

Reflective Essay

Assessment in History: The case for “decoding” the discipline

David Pace¹

I have been a professional historian for more than four decades, and I never once envied physicists — at least not until I encountered the “Force Concept Inventory” (Hestenes, Wells, & Swackhamer, 1992). This set of questions allows teachers of physics to determine whether students have internalized the basic Newtonian model taught in physics courses or whether they automatically fall back on “Aristotelian” notions of objects moving in space. It gives instructors in the field a simple instrument for determining the sophistication of students entering their classes and for evaluating the success of their own teaching strategies over a semester. This kind of disciplinary consensus about learning goals allows for the kind of far reaching and impressive assessment of learning that has allowed scholars of teaching and learning like Richard Hake to make convincing claims about the relative value of different teaching strategies (Hake, 1998; Bain, 2004).

A historian reading such work is apt to be immediately struck by the absence in his or her own field of this kind of agreement about what should be taught and what constitutes reasonable evidence that it has been learned. While physicists certainly argue about theoretical issues at the forefront of knowledge, they do not need to spend a great deal of effort justifying either the truth value of Newtonian mechanics or its relevance to the curriculum. In a discipline such as history, by contrast, undergraduates can enter contested spaces from their first day in the college classroom, and the subject matter that they are studying is as varied as the cultures that have left a trace on this planet. Both the ambiguity of sources and the co-existence of mutually contradictory interpretations would seem to dictate that history is and is apt to remain a “fuzzy” discipline.

This relative dearth of consensus is a result of the nature of the phenomena historians study, rather than any great deficiency in their profession, but it does lead to difficulties in creating a credible scholarship of teaching and learning. In a field in which reasoning is, of necessity, somewhat nebulous, it can be daunting to develop a clear consensus on what constitutes evidence that learning has occurred. Yet, assessment is at the core of the entire SoTL enterprise. It is difficult to imagine a robust scholarship of teaching and learning unless our work is cumulative and built on previous research and unless there is a means to systematically evaluate the validity of claims being made about student learning. In the now canonical formulation of Lee Shulman, the scholarship of teaching and learning must be “*public*, susceptible to *critical review and evaluation*, and accessible for *exchange and use* by other members of one’s scholarly community” (Shulman, 1998). Historians have made considerable progress in making SoTL public and accessible. An international society of historians working in the field has been created with its own website and newsletter, (<http://www.indiana.edu/~histsotl/>) and there is a growing programmatic literature exploring how the discipline might respond to the challenge of SoTL (Booth, 1996; Calder, Cutler, & Kelly, 2002; Pace, 2004, 2008; Brawley, Kelly, & Timmins, 2009).

¹ Professor of History, Indiana University Bloomington, dpace@indiana.edu

However, we have, as a discipline, been less successful at making our work “susceptible to *critical review and evaluation*,” in large part because of the lack of a consensus on how to assess learning. The great majority of studies involving history classrooms at the college level are based entirely on the instructor’s impression that “learning improved” or on a (generally undocumented) sense that student satisfaction increased. A journal such as *The History Teacher*, for example, is filled with wonderful ideas about how to improve instruction in the discipline, but, despite the efforts of the editors, one can read entire volumes of the periodical without encountering a single well substantiated conclusion (Calder, Cutler, & Kelly, 2002).

In earlier generations, when facts and dates occupied a much more central role in college history courses, the problem of demonstrating learning might not have been so daunting. But today, historians demand more complex cognitive processes from their students, and the subject of their inquiries is more often the perspectives, perceptions, and systems of power of earlier eras. If there is to be a credible scholarship of teaching and learning history, it will have to evaluate both what contemporary students actually have to do in the history classroom and how effectively our teaching strategies prepare them for those tasks.

Historians do, of course, evaluate their students’ performance, usually through essay exams, short answer identification questions, or multiple-choice exams. All of these provide information about student success and, thus, would seem to provide information about learning. But, at least as commonly interpreted, these instruments of assessment are too global and too impressionistic to provide the basis for a systematic scholarship of teaching and learning. Success or failure in an essay exam, for example, depends upon a host of separate skills, ranging from the ability to decipher a question or provide evidence in support of an interpretation to the capacity to manipulate English grammar or to manage time effectively — not to mention an understanding of a specific subject matter. The difference between an “A” and a “C” may be the result of a host of very different factors — emotional, cognitive, cultural, economic, or even aesthetic. Moreover, procedures for determining grades are generally shrouded in mystery and rest upon processes that may be perfectly legitimate for classroom teaching but do not provide a firm foundation for a systematic exploration of teaching and learning.

Multiple-choice questions are more focused, but, as they are commonly used, they tend to measure students’ mastery of facts or their memory of their instructor’s interpretation, rather than the ability to employ historical concepts and procedures. As has been noted repeatedly in the literature about teaching history, what really counts in the discipline is not the ability of students to repeat dates and events from memory, but rather their ability to think historically (Drake & McBride, 1997; Wineburg, 2001). Multiple choice questions can, of course, be crafted to measure higher level skills (Scott, 1983; Karras, 1984), but, even when this is done, they are typically created in reference to a set of content issues, rather than in response to a systematic analysis of the kinds of cognitive skills required in history courses.

