality of dividing educational objectives into distinct domains. Research integrity manifests itself through ethical actions, which involve the application of knowledge, skills, and attitudes.

How do these ideal objectives map onto the actual objectives that instructors have? Kalichman and Plemmons conducted phone interviews with 50 RCR instructors and found a wide variety of instructor goals and perceptions of goals for RCR instruction (21). The authors expressed concern about the lack of clarity that many instructors articulated regarding their goals, including confusing their goals with their methods.

Selecting objectives for specific programs

Several considerations guide the selection of educational objectives. First, what is the educational stage of the learners? This could affect the learning domain targeted, with early education focusing more on knowledge and later education focusing more on skills. For example, the Accreditation Council for Graduate Medical Education (ACGME) describes five competency levels in the area of professionalism ranging from 1) "Is aware of basic bioethical principles and is able to identify ethical issues in clinical situations" to 5) "Demonstrates leadership and mentorship on understanding and applying bioethical principles clinically, particularly responsiveness to patients above self-interest and self-monitoring" (32). Similarly, instructors might expect postdoctoral fellows to move from ethical sensitivity in the early stages of their program to creative ethical problem-solving in the later stages. However, a full-length course in RCR might aspire to address all three learning domains with early-stage learners but expect a lower level of mastery. That being said, it is unclear whether it is realistic to expect individuals to grow in ethicality as they progress from undergraduate through doctoral and postdoctoral levels of education. Pressures in the climate, self-serving biases, poor mentoring, and competing interests may detract from successful problem-solving at any level of training (1, 2, 24); in fact, moral regression is regularly observed in some fields (e.g., during years of medical training) (20).

Second, what is feasible given limitations in resources, time, and instructors? Does the learning space or format permit interpersonal dialogue and debate? If not, it will be challenging to foster the cognitive dissonance necessary to encourage learners to question their assumptions and seek new ways of approaching problems (35). Are the instructors trained to use case studies to foster ethical problem-solving? Does the course provide enough contact time to do more than convey basic knowledge?

While it is difficult for one RCR course to meet a broad range of robust objectives, ideally research training program directors seek to develop an array of programs aimed at fostering development across learning domains—including not only formal coursework, but also mentoring and informal programming (6). Table 1 provides examples of educational objectives aligned with instructional methods and describes how these might influence assessment.

Why assessment matters

Two kinds of educational assessment exist: summative assessment is used to measure achievement after learning has taken place (e.g., a final exam), and formative assessment is used to provide feedback on student progress to support ongoing learning and course improvement (e.g., weekly reflective journaling or short quizzes with corrective feedback). We focus on summative assessment, but instructors should incorporate formative assessment into learning activities throughout the course.

Gauging student learning through summative assessment provides data to address the questions:

1. Is this individual learner achieving the objectives of the course?
2. Is this course effective overall in meeting some or all its objectives?

Answering the first question can help individuals learn by providing feedback on mastery. It may also help instructors determine when an individual has made sufficient progress to complete training. Answering the second question guides improvements in instructional design.

This latter task is essential in RCR instruction. It is easy to assume that RCR courses have positive effects. Knowledgeable people usually design them with important objectives in mind. However, the history of educational interventions is marked by expensive and time-consuming projects that fail to demonstrate positive effects (37, 17). In RCR education, a systematic review and meta-analysis of RCR programs found that many programs had no positive effects, and some were associated with negative outcomes such as decreases in perspective-taking in decision making and increases in deceptive responses to ethical problems (4, 5).

If a measure is not valid, you do not know what you are measuring

All measurement is based on the operationalization of a concept, which involves value judgments and reduction. Thus, research to establish the reliability and validity of scores is essential to drawing meaningful conclusions from them. There are many different kinds of validity, but ultimately, they all relate to "construct validity" (27). Construct validation data help to answer the question, "What does this score mean?" Does it mean the same thing for different groups? Is it related to scores on similar tests? Does it predict any observable behaviors?

