Evaluation for Smarties

Lisa Beck McCauley (lisamccauley@verizon.net)
William Grisham (dr.billgrisham@gmail.com)
UCLA Modular Digital Course in Undergraduate Neuroscience Education (MDCUNE)

Why evaluate?

One reason for conducting an evaluation is to fulfill your obligations to the granting agency. An even better reason is that ANY course or program can benefit from good evaluation practices. The information you gain will inform your teaching, course design, and test preparation. Evaluation gives you the tools to improve and adapt your course from year to year.

What are your goals for the course and the evaluation process?

This step begins with the Principal Investigator (PI). Decide what content is central to the course/project, what challenges you anticipate, and what you expect from your students/project. Once your goals, priorities and expectations are clear, you can begin developing assessment items, in collaboration with the evaluator.

What should you evaluate?

- Affective responses

Scientists tend to downplay affective data in favor of "hard" numbers like test scores, but students' subjective experience goes hand in hand with content learning.

There's no secret to making affective items work for you. Use the items that you want to know the answer to such as:

- How well did the textbook work with class lectures and exams?
- Were online/digital materials worth the trouble?
- How many hours each week did you spend on this course?
- Did the prerequisites prepare you well for this course?

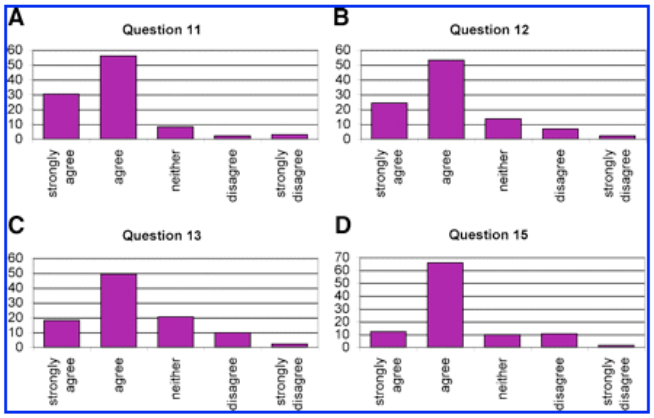


Figure 9. Percentage of respondents (n = 132) as a function of scale points. Questions were worded as follows. (A) Question 11: My understanding of bioinformatics databases was enhanced by actually doing the computer tasks and examining their data. (B) Question 12: My understanding of genetics was enhanced by the QTL (Bioinformatics) module. (C) Question 13: My understanding of statistics was enhanced by the QTL (Bioinformatics) module. (D) Question 15: I learned something about molecular biology from the QTL (Bioinformatics) module.

IDEAS: Correlate the responses with each other and with test scores; compare means for responses from students with high and low grades or test scores.

- Open-ended qualitative feedback

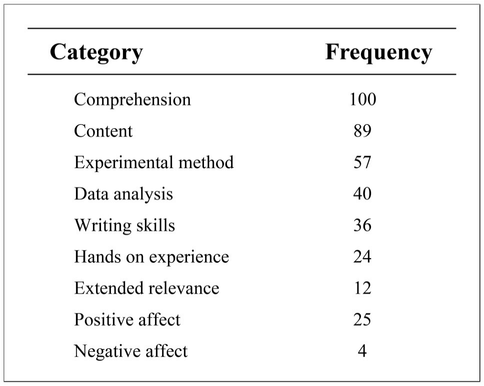
In general, people appreciate the opportunity to say what's on their minds. Take advantage of this contribution from your students, and try to make the most of their comments.

An open-ended item such as this invites students to reveal what THEY thought was most important about the class. Responses on a recent posttest ranged from general content ("sex differentiation of the zebra finch") to specific skills (writing in journal article style; data analysis), relevance beyond class, and hands-on activities. A tally of comments by category provides a students' eye view of the entire course, and an opportunity to find out how their perspective differs from your goals.

The following are responses to the open-ended question "Please describe the purpose of using the Bird Song System module from a learning standpoint."

- <u>Student 1 Response</u>: "*The Bird Song System module allowed students to conduct actual research and experience each step of the process. From slicing and staining the brains to measuring the digitized images of the song nuclei and analyzing the data, we gained hands-on experience on how to study a neurological basis for bird song. We also learned the effects of steroidal hormones on sexual differentiation in zebra finches and how we can manipulate the sexual development of a bird*."

- <u>Student 2 Response</u>: "*It was very fun and enjoyable and Dr. Grisham rocks at teaching*."



