

Evaluating the Effectiveness of Training Strategies: Performance Goals and Testing

Wellesley R. Foshay and Peggy T. Tinkey

Abstract

The Public Health Service policy, Animal Welfare Act regulations, and the *Guide for the Care and Use of Laboratory Animals* all require that institutions provide training for personnel engaged in animal research. Most research facilities have developed training programs to meet these requirements but may not have developed ways of assessing the effectiveness of these programs. Omission of this critical activity often leads to training that is ineffective, inefficient, or unnecessary. Evaluating the effectiveness of biomedical research and animal care training should involve a combination of assessments of performance, competence and knowledge, and appropriate tests for each type of knowledge, used at appropriate time intervals. In this article, the hierarchical relationship between performance, competence, and knowledge is described. The discussion of cognitive and psychomotor knowledge includes the important distinction between declarative and procedural knowledge. Measurement of performance is described and can include a variety of indirect and direct measurement techniques. Each measurement option has its own profile of strengths and weaknesses in terms of measurement validity, reliability, and costs of development and delivery. It is important to understand the tradeoffs associated with each measurement option, and to make appropriate choices of measurement strategy based on these tradeoffs arrayed against considerations of frequency, criticality, difficulty of learning, logistics, and budget. The article concludes with an example of how these measurement strategies can be combined into a cost-effective assessment plan for a biomedical research facility.

Key Words: animals, laboratory; assessment; education; education, continuing/methods; laboratory personnel/education; measurement; testing; training

Wellesley R. Foshay, Ph.D., CPT, is Senior Scientist at The Foshay Group, Dallas, TX. Peggy T. Tinkey, D.V.M., DACLAM, is Associate Professor of Comparative Medicine at The University of Texas MD Anderson Cancer Center, Houston, TX.

Address correspondence and reprint requests to Dr. Wellesley R. Foshay, The Foshay Group, 9614 Hillview Drive, Dallas, TX 75231, or email rfoshay@foshay.org.

Why Evaluate Effectiveness?

In a biomedical research facility, it is essential for personnel to do their jobs well. Appropriate training may be a prerequisite, but it is not the goal. Depending on how it is done, training may or may not lead to competence in performing the procedures of the research laboratory or animal facility. Conversely, competence in these areas may or may not result solely from training. Furthermore, those who are competent (i.e., know how to do the tasks) may or may not actually perform well in the laboratory or facility. The distinctions between training, competence, and performance are fundamental to an understanding of how to evaluate the effectiveness of training in a laboratory environment. We believe that this understanding should be based on assessment and evaluation strategies that are substantially different from what is commonly used in academic contexts. The types of tests one typically takes in college are not a good model for evaluating animal laboratory performance.

In this article, we first compare the tradeoffs among various performance measurement options and then examine the problem of what to measure. Finally, we describe an example of a performance measurement solution strategy that combines methods in a cost-effective way. The reference list includes sources for additional information on all of these topics.

Tradeoffs in Performance Measurement

The purpose of performance assessment in a research context is to provide feedback to laboratory or facility managers and workers on the quality of the work performed in particular areas, with the goal of achieving and maintaining high-quality performance. For this reason, it is necessary to measure those aspects of performance that are the most important. The strategy should be to assess on-the-job performance, which is defined by performance objectives—descriptions of each laboratory task and its necessary level of performance. Performance objectives should be developed for every research task that it is worthwhile to manage through a process called *job task analysis* (JTA) and, in some cases, a more elaborate process called *cognitive task analysis* (CTA) (Jonassen et al. 1999).

Such a measurement strategy differs in important ways from conventional testing performed in academic contexts. The familiar “course final examination” is usually some kind of knowledge test. Often the questions are posed in the

abstract, rather than as requests for use of the knowledge in context. For example, an abstract knowledge question might ask, “What regulations apply to disposal of animal waste?” By comparison, in placing the question in context, one might ask, “It’s time for Mary to collect the waste from the rabbit cages. Here’s how she does it: [Provide description of Mary’s performance here.] How well does Mary’s method correspond to the regulations for disposal of animal waste?” The contextualized question should require use of the knowledge in a way that is as similar as possible to the way the knowledge is used in the animal facility. In this example, the knowledge question tests for recall of regulations, whereas the contextualized question tests for deeper understanding because it asks for application of the regulations in a situation that is somewhat similar to the way the knowledge is used in an actual facility to monitor performance. For more information on test design, see Shrock and Foshay (1984) and Haladyna (2004).