TABLE 1.
Aligning instructional objectives, methods, and assessments.

Learning Outcome	Ethical Problem-Solving Skill	Ethical Sensitivity Skill	Knowledge of Research Ethics	Attitudes and Values
Instructional Objective	Foster ethical problem-solving skills in the conduct of research	Increase the ability to recognize ethical issues in the design and conduct of research	Identify and understand research ethics regulations, policies, and resources	Cultivate constructive attitudes towards research ethics and compliance
Rationale	Researchers confront complex problems involving ethical, regulatory, and interpersonal dimensions. Specific strategies can be taught to improve the quality of decisions.	Researchers must recognize the presence of an ethical issue to engage in problem-solving. Researchers may also require sensitivity toward compliance, professionalism, and broader interpersonal issues to be fully effective. Ethical sensitivity skills are intertwined with knowledge, problem-solving skills, and attitudes about research ethics.	Researchers require foundational knowledge about the rules and regulations of the research enterprise. This knowledge provides a basis for ethical sensitivity and problem-solving.	To motivate ethical action, individuals must appreciate the importance of RCR and fostering research integrity. Attitudes influence action subsequent to instruction and influence the learning process itself through motivation and engagement. Attitudes are closely linked to values and biases, and researchers may not be fully aware of them or their influence.
General Instructional Approach	Activities must activate the multiple, complex skills associated with ethical problem-solving, such as considering the impact of actions on others, predicting downstream consequences, and applying relevant ethical principles and regulatory rules. Instruction should involve practicing skills through active case discussion or role plays. Case scenarios should not describe flagrant misbehavior, but present complex, "gray" areas that require problem-solving.	Activities and instruction should encourage creative thinking. Students should engage "what if" scenarios to explore multiple possibilities. The notion of particular "correct" answers should be suspended in favor of a focus on multiple competing principles, goals, and concerns. The learning environment must feel open and accepting so that all learners are comfortable sharing ideas.	Traditional lecture format may be effective to deliver key content; however, engaging students in discussions to reinforce concepts and make the topics more personally relevant facilitates learning. For this learning outcome, it may be appropriate for the instructor to think about the traditional model of an expert "delivering" content. However, for the other learning outcomes, the instructor is a facilitator or guide.	The instruction must challenge people to question and test their assumptions about the world, themselves, and others. Activities should challenge students to engage in self-assessment or self-reflection about their values, assumptions, or beliefs. Discussions should engage classmates in debates and sharing related to attitudes toward research ethics and the responsibilities of researchers. Instructors and mentors should model core values and positive attitudes.
Sample Instructional Methods	Written case analysis; small and large group discussion; role-play, video case analysis, student-generated case writing; online/video simulations	Written case analysis; small and large case discussion; role-play, video case analysis, student-generated case writing	Readings; informational lectures; PowerPoint slides; question-and-answer sessions; quizzes (graded or ungraded); independent study and research; student-led lectures/teaching others; individual or group written reports; worksheets; concept mapping	Perspective-focused lectures; reflective writing; debate; discussion; blogging; service learning; role modeling; interaction with non-experts (e.g., community members); self-assessments/awareness exercises; peer feedback; creative exercises such as drawing or acting; interviewing others; films; storytelling

TABLE 1.
Continued

Learning Outcome	Ethical Problem-Solving Skill	Ethical Sensitivity Skill	Knowledge of Research Ethics	Attitudes and Values
Possible Assessment Approaches	Engage the learner in the psychological activities that would underlie real-world ethical problem-solving by presenting scenarios that are interesting, relevant, and engaging. Objective tests should present response options that are all plausible, with some better and some worse. Qualitative approaches should develop detailed coding guides that reflect criteria for good decision making.	Present a realistic scenario followed by an open-ended prompt asking participants to indicate issues within the scenario; trained raters code the responses according to the issues identified.	Multiple-choice items with one best response or fill in the blanks. True/false items are generally not as effective as multiple-choice items in validly discriminating between those who know and do not know material. "Tricky" items should be avoided, as well as response options that are not plausible.	Brief statements followed by Likert-type scale responses to indicate agreement or disagreement with statements. Presentation of value statements or value names that can be rank ordered. Projective measures may involve picking a number of values from a longer list and placing them inside concentric circles.

