TEACHER-MADE TESTS IN HISTORY

Douglas D. Alder
Michael L. DeBloois
J. Nicholls Eastmond, Jr.
Utah State University

As history teachers we often write our own tests. But are we aware of the many testing options available? Perhaps one purpose we have in mind is to pace students through the subject matter. This approach assumes that tests can provide the impetus for students to study regularly. We also use tests legitimately as credentialing or ranking devices, measuring student achievement and awarding an appropriate grade. In addition to these two uses, pacing and ranking, historians find tests to be useful as diagnostic tools or as reviews. These several uses can be intermingled and, unintentionally, often are.

Sometimes we assume that our many years as history students qualify us as test writers. Perhaps some teachers understand testing without ever having read the professional literature on educational evaluation. Often when we do read in this material, we find jargon-laden observations that seem obvious and boring. Yet there are insights in the testing and measurement field that could improve our historical teaching and learning.

For example, pre-testing can serve as a diagnostic device. Few historians employ such options even though students in any given class enter with a wide variety of backgrounds. A teacher would be well advised to discover how much each individual already knows about the intended course content. This pre-test, administered at the beginning of the course, merely gathers useful information; students are neither graded nor ranked by it. Once such information is known, the history teacher may prescribe alternative learning activities for those who need remediation and especially for those who are already advanced in the subject. Another use for pre-tests may be to establish a base for measuring learning growth when compared to a concluding examination.

Weekly quizzes are a testing option that many history teachers use to help students review. A common purpose for such short tests is to give students feedback on their learning effectiveness. Since many students are apprehensive about their ability to learn, quizzes can tell them whether their studying is focused correctly. These little tests can build confidence and mold the students' motivation so they will study more and thereby perform better on the major examinations. They also help instructors determine whether their teaching is "on target" or "missing the mark."

The Essay Test

For historians, tests are commonly used to measure student learning. As such, one of the central practical questions is whether to use the essay or the so-called objective test. The former enjoys the favor of tradition. It requires students to present their answers in the highly respected format that historians use for professional communication: written narrative. Thereby, their performance is closely allied to the thought processes of the humanities which reward nuance more than specificity, complexity more than simplicity.

The essay examination is especially functional for history teachers. Essay questions can ask a student to describe principles and support them with accurate data. They can call upon students to build a case using contrasting arguments, and even give priority to the alternatives. Analysis

and synthesis skills can be tested. The wedding of data mastery with critical thinking can be promoted.

The chief characteristic of a good essay examination as well as an essay answer is clarity. To a large extent, the writing of straightforward answers is made possible by unambiguous questions. To facilitate a lucid answer, the question should be written in terms that require thought, employing such verbs as "compare, justify, explain, criticize," etc. They should avoid straight-fact questions because the purpose of the essay examination is to measure higher level cognitive processes. Clear answers can also be encouraged if examination writers include structuring suggestions such as "compare 'X' with 'Y' on at least three criteria" or "analyze the impact of 'Z' on the 'Q' decision including both political and economic factors."

While working on the clarity of each question the test writer would do well to simultaneously prepare an answer key. It should include some of the basic answers anticipated so that the responses can be graded against some standard, minimizing the subjective nature of essay grading.

The Objective Test

The type of test which includes many specific questions has attracted increasing admiration, particularly from professional testing and measurement scholars, because these examinations allow for wider coverage of content material. The answers are readily observable and the subjectivity in the correcting process is minimized. Questions developed for such tests can be retained for continued use in an expanding pool and they can be subjected to item analysis to evaluate their effectiveness. They can even be adapted to computer-assisted instruction. The speed of correction, mechanically or manually, is still another advantage. Advocates for objective tests argue that such specific questions develop a reverence for data and for verification, a key value to historians. They further feel such discrete questions promote problem-solving ability.

Many teachers feel the use of objective test items will result in simply measuring the trivial, rewarding recall and recognition abilities, while ignoring the higher levels of thinking. Since these teachers try not to thwart divergent, creative and free thinking, they worry about using tests that require uniform answers. On the other side of the question, some teachers need to realize that students are nearly professional test-takers; if they are verbally skillful or write well, able students may "get by" on essay tests through these abilities rather than by mastering the historical information. Since the study of history requires skills in data recall and information retrieval, it is defensible to legitimize those skills by measuring them in the tests. Objective tests are an efficient way to do so, independent of verbal and writing abilities. Students will likely study specific data more diligently if they know that objective-type questions will be included on the test.