Of course it is possible for a laboratory animal worker to correctly answer a written question such as the one described above but still be unable to handle animal waste

correctly in an actual research facility. At best, any written test is an indirect measure of performance. For this reason, the best measure of performance is to observe it directly in a real situation, with all of the complications of a real work environment that compete for the worker’s attention, time, and resources. In other words, direct measures are usually more *valid* than indirect measures. However, direct observations (often done by a trained observer with the aid of a checklist) are more time consuming than written tests. Thus, there is often a tradeoff between validity of the performance measure and its cost. The comparisons in Table 1 illustrate several tradeoffs. The example in the table is of a simple, routine task—changing mice into clean cages (cage change). Such tradeoff factors are further discussed by Shrock et al. (2000).

Additional tradeoffs also exist and must be considered. For example, one must remember that all measures have error. This caveat applies to written tests and performance observations as well as to a pH meter or a balance. In tests and performance measurement, error is quantified as *reliability*. Techniques for maximizing the reliability of each

Table 1 Tradeoffs between knowledge recall tests, knowledge application tests, and performance measurements

Type of test	Example design	Level of validity	Cost/time
Written knowledge recall test	A series of written questions asking the student to write the steps in the procedure, or to select the steps and place them in order	Low: Ability to recall and verbally describe the procedure is independent of the ability to do the procedure in the animal room	Low: Typically about 20 min per question to draft, and a few seconds or minutes per question to score
Written knowledge application test (performance-based written test)	A verbal, graphic, and/or video depiction of a cage change. The student is asked to select the correctly performed steps and place them in order, or to identify procedural errors and select or describe correctly performed steps. May allow use of a reference card or other performance aid available in the animal room	Moderate: Tests ability to identify and describe examples of performance, which is a better indicator of understanding than is a recall test. However, this test requires greater verbal skill than the actual job performance, which weakens validity.	Moderate: Typically ½ to 1 hr per question to draft. Scoring can be done in a few seconds or minutes per question.
Performance measurement	Intermittent observation of the performance of the procedure during normal work hours, by verifying that the person is in his/her assigned work area during expected times, and by monitoring the number of cages per day that the person is changing. The supervisor also spot checks after a change-out to ensure that cages are not being skipped, that the proper level of food and water has been provided, and that the equipment is being assembled correctly.	High: Directly measures actual performance across a series of data points (to reduce the likelihood of stimulating above-average performance as a result of being observed)	High: Usually done with an observer’s checklist, which can be developed in a few minutes per point. Observers must be trained to use the checklist uniformly. Each observation is time consuming.

type of measure are well understood, and they apply both to the design of measurements and to their use. Ironically, it is usually true that written tests can attain a higher degree of reliability compared with observational tests because observations have more potential sources of error. For example, different supervisors may have different subjective standards of how each task should be performed; supervisors may be under pressure to approve the work of their own employees; or supervisors may feel time pressure when doing the observations, and tend to rush or be distracted. For these reasons, practitioners not only should take precautions to maximize reliability but also should carefully consider the tradeoffs between reliability and validity when designing an assessment strategy.

The issues of cost associated with typical measures are also compared in Table 1. In general, direct observation is the least expensive method to develop but the most expensive to use. Written tests are moderately costly to develop, and the cost of using them can be low if the tests are machine scored and comprised of multiple-choice questions. However, open-ended questions typically require scoring by a human rater (although technologies for machine scoring do exist), so the cost of using them is higher.

Although Table 1 includes only paper-and-pencil tests and direct observations, other measurement strategies are sometimes acceptable. For example, some animal care tasks result in work products in the form of photographic and written logs, test results, and the like, which may provide acceptable indicators of work quality. It also may be possible to examine measures related to the job tasks such as resources consumed, time records, quality control records, defect rates (e.g., iatrogenic infection records), animal health records, animal excretions. These measures are most commonly indirect indicators of job task performance and thus may be subject to validity concerns. Nevertheless, they often are available at little or no incremental cost and thus may be useful as early indicators of performances that are out of compliance with job task standards and merit further assessment.