If the question, "What does this score mean?" cannot be answered with validation data, then we do not know what we are measuring (despite the intentions of the test developer). One of the challenges of assessment is that a score can mean multiple things, and scores can be affected by many factors, such as intelligence and socially desirable responding.

What is needed to develop a reliable and valid measure? The first step is to systematically define the construct (i.e., the knowledge, skill, or attitude) to be assessed, followed by systematic item development to ensure appropriate, comprehensive content. In general, test developers must have experience formulating items according to rules that maximize reliability, and they must develop, at least initially, multiple items to assess each construct, or sub-dimension, of interest (18, 13). Typically, a large sample (generally 200 to 400) is needed, and participants must complete multiple validated instruments that measure variables that should (and should not) be related to the current variable(s) of interest. A test cannot be valid without first establishing that it is reliable. Different types of reliability are appropriate for different situations, but they all provide an estimate of the degree to which a measure produces stable, consistent results. Additional validation evidence is established when scores predict some external outcome, criterion, or behavior that they should theoretically.

Proper test development and validation will typically require that RCR instructors collaborate with individuals who possess expertise not only in statistics and research methodology, but more specifically measurement and psychometrics.

Aligning learning objectives with assessment measures

What follows is a discussion of examples of measures for only four objectives in RCR education to illustrate how complex objectives might be operationalized or translated into measurable traits. Table 2 provides available information about the measures and their validity; most are in the earliest stages of validation. There is no perfect measurement tool. A measure cannot be absolutely "validated," especially as measures are used in different contexts, are used with different groups, and become outdated. Furthermore, all measures require tradeoffs (e.g., between length of time to complete and the information generated, or between face validity and variance).

Ethical problem-solving skills in research

Two measures exist that operationalize ethical problem-solving by evaluating the degree to which the decisions an individual selects in response to professional problems illustrate the use of "sensemaking" or professional decision-making strategies, such as considering consequences to oneself and others, seeking help, managing emotions, questioning one's assumptions and motives, and recognizing relevant rules. These measures illustrate a limitation to measurement: if one has a different philosophy of professionalism, then one might disagree that these tests accurately measure ethical problem-solving in research.

The Ethical Decision-Making Measure (EDM) presents research vignettes to examine the ethicality of decisions across four domains of research behavior, as rated by expert judges based on field norms and guidelines. Additional scores illustrate respondents' endorsement of seven sensemaking strategies (30). Validation evidence has accrued through a number of studies with these measures. A summary of this research and the newest, refined versions of the measures are available online (http://ethics.publishpath.com/).

The Professional Decision-Making in Research (PDR) measure is similar to the EDM in its structure (14). It

presents research vignettes followed by six response options, and respondents pick the two options that best describe what they might do in each situation. High professionalism responses incorporate the use sensemaking strategies, and low professionalism responses violate one or more of these strategies for professional decision making.

TABLE 2.
Sample assessment measures in the four domains.