Category	Frequency
Comprehension	100
Content	89
Experimental method	57
Data analysis	40
Writing skills	36
Hands on experience	24
Extended relevance	12
Positive affect	25
Negative affect	4

The table categorizes the 136 student responses (N = 136) to the open-ended question "Please describe the purpose of using the Bird Song System module from a learning standpoint." Note that a given student's response can be coded in more than one category.

- Content learning

This is what most people associate "assessment" with. To make sure you're making the most of your time and your students' time, plan the test as carefully as you plan your lectures. Be sure to consider:

- reliability—see the **<u>Reliability</u>** section below for more information.
- validity—see the **<u>Validity</u>** section below for more information.
- item analysis—see the **<u>Item Analysis</u>** section below (page 5) for more information.

What are your goals for the course? What knowledge and skills do you need EVERY student to take away from your class? What *additional* benefit do you hope the BEST students will gain? Distill your expectations into 10 or 15 test questions.

<u>Review and Rework</u>

You may have what you need to submit your progress report, but your evaluation data still has plenty to offer. Spend some time with the data and the evaluation report, and consider what you might do differently next time in order to:

- reach more students.
- select textbooks and other course materials.
- Improve tests to conform to your teaching goals and practices.

And while you're at it, sit down with your evaluator to make plans for the next round. The evaluation process can improve from year to year, along with the program itself.

Reliability

There is NO VALIDITY without reliability! Are your assessment instruments reliable and appropriately tuned/calibrated? How can you tell?

Get a reliability check on your instrument such as:

- Split half/Spearman-Brown/Chronbach's alpha: Your evaluator ought to provide you with one of these measures—1.00 is the max (range is -1.00 to 1.00, negative values extremely bad!) What is the good range?—should be .70 and above.
- Test-retest

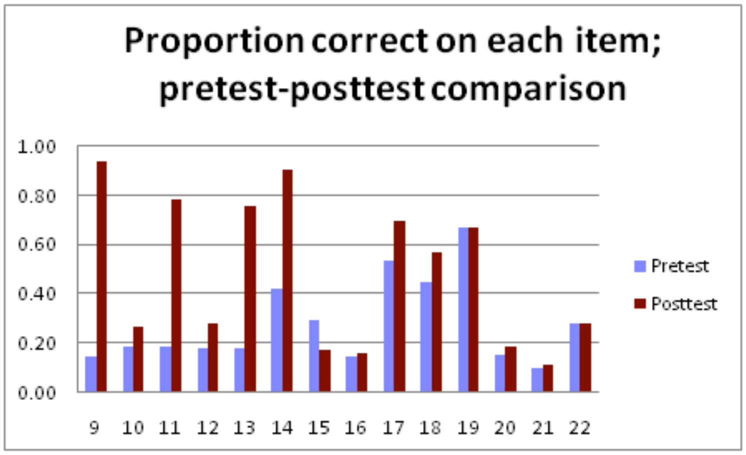
Do an item analysis—WEED OUT THE CLUNKERS. See the **Item Analysis** section below.

Item Analysis

Item analysis is useful for improving items. In particular but it can eliminate ambiguous or misleading items, which may help with reliability. Also, item analysis is helpful in identifying content areas that need greater emphasis or clarity.

Item analysis is a statistical assessment that looks for items that need further examination—items with unexpected patterns of response. Unexpected patterns call for closer examination of an item, but not necessarily for dropping the item.

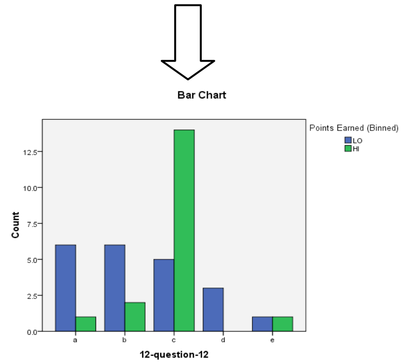
- A good item will show a change from pretest to posttest: If there's no improvement, then either the item or the course material needs to be reconsidered.



In this example, notice that there are a few items showing little or no improvement from pretest to posttest. Scores on *Item 15* actually decline from start to finish.

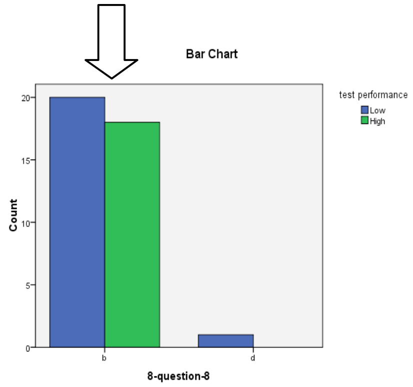
- Compare high- and low-performing students' responses: Students who do well on the test as a whole should also perform better on an individual item than students who fare poorly on the test.