A simple self-report checklist may be sufficient for some purposes. Simply asking trainees questions such as, "I know how to clean a rat's cage properly [yes/no]" can be a relatively valid indicator of competence *if* the rater understands the standard being applied (in this case, standards for clean cages) and *if* the trainee does not anticipate that negative consequences might result from a negative self-assessment. Thus, self-reports are often useful for quickly assessing training needs or for providing a crude inventory of competence in noncritical tasks (Norwick et al. 2002). Self-reports should be used with caution, however, in assessments that involve a high level of criticality.

It may be possible to justify the cost of utilizing electronic delivery of tests and simulations in the following cases: (1) if the number of people to be measured is large or dispersed by geography or time; (2) if the task is of high criticality and/or high cost; or (3) if the task is performed only rarely. It is possible to efficiently administer electronic

tests remotely and on demand, whereas for security reasons, it is usually necessary to administer written tests to groups in a few well-controlled locations at a few well-controlled times. In addition, with electronic tests it is possible to use a variety of questioning techniques that are more sophisticated than what can be accomplished on paper. One such technique is *adaptive testing*, which dynamically varies the questions asked according to the performance of the testee. The result is greatly improved reliability and validity without lengthening the test unacceptably. Suitable precautions for online test security are needed. For performance measures, simulations can sometimes serve as valid substitutes for direct work observation. Although the cost of developing simulations and adaptive tests is high, the cost of administering these tests is low. Under some circumstances, this tradeoff profile may be attractive.

In any animal research facility, there are far more tasks than it is feasible to assess. Therefore, another set of tradeoffs to consider involves deciding what to assess and how frequently. In general, it is wise to select the job tasks that are characterized in the following ways:

- The most frequently performed;
- The most critical to the mission of the facility, to safety, or to regulatory compliance;
- The most difficult to learn.

It also may be useful to develop a *sampling plan*, in which the tasks of greatest concern are assessed at a regular interval compared with other tasks, which are assessed less frequently and randomly (see Shrock et al., op. cit.).

To summarize, various indirect and direct measurement techniques are described in the literature. Indirect measures include written or electronic tests of knowledge recall or knowledge application, indirect performance measures such as work products and processes, and simulations of different types. Direct measures usually are performance observations. Each measure has its own profile of strengths and weaknesses in terms of measurement validity, reliability, and costs of development and delivery. It is important to understand the tradeoffs associated with each measurement option and to make appropriate choices of measurement strategy based on these tradeoffs while at the same time giving careful consideration to frequency, criticality, difficulty to learn, logistics, and budget. A common error is to measure what is easy or inexpensive to measure without considering measurement validity, reliability, or importance. Such an ill-considered strategy produces misleading information on laboratory and worker performance.

Performance Versus Competence Versus Knowledge

Having examined the alternatives for measurement strategy, it is now important to discuss the important distinctions between performance, competence, and knowledge. Evaluating the effectiveness of animal care training typically

should involve assessments of all three areas, and it is important to understand the value of each type of assessment. The distinctions are summarized in Table 2 and are further discussed by Hale (2002) and Addison and Haig (2006).

Performance

Performance is what is actually done in the research laboratory or animal facility. Knowing how to perform a research or animal procedure is an essential, but not sufficient, prerequisite to performance. The following simple example is provided:

Animal care employees are often faced with conflicting priorities. Two common conflicts are the need to accomplish a large amount of work while, at the same time, adhering to time-consuming but important safety regulations. The use of personal protective equipment is required in biocontainment facilities. This equipment often includes protective jumpsuits, head and shoe covers, and specially fitted facemasks. Employees who need to move in and out of biocontainment facilities frequently spend significant time simply putting on and removing this gear. The pressure to accomplish work in a time-efficient method can result in an employee making the choice to forego the protective clothing requirements for tasks that they judge to be of low risk such as entering the facility to check room temperatures or lighting levels.

As the example shows, a host of factors influence performance. These factors include but are not limited to the time pressure of competing tasks, the availability of necessary tools and resources, competing incentives for good or poor performance, and personal motivation. Thus, in the context of the real world of work, training is not sufficient to ensure performance.