Measure Name	Description	Preliminary Validation
Ethical Problem-Solving		
Ethical Decision-Making Measure (EDM) (30)	25 vignettes specific to biological, health, or social sciences; pick two of eight options; about 45 minutes to complete. Produces multiple scores: four ethicality scores across four domains of research behavior—data management, the conduct of human or animal research, professional practices (e.g., treatment of staff and peer review), and business practices (e.g., conflict of interest). Also produces seven scores that reflect use of sensemaking strategies. Items may also be scored for endorsement of social-behavioral responses, such as deception and retaliation.	Beta version validated in sample of 102 doctoral students; demonstrated adequate reliability and correlated appropriately with the other psychological measures (e.g., intelligence, narcissism, self-deceptive enhancement) included to examine construct validity. Subsequent research using this measure in a sample of 252 doctoral students demonstrated that scores on the EDM were related, as expected, to environmental variables, such as laboratory climate and exposure to unethical behavior (29). A sample of 59 training participants also revealed that the scores on the measure changed as a result of training focused on a sensemaking framework (28). Subsequent updated versions of the test used in training at University of Oklahoma with >1,000 graduate students and in studies elsewhere (26).
Professional Decision-Making in Research Measure (PDR)[b]	16 vignettes relevant across human subjects, animal subjects, and translational research; pick two of six options; about 20 minutes to complete. This research is recent and ongoing, but preliminary evidence provides solid support for the validity of the measure (14). Available in parallel pre- and posttest forms.	Preliminary validation study with 300 NIH-funded researchers using a battery of measures to examine convergent validity. This stage of validation research demonstrated promising evidence for its validity—scores were not correlated with socially desirable responding, they were moderately correlated with narcissism and cynicism, and they were strongly correlated with a measure of moral disengagement in research. Ongoing research will seek to collect normative data in a sample of 400 NIH-funded researchers to establish "typical" scores.
Ethical Sensitivity		
Test for Ethical Sensitivity in Science (TESS) (11)	Adapted from Bebeau's Dental Ethical Sensitivity Test (8) to assess sensitivity among undergraduate students in life sciences and evaluate an ethics program using written responses instead of relying on interviews and interview transcription. One scenario about genetic testing in an animal followed by a prompt to write issues identified; coded by trained raters with a structured coding guide.	No inter-rater agreement estimates provided. A sample of students in an ethics program (n = 133) was compared to a control group (n = 134) using a pre/post design. The training sample scores increased after the course, and the control group scores went down on the posttest.
Test of Ethical Sensitivity in Science and Engineering (TESSE) (10)	Seven scenarios related to professional practice in science and engineering followed by open-ended space to comment on professional ethical issues and a set of eight statements. Participants were asked to rate each statement on a Likert-type scale according to whether they agree/disagree that it corresponds to an ethical issue in the scenario. Three of the seven scenarios are ethically neutral, and each scenario includes distractor responses that sound important, but are not relevant to the scenario. Authors aim to remove the open-ended portion after initial pilot studies.	No reliability estimates provided. Analyses using a pre/post test design indicated no change in scores from pretest to posttest in the control or experimental groups. Authors recommend instrument revision and further validation studies.

TABLE 2.
Continued

Measure Name	Description	Preliminary Validation
Knowledge of Research Ethics		
Research Ethics Knowledge and Analytical Skills Assessment (REKASA) (36)	33 multiple-choice, true-false, and short-answer items mapped to research ethics knowledge (e.g., IRB procedures, regulatory requirements), in addition to two cases with four open-ended ethical analysis questions each (for 41 items total).	Content validity established by extracting 271 available quiz items and mapping items to testing domains and to learning objectives. An initial pilot of 74 items (split into two assessment tools) was given to a group of 58 researchers before and after a research ethics course. Item discrimination was calculated for each item, and item discrimination greater than 0.2 allowed an item to be retained for the final version. The final version, consisting of 41 items, produced a Cronbach's alpha reliability coefficient of 0.84. The reliability coefficients of the shortened versions of the test without the case questions (α = 0.72) and the short-answer knowledge questions (α = 0.67) were also estimated.
RCR knowledge items indexed to Delphi topics[a]	125 multiple-choice items with one best choice among four options. Content of items indexed to specific topics within seven core areas of RCR instruction identified by a Delphi panel (15).	Items developed to cover core RCR content areas. Correct answers were indexed to five leading RCR textbooks or online courses. Preliminary reliability testing was conducted by dividing the 125 items into five test booklets consisting of 25 items and administering to 232 graduate students at the University of Oklahoma from 2009 to 2011 following RCR training. The average Cronbach's alpha across the five test booklets was good (0.71) and the Spearman Brown correction for test length provided a stronger reliability estimate (0.92). The average number of participants answering an item correctly was 67%.
Attitudes and Values		
The How I Think about Research (HIT-Res)[b]	Assesses the use of cognitive distortions (e.g., assuming the worst, blaming others, minimizing, and self-centered thinking) to disengage from research integrity and compliance (14). The test is comprised of 45 Likert-type items; higher scores indicate a greater level of disengagement from integrity and compliance in research.	Preliminary validation data from 300 NIH-funded investigators and trainees indicate excellent internal reliability and that the HIT-Res is strongly correlated with a general measure of moral disengagement.
Norms and Counter-norms of Science Survey (3)	Presents 16 items, each representing a norm or counter-norm in science (e.g., "Scientists openly share new findings with colleagues" vs. "Scientists protect their newest findings to ensure priority in publishing, patenting, or applications"). Using three sets of three-point scales, participants indicate the degree to which the norms should represent behavior of scientists, do represent the behavior of scientists, and represent their own behavior.	Content validity established through literature reviews and focus groups. Items administered to approximately 3,650 participants to examine variation of norms across disciplines and career stage. However, focus was not on item reliability or measure validation. Reported data focus on frequencies and differences between groups.