Such objective questions are written either as multiple choice, matching, true-false, listing, or fill-in questions. The first three require selection skills and as such demand less rigorous thinking than recall questions. The students do not really have to draw on what they remember; they only have to recognize it from answers that are supplied. So questions that ask students to identify or list or give short paragraph answers may measure recall better.

In these, as well as the essay test, the key to success in writing questions is that they be especially clear. Test writers would be well advised to ask a colleague or student to criticize the questions for clarity before they are used. Here are some specific suggestions for writing good objective questions.[1]

True-False Questions

--Correctness should not be based on a minor point.
--Questions should not include more than one idea that is correct.
--Superlatives such as "all," "never," and "always" should be avoided.

Multiple Choice Questions

--Each choice should include just one clear idea.
--Choices should be similar in length.
--Wrong choices should seem possible to the uninformed.
--Using "all of the above" or "none of the above" as an answer should be avoided. It increases the chance of guessing the right answer.

Completion Questions

--Words to be placed in the blanks should be important ones, not trivial ones.
--Clues in the sentence or give-aways which make the question not really a test should be avoided.
--Each blank should elicit only one possible correct answer.

Matching Questions

--The items in the right column should exceed the number of items in the left column to minimize the opportunity to guess the answers.
--The test writer should state in advance if more than one combination is acceptable.

Identification Questions

--The required response should consist of only a few sentences.
--The criteria against which the answer will be judged should be stated.

The strength of the objective test is reliability, because a wide variety of information can be tested in a short time using a fairly consistent measure. As a result the objective test can be both thorough and efficient.

Take-Home Tests

Since historians usually have reference works, libraries, and colleagues available when they are engaged in the historian's craft, it seems realistic to test students occasionally in the same atmosphere. Such a test needs to emphasize the process and inquiry approaches to history more than the information retrieval. A take-home test can cause students to evaluate evidence more than recall it, to think more than write, to organize more than to remember. Since the student will take the test out of the classroom and compose an essay answer away from the teacher, the test writer needs to solve the cheating potential in advance. This is normally done by designing the questions to call for analytical or interpretive responses. Nonetheless, the teacher will need evidence that the student actually wrote the paper. One approach is to conduct a brief oral review of each student's paper.

Competence Tests

Some student achievement does not lend itself to evaluation in tests. This can be examined in various ways--diaries, video tapes, interaction analysis, or other means of systematic observation. Such competency examinations are being used increasingly in many fields to get closer to actual achievement. This form of evaluation may even reward students who are not especially talented in verbal skills. In history, one such option would be having students solve information retrieval or research problems that are given to them in a laboratory course by administering the test in the library or archives. The students would be asked to actually do a task described in a test and bring the documentary results to the teacher for credit rather than to write about it.[2] Another more traditional and verbal approach is to have students produce some history, in a standard research paper, or a less ambitious experience. This requirement has been used by historians long before the competency examination was conceived, but it is essentially an evaluation of a student's ability to "do" history rather than to take tests.

Oral Examinations

Historians often wish to examine a student's ability to "think on his feet." An oral examination tests the student's ability to think quickly and express his views extemporaneously rather than rely upon writing skills. Not only does the examiner witness the student composing an answer, but the situation encourages respondents as well as the examiners to clarify or modify questions and answers during the process. Oral examinations are standard practice in tutorials, admission exams, and thesis defenses, but they have certain drawbacks. The major one is injustice. Too often questions posed at oral examinations are spontaneous. They lack the benefit of planning and may not tie systematically to the purpose of the exam. When several faculty members form the panel of examiners, the questions may be posed to impress colleagues as much as to test the student. Sometimes questions are based on the examiner's current reading instead of what the student could be expected to know. The student who has a wide background of interests and can think well on his feet is obviously at a marked advantage. Oral exams, as much as any, should be planned in advance, with criteria stated explicitly and not be subject to whim to establish a passing grade.

Correcting Examinations

Historians could well make their method for correcting their exams public. Is the task intended to be criterion-referenced or norm-referenced.* If the former approach is used, a very tight statement of criterion should be formalized and announced. In either case, the teacher would do well to prepare an answer key before grading the test. If not, he may rightfully be accused of adjusting the target after the shot was fired.