Because actual performance in the facility is what ultimately matters, it is important for the assessment plan to include measures of actual performance. However, a common error is to assume that when the measures reveal a performance problem, there is a need for more or better training. Often the cause is not related to training. In such as case, if corrective action is limited to training, performance will not improve and the investment in further training will not be justified.

Competence

Competence is knowing how to accomplish each animal care or research task. It is the ability to perform the required task under controlled circumstances, without the competing pressures of subsequent, independent (“real-world”) work. The following example will illustrate:

Most hands-on technical procedures of biotechnology are taught “apprentice-style.” These procedures include most animal manipulations such as blood collection or substance administration. An instructor who is competent in the procedure demonstrates the procedures to a small number of students. Instruction may involve some background reading material provided or a video demonstration of the procedure, but most of the training consists of verbal instruction in conjunction with actual demonstration of the procedure. The instructor then requires the students to perform the procedure on an animal under working conditions while he observes. Initial training may begin with inanimate objects to allow the students to become comfortable handling a syringe and needle and then progress to a live animal, as appropriate. When the instructor has observed the successful completion of the procedure, the student is then allowed to perform that procedure independently. Follow-up training may occur at the request of the student or if a problem is seen with the animals that are handled by a particular student.

Assessment of competence involves performance of the technical procedures under the controlled circumstances of a training session or an apprentice relationship. If the students perform the tasks correctly, then the instructor is justified in concluding that the students *know how* to do the tasks. However, the instructor cannot conclude that the tasks are actually performed the same way outside the training/apprenticeship context (i.e., in real-world work). Only an on-the-job performance measure can confirm that level of performance.

Again, the distinction is important because effective training should result in competence but it cannot guarantee performance. Competence is a necessary, but not sufficient, component of performance. If the goal is to evaluate per-

Table 2 Comparison of performance, competence, and knowledge references

	Definition	Example
Performance	Performing a task in a real-world work situation	Cleaning a room full of cages as part of a daily work schedule
Competence	Performing a task under controlled conditions	Cleaning a single cage while being observed by an instructor
Knowledge	Understanding a task	Describing the steps in cleaning a cage and/or separating properly cleaned cages from improperly cleaned ones

formance, then some combination of measures of competence and performance are often used. However, if the goal is only to evaluate the effectiveness of training, then competence measures alone may be sufficient.

Knowledge

Knowledge concerns how well students understand what is being taught. Knowledge involves the following:

- “Knowing what” (e.g., familiarity with laboratory animal science terminology and specifications)—*facts*;
- “Knowing why” (e.g., understanding the causes of iatrogenic infection)—*concepts and principles*; and
- “Knowing how” (e.g., the ability to translate an understanding of the causes of iatrogenic infection into procedures for handling laboratory animals)—*procedures*.

Cognitive learning theory often groups fact, concept, and principle learning as “declarative knowledge” and makes an important distinction from procedural knowledge. Often “understanding” is taken to mean a combination of knowing why and knowing how—or declarative knowledge. However, learning psychologists have demonstrated that declarative and procedural knowledge are relatively independent of each other, so that it is quite possible to know what and know why without knowing how. By these definitions, it is possible to understand a procedure and still not be able to do it, and it is possible to be able to do it without understanding it. These distinctions are elaborated by Gagne and Medsker (1995) and Anderson et al. (2001).

The following example illustrates the distinction:

Many programs require technicians to be certified by the American Association of Laboratory Animal Science (AALAS¹). The AALAS technician certification program is a self-paced program that consists of material contained in textbooks. Individuals who believe that they have mastered the material are required to take and pass a computerized test (questions and answers, multiple choice) to receive certification. The bulk of the testing focuses on facts (knowing what) and principles (knowing why) rather than procedures (knowing how). The reasons for this focus is that the test must be administered to people from a wide variety of biomedical research facilities, and individual facilities use a vast array of procedures specific to their own business practices. Thus, the knowledge of terms and principles regarding effective equipment sterilization practices can be tested, but the actual procedures for accomplishing the act of sterilization may be different at each facility. Furthermore, knowing the definition of “steam sterilization” versus “chemical sterilization” and the parameters

required to accomplish effective sterilization (e.g., the minimum time and temperature required to achieve sterility using steam methods) does not ensure that an individual can operate the equipment in his or her own facility.