[a] Measure developed by James DuBois and Holly Bante. Measure is owned by the U.S. Office of Research Integrity but may be made available by contacting the lead author at jdubois@wustl.edu.

[b] Articles on the HIT-Res and PDM validation studies are currently in preparation. Further information available by contacting the lead author at jdubois@wustl.edu.

These examples demonstrate that even two vignette measures aimed at assessing the same construct can vary a great deal. The PDR represents more of a mastery test that demonstrates whether a respondent has or has not grasped professional decision making in the research setting. The PDR presents some advantages: it is appropriate across general fields of research, it requires approximately 50% less time to complete, and its reading-level is substantially lower than the EDM, making it more suitable for researchers who speak English as a second language. However, the EDM presents more nuanced responses and provides multiple scoring systems. Thus, the EDM should be more sensitive to detect changes due to instruction with "normal" populations (vs. outliers or those requiring remediation), and it provides more specific insight about where instruction might require modifications.

Thus, instructors must consider the tradeoffs inherent in the measures they select, and they must be explicit about the assumptions of a test.

Ethical sensitivity in research

Ethical sensitivity describes an individual's ability to recognize the ethical issues embedded in a situation, which is essential before one can then go about addressing them (11). Several researchers contended that this skill should be assessed separately from ethical problem-solving (11, 8, 31). However, because traditional measurement tools relied on time-consuming coding of transcribed interviews or written responses, this measurement approach was cumbersome. Borenstein and colleagues' work aimed to address this limitation by providing a more objectively scored measure that presents options regarding ethical issues in scenarios, followed by respondent ratings of their relevance (10). More research is needed to determine whether the validity of sensitivity scores can be maintained with this testing format.

Knowledge of research ethics and regulations

Most tests of knowledge are developed by instructors; this is legitimate, as knowledge is the most straightforward objective to assess. However, there are guidelines for writing valid items that are frequently violated. As a general rule, to improve item reliability, items should avoid: true/false format; extensive use of options such as "all of the above" or "none of the above"; item stems that ask learners to identify the option that is not true; item options that are of unequal length or nonparallel forms (18). Also, constructing a knowledge test requires considerations regarding the breadth and depth of topics to be included. Are all objective knowledge topics equally important to assess? What depth of knowledge is necessary? Is advanced or cursory knowledge of this content necessary? Examples of knowledge tests are discussed in Table 2; however, they are not widely distributed (36).

Attitudes toward research ethics and compliance

Changes in attitudes are often desirable learning outcomes (21), but they have received limited attention in RCR assessment. Attitudes shape thinking and motivate behavior, so an instructor might reasonably ask: Did students gain a greater appreciation for the significance of ethics in research? Do students believe that unethical behavior is a concern for a select few "bad apples," or do they believe that the pressures of science can influence any researcher to make a career misstep?