  The following graph displays the number of students choosing each option on a multiple-choice test. The dark bars represent students with low scores on the complete test; the light bars represent high-scoring students.



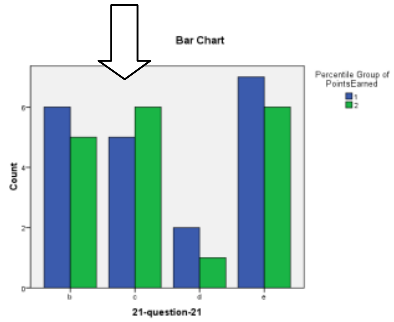
The high-performing students overwhelmingly choose the correct option for this item. In contrast, the responses of the low-performing students are pretty evenly distributed among the incorrect choices. This one's a keeper!

- Watch out for items that everyone gets right, or gets wrong: It's a good idea to include a couple of easy items to keep morale up, but these should be balanced with moderately difficult and downright hard items. A good test will provide a range of scores, from low to high.



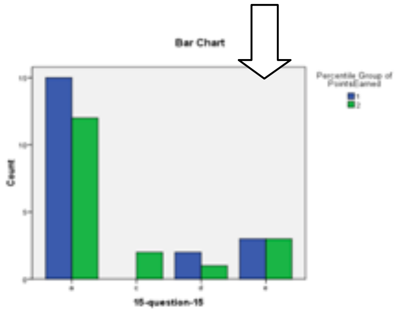
All but one student chose the correct answer here.

- Determine if the distractors are pulling their weight: The incorrect options should be reasonable enough to require thought, but not so misleading as to confuse well-prepared students who know the material.



This item may be problematic. High and low performing students respond to each of the options at about the same rate, and the correct option is lost in the crowd. A careful look at the item should reveal whether it's confusingly worded, unprepared for, or just difficult. This outcome may indicate a need to adjust teaching strategy on this point.

- Maybe it's not the item's fault: You wrote each item to reflect an important aspect of course content. Before discarding an item with a surprising pattern of responses, consider the possibility that the material was not covered adequately in class.



It's that pesky *Item 15* again. Why are both high and low performers so certain that "a" is the correct answer? This outcome may indicate a need to adjust teaching strategy on this point.

Validity

Test Validity:

- Variations of validity

Validity comes in different flavors:

- Content Validity: Is the test an adequate representation of the material?

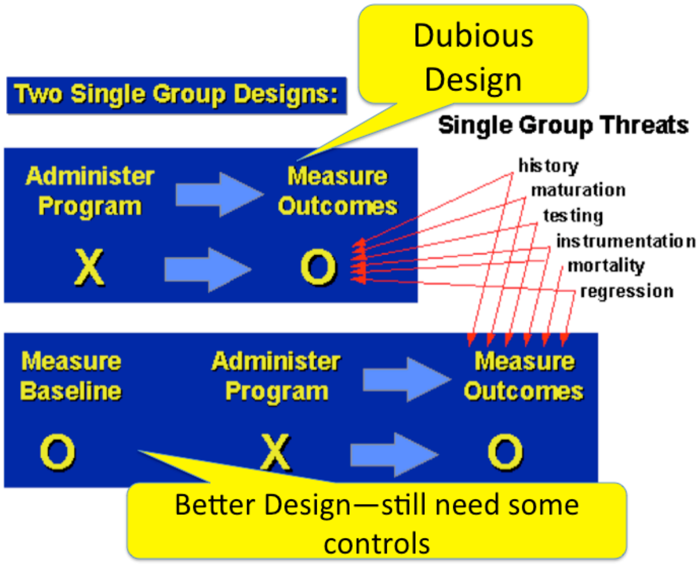
  → Get a panel of experts to judge test items.

- Construct Validity: Does the test measure what it is intended to measure?

  → A "construct" is something that we believe that people have in their heads but we can't directly observe (e.g. engineering knowledge, critical thinking, intelligence).