Because animal research procedures often involve tasks such as handling animals, cages, feed, and medications, it is also important to understand the distinction between the *cognitive* domain and the *psychomotor* domain in learning. Cognitive learning typically involves verbal, auditory, and visual learning. Psychomotor learning typically involves muscle movements with associated senses of touch and smell. The two types of learning are quite independent of each other. For example, the ability to verbally describe the factors that contribute to an assessment of the condition of a laboratory animal is essentially independent of the ability to actually make the assessment, by handling real animals in a real research environment.

Written Tests and Time Intervals

The distinction between the types of learning described above is important because written tests require knowledge of the respective language. Written tests often produce results that are related only to declarative knowledge (written tests of procedural knowledge are sometimes possible, but they are relatively uncommon). Thus, written tests are often removed from measurement of competence by two degrees because there is often no test of procedural knowledge and no test of psychomotor knowledge. Furthermore, verbal ability mediates skill in taking written tests. An animal care worker may be excellent in handling animals but have a limited ability to read, write, or even converse—whether in English or another language. For such people, the validity of a written test is questionable.

Another factor to consider is the time interval between training and measurement. Knowledge that is not used is rapidly forgotten; a test administered upon completion of a training course typically shows much higher scores than a similar test administered days or weeks later. Because it is common for performance, competence, and knowledge measures to be used at different time intervals, it is scarcely surprising that measurement results can lead to very different conclusions about the effectiveness of training.

We are not arguing against the use of written tests; instead, we are describing what they can—and cannot—do well. Written tests can be very useful when it is necessary to measure only the understanding of declarative knowledge in the cognitive domain, and when questions about the student’s verbal ability are minimal. Thus, written tests are likely to be appropriate measures of the effectiveness of some kinds of training experiences that are part of facility training. Nevertheless, it is unlikely that a written test alone will suffice as the sole measure of effectiveness of training for a laboratory animal training program.

¹Abbreviation used in this article: AALAS, American Association for Laboratory Animal Science.

Table 3 Plan for evaluating laboratory training at Global Pharma, Inc.

Training component	What to assess	How to measure
Course #1: Animal terminology and concepts	Knowledge: Basic husbandry, handling, routes for substance administration, routes and methods for body fluid sampling (e.g., blood, urine), ability to perform drug dosage calculations	Written test of terms and concepts involving generation or identification of concept examples and principles, drug dosage calculation, word problems
Apprenticeship #1: Basic care of laboratory rats and rabbits	Competence in procedures: Assessment of normal posture and behavior, handling methods, weighing, administration of substances by oral and parenteral routes, blood collection, calculation and preparation of drug dosages	<ul style="list-style-type: none"> ● Instructor observes students performing procedures to ensure correct performance. ● Instructor reviews drug calculations and prepared doses for accuracy.
Performance evaluation #1a: Review of animal records	Performance assessment: Indirect measure of performance based on adverse health events reporting	Animal health records are reviewed for trends that indicate performance problems (e.g., trauma from poor handling or infections from poor substance administration techniques).
Performance evaluation #1b: Observation of animal procedures during ongoing animal study	Performance assessment: Judgment of quality of demonstrated techniques of animal handling, substance administration, blood collection, and condition of animal	Veterinarian or veterinary technician spot checks scheduled procedures being performed by animal care staff. Assessment is made of adherence to required technique, ease of performance, impact on animal, and overall quality of performance. Drug doses and volumes are spot checked for accuracy.
Course #2: Radioactive materials safety course	Knowledge: Types of radiation, units of measure, equipment for measuring radiation, regulations for exposure limits	Written test of terms and concepts
Apprenticeship #2	Hands-on teaching laboratory to demonstrate competence in operating radiation detection equipment, estimation of exposure level, demonstration of shielding effectiveness. In training, instructor demonstrates Geiger counter operation	<ul style="list-style-type: none"> ● Instructor observes student operating equipment. ● Student records radiation readings of unknown samples and submits written results to instructor.
Performance evaluation #2a	Performance assessment: Checking to ensure that required safety measurements are being conducted	<ul style="list-style-type: none"> ● Supervisor reviews required radiation clearance forms for completeness and accuracy. ● Supervisor compares procedure records with radiation clearance forms to confirm that forms are being completed as required after each procedure.
Performance evaluation #2b	Performance assessment: Observation of radiation clearance procedure	<ul style="list-style-type: none"> ● Supervisor drops in to observed scheduled procedure, observes use of Geiger counter, confirms correct use of instrument, reviews completed forms for accuracy. ● Supervisor may ask procedural questions to assess understanding of procedure.