Perhaps these questions have not been examined because they appear rather subjective. Scientists are accustomed to assessing objective outcomes with right or wrong answers. How one determines the "right," or ideal, answer on an attitude test is partly a matter of judgment.

The How I Think about Research (HIT-Res) instrument described in Table 2 provides an example of a measure that an instructor might use to gauge a researcher's commitment to various research ethics and compliance expectations (14).

A second measure in Table 2, the Norms and Counter-Norms in Science survey, assesses respondents' perspectives on behaviors that represent norms and counter-norms in science (3). It elicits information on the norms participants think should represent behavior in science, those that do represent behavior, and those that represent their own behavior.

An ongoing project by the authors of this paper (IR-ORI-14-001-018712) will develop two measures: the Evaluating Rules and Norms in Science Task (ERNST) and the Rating Values in Science Task (RVST). The ERNST will examine the importance researchers attach to statements illustrating research regulations, norms, and counter-norms and the importance they think research administrators attach to the same. The RVST will assess the importance researchers attach to different general values in science.

So, how do I use such measures in assessment?

The most common way of using educational tests with validated psychometric properties is to administer a pretest before a course (or educational intervention) and a posttest after the course. Paired sample *t*-tests will indicate whether scores are significantly different following the intervention and whether they moved in a positive or negative direction.

In deciding whether to use a measure to evaluate individual learners (e.g., assigning grades), consider (a) whether it is reasonable to hold the learner accountable for making progress on the underlying trait (such as an attitude or problem-solving skills) based on the intervention you provided (preliminary data will help in this determination), and (b) whether the measure is sufficiently valid and reliable to use for this purpose. Consider whether it is appropriate for learners to receive completion credit, even if individual scores are not used to assign grades, particularly if there is a substantial time burden associated with completing the tests.

Concluding reflections

We strongly support the growing attention paid to moral climate, stress management, and interpersonal skills such as conflict resolution and leadership (25, 12); yet, given space limitations, we have focused on just four traditional objectives for RCR instruction. These learning outcomes enable and support research integrity. But, can we go further? Is it possible to assess whether RCR instruction increases research integrity?

Often the question posed is whether RCR instruction reduces misconduct (fabrication, falsification, and plagiarism). Measuring behavior is problematic, but measuring misconduct is particularly problematic (34). These behaviors are rare and difficult to detect in a timely manner. On the other hand, it might be feasible to assess whether RCR instruction influences observable good behaviors and best practices for responsible conduct, such as holding regular project team meetings, keeping good records, or sharing written data management procedures among team members. Self, peer, or mentor reports could capture these behaviors (although not without limitations—thus the need for validation).

The points made in this article will seem obvious to those trained in educational psychology or measurement. Nevertheless, there are several reasons why we believe these points need to be disseminated broadly within RCR education.

First, the published literature indicates that some instances in which RCR education fails to demonstrate positive outcomes are due to a mismatch of objectives with assessment. For example, courses that focus on fostering ethical sensitivity and knowledge of rules for research should not be expected to increase principled moral reasoning as measured by the Defining Issues Test (5).

Second, many RCR programs are not assessed at all. An informal survey (approved by the Vanderbilt University Institutional Review Board) of RCR instructors at institutions with NIH Clinical and Translational Science Awards (CTSAs) found that only 2 of 37 respondents reported using a validated measure to assess learning outcomes; most use only quizzes developed by instructors (which may be fine for assessing declarative knowledge) and course evaluations (which provide student satisfaction data) (J. M. DuBois and E. Heitman, unpublished data).