The next requirement for fair correcting is that the papers should be kept anonymous during the grading. Names can be covered up, or better still, students can be assigned numbers to identify their test so that their name is not even on the test paper. Another sensible tactic in correcting is to grade one essay question on all papers and then go on to the next. This avoids biasing the reading of a student's answer on one question by the quality of the previous answer, particularly in essay grading. It is useful, after reading all questions to return and re-examine each paper in light of

---

*For a discussion of these terms, see below.

the others. In so doing, a teacher can judge relative merits more in relation to the entire set of responses.

One important concern of teachers in considering the type of test to be used is the time involved in scoring them. Usually, essay tests are easy to construct but difficult to score, while objective tests take considerable time to construct but are simple to score. There is something of the mass production logic in the use of the objective test: the teacher invests a considerable amount of time in the test production, and from that time on, scoring and generation of revised or alternative forms of the test become relatively easy, regardless of the numbers of students.

The question of returning the test for students to keep raises some controversy. Where considerable time has been spent developing test items, the teacher may feel that giving away the test items would invalidate their use for subsequent classes or sections. On the other hand, there is much that the student can learn if he is allowed to keep his exam and refer back to it. The question boils down to a choice between what the student gains from the feedback versus the effort the teacher has to devote to making new exam questions.

Some Principles from Testing Technology

Examinations with the purpose of demonstrating what a person knows or can do relative to a standard of performance are known as the mastery tests or Criterion-Referenced Tests. On the other hand an examination which compares a person's performance to the performance of other people in the group is known as discriminatory or Norm-Referenced Test. A professor can markedly improve his tests if he consciously designs a test to fit either of these purposes.

If the norm-referenced approach is selected, there should be a conscious expectancy designed into tests concerning item difficulty. For example, there should likely be some easy parts of the test but they should only account for 15% to 20% of the test. Thereafter, a scale of increasing difficulty could be established with modestly hard questions and hard questions providing the bulk of the examination and the very demanding portion limited to 10% to 15% of the credit. The result should be a wide range of test scores.

Another important step in developing quality examinations is to make certain the test actually measures what the professor intended the student to learn. To make sure this happens, the following sequence for constructing a test is recommended:

1. DESCRIBE IN WRITING WHAT IT IS YOU WANT THE STUDENT TO KNOW, FEEL, OR BE ABLE TO DO UPON SUCCESSFUL COMPLETION OF YOUR COURSE.
2. WRITE THOSE EVALUATIVE ACTIVITIES OR TEST ITEMS WHICH WILL INDEED MEASURE WHETHER YOUR INSTRUCTIONAL INTENTS WERE ACHIEVED.
3. DEVELOP YOUR LECTURE NOTES OR OTHER LEARNING ACTIVITIES IN SUCH A MANNER AS WILL ACCOMPLISH YOUR OBJECTIVES AND CAN BE EVALUATED AS DESIGNED.
4. PERFORM YOUR INSTRUCTION, GIVE YOUR LECTURE, OR MANAGE THE OTHER LEARNING ACTIVITIES.
5. USING THE PREVIOUSLY DESIGNED TESTS OR EVALUATION ACTIVITIES ASSESS WHETHER YOUR OBJECTIVES WERE MET OR NOT.
6. INTERPRET THE PERFORMANCE OF YOUR STUDENTS AS MEASURED BY YOUR TESTS.

Professors who develop their tests after they have performed instruction often commit a deception because the test may be based largely upon the memory of what the professor thinks he taught. Were teachers to write the tests before teaching the course, they would then feel contractually bound to see that the student had a fair chance to learn everything that would be tested. This need not cause instructors to teach to the test, but it can increase student productivity. There is simply no substitute in human performance for letting people know what is expected of them.

Historians will object to slavish adherence to this principle and rightly so. Sometimes the best teaching occurs as a diversion from the plan, and often the best learning results when students find a topic so interesting that they pursue it well beyond the course outline. Similarly, historians can argue that history is so broad that everyone in a class should not be held to one set of data for each period or theme. In some courses, students should be allowed to pursue their own interests within the subject parameters. If that is the case, then the mentor should say so clearly at the beginning of the course instead of keeping his views as an implicit attitude to be revealed only in his correction of tests. If pressed, the teacher would likely admit that there is some core body of data which serves as a basis for the branching out. That core can be pre-stated and a test written to measure its achievement. Then enrichment material can be evaluated separately. The principle thus still holds in general: make the goals or objectives explicit and develop test items to measure them.