Summary

We have described a hierarchical relationship between performance, competence, and knowledge in addition to distinguishing between different types of knowledge.

Cognitive knowledge is to a large degree independent of psychomotor knowledge. Moreover, with the domain of cognitive knowledge, an important distinction exists between declarative knowledge and procedural knowledge. To evaluate the effectiveness of biomedical research training

accurately, it is important to use a combination of performance, competence, and knowledge assessments, to use appropriate tests for each type of knowledge, and to administer tests at appropriate time intervals. A written test alone, no matter how carefully constructed, cannot serve as a complete measure of the effectiveness of an animal research training program.

Combining Performance Measures to Fully Evaluate Effectiveness of Laboratory Training

We conclude with a hypothetical example of a combined assessment strategy for a typical laboratory animal technician training program designed to cover basic animal handling and radiation safety. This example is presented in Table 3, in which assessment progresses from written knowledge tests, to observational competence tests in apprenticeship settings, to spot check observations of actual work in the animal facility. The assessment strategy includes both cognitive and psychomotor measures, and uses each measure for appropriate content assessment. Performance measures include both direct observations and indirect measures of work products. The resulting assessment plan strikes an appropriate balance between cost and validity of measurement, with the goal of providing an acceptably valid assessment of performance but at a reasonable cost of development and use.

Conclusion

A successful strategy for evaluating the effectiveness of animal technician training requires careful analysis of the job tasks for which the training will take place. It is important to carefully prioritize the evaluation so that its primary focus is on the animal care tasks that are most frequent, critical, and difficult to learn. Because the training program as a whole is effective only if it results in good animal research performance, it is important to measure actual performance of the animal care tasks as well as trainee competence and underlying knowledge. We recommend using a combination of measures, which will provide the most valid and reliable data on the aspects of performance that are most important at an acceptable cost.

A well-designed and executed evaluation yields the following important benefits:

1. Trainees and their supervisors receive accurate feedback on what has been learned and what needs to be learned. This information increases the ability of instructors to focus on the training needs of each individual trainee.
2. Research facility managers know the competence level of their animal care staff. Managers can use this information to project the performance capability of staff and to know what training will be needed to implement current or new procedures.
3. If there is a discrepancy between competence and performance, then facility managers will know to look for nontraining impediments to effective performance in the laboratory.
4. The evaluation is useful for justifying the financial investment in adequate training in that it identifies the relationship between training and actual research performance. If there is a weak relationship between knowledge and competence, then the evaluation will help show how the training program should be improved.

The end result is improved animal care performance, often with a more cost-effective training program.

References

- Addison RM, Haig C. 2006. The Performance Architect's Essential Guide to the Performance Technology Landscape. In: Pershing JA, ed. *Handbook of Human Performance Technology*. San Francisco: Pfeiffer. p 35-54.
- Anderson LW, Krathwohl DR, et al. 2001. *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*. New York: Longman.
- Gagne RM, Medsker KL. 1995. *The Conditions of Learning: Training Applications*. Boston: Wadsworth Publishing.
- Haladyna TM. 2004. *Developing and Validating Multiple-Choice Test Items*. Mahwah NJ: Lawrence Erlbaum Associates.
- Hale J. 2002. *Performance-Based Evaluation: Tools and Techniques to Measure the Impact of Training*. San Francisco: Jossey-Bass Pfeiffer.
- Jonassen DH, Tessmer M, Hannum WH. 1999. *Task Analysis Methods for Instructional Design*. Mahwah: L. Erlbaum Associates.
- Norwick R, Choi YS, Ben-Shachar T. 2002. In *Defense of Self-Reports*. *The Observer* 15(3).
- Shrock SA, Coscarelli WCC, et al. 2000. *Criterion-referenced Test Development: Technical and Legal Guidelines for Corporate Training and Certification*. Washington DC: International Society for Performance Improvement.
- Shrock SA, Foshay WR. 1984. Measurement issues in certification. *Perform Instruction* 23:23-27.