Thus, while a consensus exists that RCR education should address more than declarative knowledge, few programs aspire to assess more robust objectives, and those that do frequently use instruments developed by instructors that lack validation evidence. Why do programs fail to conduct assessment or use inappropriate measures? Several potential explanations exist. Experts in a particular scientific field typically instruct ethics courses, but they are not trained in methods for measurement, assessment, and educational evaluation. RCR programs also encounter time and resource limitations, and effective instructional design and assessment are resource intensive. Often course content becomes a focus with assessment an afterthought. Furthermore, instructors and program directors may focus most directly on complying with training mandates versus demonstrating program effectiveness. As educators, we tend to assume that some education is better than none. But, we cannot assume that any kind of RCR education is better than none (4).

It is necessary for instructors and program directors to be patient with assessment. Initial results may be disappointing. If so, this information should provoke questions such as: Are the right outcomes are being assessed? Are learning methods aligned with learning objectives? How might the course be revised?

We owe busy trainees and researchers instruction that is informed by data. It is time for RCR education to become evidence-based.

ACKNOWLEDGMENTS

This work was supported by NIH CTSA Grant Number ULI TR000448. The authors declare that there are no conflicts of interest.

REFERENCES

1. **AAMC-AAU.** 2008. Protecting patients, preserving integrity, advancing health: accelerating the implementation of COI policies in human subjects research. AAMC-AAU, Washington, DC.
2. **Anderson, M. S., A. S. Horn, K. R. Risbey, E. A. Ronning, R. De Vries, and B. C. Martinson.** 2007. What do mentoring and training in the responsible conduct of research have to do with scientists' misbehavior? Findings from a National Survey of NIH-funded scientists. Acad. Med. **82:**853–860.
3. **Anderson, M., E. A. Ronning, R. De Vries, and B. Martinson.** 2010. Extending the Mertonian norms: scientists' subscription to norms of research. J. High. Educ. **81:**366–393.
4. **Antes, A. L., X. Wang, M. D. Mumford, R. P. Brown, S. Connelly, and L. D. Devenport.** 2010. Evaluating the effects that existing instruction on responsible conduct of research has on ethical decision making. Acad. Med. **85:**519–526.
5. **Antes, A. L., et al.** 2009. A meta-analysis of ethics instruction effectiveness in the sciences. Ethics Behav. **19:**379–402.
6. **Antes, A. L.** 2014. A systematic approach to instruction in research ethics. Account. Res. **21:**50–67.
7. **Bates, R.** 2004. A critical analysis of evaluation practice: the Kirkpatrick model and the principle of beneficence. Eval. Program Plann. **27:**341–347.
8. **Bebeau, M. J., J. R. Rest, and C. M. Yamoor.** 1985. Measuring dental students' ethical sensitivity. J. Dent. Educ. **49:**225–235.
9. **Bloom, B. S., M. D. Engelhart, E. J. Furst, W H. Hill, and D. R. Krathwohl.** 1956. Taxonomy of educational objectives: the classification of educational goals. Handbook I: cognitive domain. David McKay Company, New York, NY.