  → To estimate construct validity, see if the test correlates with measures intended to assess the same thing. For example, if your test assesses mastery of content in the materials, it ought to correlate well with other measures of the same construct, such as resulting course grades. (In our instance, posttest scores did correlate with the grades on the unit (r_{84} = 0.537, p < 0.001) when grades were determined by a multiple-choice and short-answer exam.) Don't be dismayed by low correlations—the relationship between intelligence and grades is about r = 0.30—constructs can be tough to measure as well.

- Challenges to validity



The following are some challenges (threats) to validity with single group designs:

- History: Events occurring between the first and second observation can impact scores. Imagine if you were doing a posttest at Columbine High School just after the massacre. Your intervention would appear to diminish test scores. **Need a control (comparison) group that does not receive intervention.**

- Maturation: Scores can change due to changes occurring in subjects due to the passage of time. i.e. Little kids are going to get more coordinated no matter what you do—including nothing. **Need a control (comparison) group that does not receive the intervention/course.**

- Testing: The effects of taking a test on second test's scores—pretest sensitization or just getting more "test savvy" can raise scores apart from interventions. **Need a control group that doesn't get the pretest but does get the intervention (cf. below).**

- Instrumentation: The changes in the instrument, observers, or scorers which may produce changes in outcomes. **Don't do this!!!**

- Experimental mortality: Not necessarily literal death but can also be dropping-out of subjects. Big problem in educational research because the weaker students tend to drop out, making any intervention look good by their absence. **Best control: throw out the data of those that do not complete the course/intervention so that you get a fair assessment of impact of intervention.**

- Statistical regression: Also known as regression to the mean. Extreme groups will tend to score closer to the mean on retest regardless what you do—part of their scores were due to "bad luck" (for poor performers) or "good luck" (for good performers). **If using extreme groups, you need a control (comparison) group that doesn't receive the intervention/course or characteristics.**

- Selection of subjects: Non-random selection of subjects biases the characteristics of group, which may confound interpretation. Self-selected students may be more motivated and perform better on posttest just because they are more motivated, not due to intervention. **Random assignment to groups counters this threat. At very least, use groups that are likely to be equivalent on important dimensions at outset.**

- Proposed designs to meet challenges

The following designs control for threats to internal validity—more work, but worth it!!!

- The Pretest-Posttest Control Group Design: Two different groups, only one gets the intervention.

  (**R** represents random assignment to condition, **O** represents observation (test), **X** represents intervention)

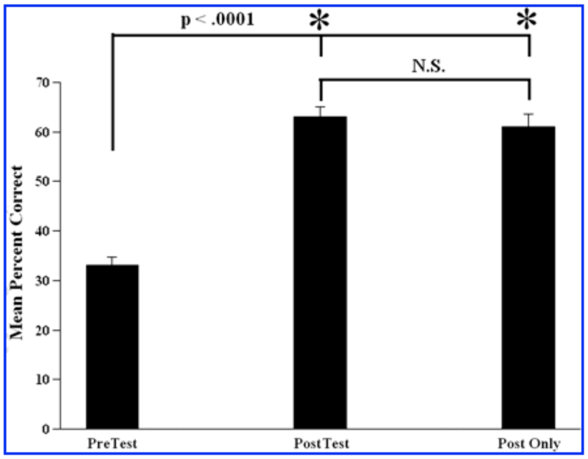
R- Exp group	**O_1**	**X**	**O_2**
R-Contol Group	**O_1**		**O_2**

  This design controls for threats to validity discussed above.

- The Pretest-Posttest and Posttest-Only Design: Two different groups, both get the intervention in Posttest, but only one gets the Pretest.

  (**R** represents random assignment to condition, **O** represents observation, **X** represents intervention)

R	**O_1**	**X**	**O_2**
R		**X**	**O**



Optimal outcome!

Figure 8. Mean (± standard error of mean) percentage correct on pretest and posttest given in one academic term and a posttest alone given in a subsequent term. Asterisks indicate significant differences as determined by a paired *t* test (for tests given in same term) and an independent *t* test for tests given in different terms. N.S., difference not significant.

Recommended Resources

Anne Anastasi (1976). Psychological Testing (4th ed).

Cook & Campbell (1979). Quasi-Experimentation: Design and Analysis for Field Settings.

Chong-ho Yu & Barbara Ohlund (2010). Threats to validity of Research Design. http://www.creative-wisdom.com/teaching/WBI/threat.shtml

Lee J. Cronbach and Paul E. Meehl (1955). Construct Validity in Psychological Tests. http://psychclassics.yorku.ca/Cronbach/construct.htm

Trochim, W.M.K. (2006). Single Group Threats. www.socialresearchmethods.net/kb/intsing.php