A major issue in test construction is test validity: does the test item really measure what the test writer intended? Validity is by far the most important single element to be considered in designing an evaluation instrument. Professional test writers have developed a rule which guides them in constructing valid tests. They first require the purposes or objectives of the learning unit to be explicitly written. Then they look at the verb in each objective and write the test item to match the verb. If the objective is to have students "distinguish liberal and conservative interpretations," then the test question must cause the students to make a choice. If it has the students "analyze the liberal position," then the test question must cause the student to demonstrate analysis skills by discussing the various parts of the whole, a more exacting question. The test developer, seeking valid measures, must concern himself with the meaningfulness of the relationship of the test item and some independent criterion (his pre-stated objectives).

In his handbook on testing, Leslie Briggs provides a helpful categorization to understand test validity.[3] He suggests that some exams are intended to test reproductive learning. In that situation, the teacher tells the students the exact content of what is to be tested. Students then practice or drill to learn that content. In that case the test must call for that information and that skill specifically, in order to be valid. The other category is productive learning, instead of reproductive. Students are taught principles or concepts and asked to apply or analyze them. The practice students should engage in is applying the principles. The test would then call for that same nature of application, but the specific situations would not be known until the questions were given to the student.

A further requirement for a professionally acceptable test is that it be reliable. A reliable measuring instrument is one which achieves similar results upon repeated use, such as a test for color blindness. Reliability is a necessary condition for validity. That is, a testing device which is not reliable obviously cannot be used to measure anything. However, a reliable instrument may have low validity: for example, a color blindness

test may not be a good indicator of skill as an art critic, even though common sense might at first suggest it.

Some causes of unreliability are the examinee's fluctuating mood, motivation, or level of fatigue, unsystematic variations in the conditions under which the exam is taken, ambiguous questions, arbitrary scoring procedures, and luck. The test is generally more reliable if it is thorough rather than brief. Short tests might hit points some students know and omit equally important issues other students have studied. The reliability of each question can be studied by tabulating how each student did on the question and comparing that with the total scores on the test. When professional testing scholars construct a standardized examination, they go well beyond that initial comparison by re-testing students and by employing internal consistency formulas to judge the reliability of questions.

Another issue which faces the teacher who makes his own tests, as opposed to using commercially available examinations, is to determine whether his questions distinguish between those who know the material and those who don't. If the question "1" was answered well by 85% of the high-scoring students and 15% of the low achievers, it is said to discriminate well; if the percentages are more like 60% and 40%, the question does not discriminate well.

There are some common sense precautions a teacher can take to insure that tests are fair: make sure the answer key is correct; don't make the test so long that it cannot be answered in the time allotted; avoid questions in which the content is so rare that it is not worth remembering; concentrate on the most significant information.

Historians may feel they cannot match the skill of professional test designers or of outside examiners. While this may be true for broad measures of content, teacher-made tests fit the objectives of specific courses better than many standardized tests. Careful attention to the preparation of history examinations will not only produce reliable and valid measures of student learning, but also will serve as powerful feedback to both students and teachers in their pursuit of historical competence.[4]

NOTES

[1]Based on Grant E. Barton and Andrew S. Gibbons, "Writing Technically Correct Test Questions" (Provo, Utah, Brigham Young University Instructional Development Program, 1972).

[2]Douglas D. Alder, "The Historian's Rites of Passage," The History Teacher, VI (May, 1973), 404-7.

[3]Leslie Briggs, Handbook of Procedures for the Design of Instruction (Pittsburgh: American Institute for Research, 1970), 52-54.

[4]Three handy how-to-do-it booklets on test writing are Lowell A. Schoer, Test Construction: A Programmed Guide (Boston: Allyn and Bacon, 1970); Max D. Englehart, Improving Classroom Testing (Washington, D.C.: National Education Association, 1964), AERA Pamphlet Series #31; and Martin Katz, ed., Making the Classroom Test: A Guide for Teachers (Princeton, New Jersey: Educational Testing Service, 1961), Evaluation and Advisory Service Series #4.