10. **Borenstein, J., M. J. Drake, R. Kirkman, and J. L. Swann.** 2008. The test of ethical sensitivity in science and engineering (TESSE): a discipline-specific assessment tool for awareness of ethical issues. Annual ASEE Conference, American Society for Engineering Education, Pittsburgh, PA.
11. **Clarkeburn, H.** 2002. A test for ethical sensitivity in science. J. Moral Educ. **31:**439–453.
12. **Cohen, C. M., and S. L. Cohen.** 2012. Lab dynamics: management and leadership skills for scientists, 2nd ed. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
13. **DeVellis, R. F.** 2003. Scale development: theory and applications. Sage, Thousand Oaks, CA.
14. **DuBois, J. M.** 2013. Validating outcome measures for remediation of research wrongdoing. Office of Research Integrity Newsletter **21(4):**2.
15. **Dubois, J. M., and J. M. Dueker.** 2009. Teaching and assessing the responsible conduct of research: a Delphi Consensus Panel Report. J. Res. Adm. **40:**49–70.
16. **Fink, L. D.** 2013. Creating significant learning experiences: an integrated approach to designing college courses. Jossey-Bass, San Francisco, CA.
17. **Gould, M. S., T. Greenberg, D. M. Velting, and D. Shaffer.** 2003. Youth suicide risk and preventive interventions: a review of the past 10 years. J. Am. Acad. Child Adolesc. Psychiatry **42:**386–405.
18. **Haladyna, T., and S. Downing.** 1986. Validity of a taxonomy of multiple-choice item-writing rules. Appl. Meas. Educ. **2:**51–78.
19. **Hicks, J.** 2013. Opinion: ethics training in science. Scientist. [Online.] http://www.the-scientist.com/?articles.view/articleNo/35543/title/Opinion--Ethics-Training-in-Science/.
20. **Hren, D., M. Marusic, and A. Marusic.** 2011. Regression of moral reasoning during medical education: combined design study to evaluate the effect of clinical study years. PLoS ONE **6:**e17406.
21. **Kalichman, M. W., and D. K. Plemmons.** 2007. Reported goals for responsible conduct of research courses. Acad. Med. **82:**846–852.
22. **Kraiger, K., J. Ford, and E. Salas.** 1993. Application of cognitive, skill-based, and affective theories of learning outcomes to new methods of training evaluation. J. Appl. Psychology **78:**311–328.
23. **Krathwohl, D. R.** 2002. A revision of Bloom's taxonomy: an overview. Theor. Pract. **41:**212–218.
24. **Martinson, B., A. L. Crain, M. Anderson, and R. DeVries.** 2009. Institutions' expectations for researchers' self-funding, federal grant holding, and private industry involvement: manifold drivers of self-interest and researcher behavior. Acad. Med. **84:**1491–1499.
25. **Martinson, B. C., C. R. Thrush, and A. L. Crain.** 2013. Development and validation of the Survey of Organizational Research Climate (SORC). Sci. Eng. Ethics **19:**813–834.
26. **McCormack, W. T., and C. W. Garvan.** 2014. Team-based learning instruction for responsible conduct of research positively impacts ethical decision-making. Account. Res. **21:**34–49.
27. **Messick, S.** 1995. Validity of psychological assessment: validation of inferences from persons' responses and performances as scientific inquiry into score meaning. Am. Psychol. **50:**741–749.
28. **Mumford M. D., et al.** A sensemaking approach to ethics training for scientists: preliminary evidence of training effectiveness. Ethics Behav. **18:**315–339.
29. **Mumford, M. D., et al.** 2009. Exposure to unethical career events: effects on decision-making, climate, and socialization. Ethics Behav. **19:**351–378.
30. **Mumford, M. D., et al.** 2006. Validation of ethical decision-making measures: evidence for a new set of measures. Ethics Behav. **16:**319–345.
31. **Myyry, L., and K. Helkama.** 2002. The role of value priorities and professional ethics training in moral sensitivity. J. Moral Educ. **31:**35–50.
32. **Nasca, T., I. Philibert, T. Brigham, and T. Flynn.** 2012. The next GME accreditation system: rationale and benefits. N. Engl. J. Med. **366:**1051–1056.
33. **Resnik, D., and G. E. Dinse.** 2012. Do U.S. research institutions meet or exceed federal mandates for instruction in responsible conduct of research? A national survey. Acad. Med. **87:**1237–1242.
34. **Resnik, D. B.** 2014. Editorial: does RCR education make students more ethical, and is this the right question to ask? Account. Res. **21:**211–217.
35. **Self, D. J., M. Olivarez, and D. C. J. Baldwin.** 1998. The amount of small-group case-study discussion needed to improve moral reasoning skills of medical students. Acad. Med. **73:**521–523.
36. **Taylor, H. A., N. E. Kass, J. Ali, S. Sisson, A. Bertram, and A. Bhan.** 2012. Development of a research ethics knowledge and analytical skills assessment tool. J. Med. Ethics **38:**236–242.
37. **West, S. L., and K. K. O'Neal.** 2004. Project D.A.R.E. outcome effectiveness revisited. Am. J. Public Health **94:**1027–1029.