Faced with this dearth of clear standards for evaluating student mastery of historical thinking, historians may assume that the only alternative is to abandon the province of their discipline and enter the alien world of classic social science methodologies. But, as so many social scientists themselves have noted, procedures such as the use of double-blind tests which measure the impact of a single variable on learning are rarely applicable to most teaching situations. There are too many variables loose in any real classroom to ever isolate one factor from all the others that have an impact on learning. Differences in the abilities or motivation of particular groups of students, in the investment of instructors in particular teaching methods, in the impact of the physical setting, or even the time of day of particular classes are extremely

difficult to control. In fields such as physics, where there are agreed upon questions, and very similar topics are being taught in hundreds of college classrooms, it may be possible to approach such an ideal. In history classes, where there is an enormous variety of subject matter and very little consensus on either central questions or how to evaluate answers to such questions, this kind of precision is unimaginable. Moreover, even if a few of us succeeded in this demanding task of methodological retooling, it is unlikely that scholarship couched in this language would have much impact on historians accustomed to approaching problems from a very different angle.

Thus, historians may seem to face the question of assessment in a field that seems resistant to systematic evaluation with tools that are both foreign to their professional training and of questionable applicability to the task at hand. It is not surprising that so many of us have chosen to ignore the entire issue of assessment –or rather have limited it to our own general impressions of success or failure. Some would argue that the problem lies in the nature of history teaching itself, that the ability to make reasoned and systematic judgments that we take for granted in the realm of our research can never find a place in that of our teaching. But such a position is thoroughly a-historical and fails to recognize that the standards employed by professional historians in judging traditional disciplinary research are the product of generations of focused cultural labor. The criteria of judgment, rules for the admissibility of evidence, and social foundations of credibility that allow the scholar to think systematically about the past are not intrinsic to the subject matter. Like the procedures of our legal systems, they arose through the need to establish agreed upon bases for decision making. We are currently facing a similar need in the scholarship of teaching and learning history, and we must begin the demanding task of establishing methods of systematic argumentation about student learning in the field.

If the scholarship of teaching and learning is to succeed in history, it will be necessary to move beyond this impasse by finding new criteria for defining the basic operations needed for success in history classrooms and for evaluating student mastery of these skills. I would suggest that we consider the following principles, when attempting to assess learning in history courses:

1. Assessment must be preceded by a clear definition of what is to be assessed. We need to have some idea of what we want to measure before we can measure it.
2. It is best to begin by focusing on the specific operations required in a history course, rather than on generalized forms of critical thinking.
3. In deciding what to assess, it is important to concentrate on measuring things that have a great impact on student success in courses in the discipline. There is always a temptation to measure what is easy to assess (e.g. students' knowledge of facts and dates) rather than the more complex forms of historical reasoning that are usually more essential to success in contemporary history courses.
4. We should concentrate our energies on aspects of history teaching that are problematic. It is less important to develop means of assessing student progress in areas where learning generally occurs spontaneously than in those in which many students are unable to master basic ways of thinking.
5. Assessment will be most effective if it is narrowly focused on particular skills or tightly related clusters of well-defined skills. As has been noted above, traditional history exams do provide a basis for judging students' global mastery of the entire set of skills required for success in history courses, but they generally provide little specific knowledge about which operations have been mastered by students.

6. Assessment in a field such as history rests on judgments that are relative, not absolute. The nature of the phenomena being observed is so complex that positivistic criteria for establishing certainty can only hinder the work. The best that we can do is to make it appear reasonable to expect that certain strategies have a positive impact on learning or that certain ways of approaching historical questions are common in particular groups of undergraduates.
7. Because of the complexity and the ambiguity of the phenomena being studied it will be best to explore a variety of assessment strategies, both quantitative and qualitative.
8. Assessment should be viewed, not simply as a means of evaluating student learning, but, whenever possible, it should serve to further that learning. Assessment should generally be a part of the learning process, not something that is added on as an after thought.

There are almost certainly multiple paths to achieving meaningful assessment of learning in history courses. But each of these will probably have to meet most of the requirements listed above. Personally, I have found it most effective to pursue these goals within the framework of the Decoding the Disciplines process. This approach, developed in the Indiana University Freshman Learning Project, suggests that faculty seeking to understand the learning processes in their courses can productively begin by defining “bottlenecks,” i.e. places where large numbers of students have difficulty mastering some concept or action that is essential to success. Then the investigator can begin the intellectually demanding process of defining the steps or operations students would need to overcome the bottleneck. Generally, this requires a painstaking deconstruction of the processes professionals in the field employ automatically, and, like the exploration of other largely unconscious phenomena, it may require the assistance of others who are less involved with the material. Once the task at hand has been broken down into its component parts, each of these can be modeled for students, they can be given opportunities for practice and feedback, and the mastery of each operation can be assessed individually (Pace & Middendorf, 2004).

The Indiana University History Learning Project has demonstrated that this process can be effective at promoting and assessing learning in history classrooms (Diaz, Middendorf, Pace, & Shopkow, 2007; Diaz, Middendorf, Pace, & Shopkow, 2008; Pace, 2008) In the pages that follow I will trace the application of this process to two interrelated bottlenecks frequently encountered in history courses: 1) students’ inability to find appropriate evidence to support an interpretation; and 2) their difficulty in making the connections between the evidence and the interpretation clear to their readers. These skills are absolutely essential to any history course that goes beyond simple memorization of facts, and yet they are not part of the skill set of many current college students.

In the description below I will focus primarily on a small seminar I taught on “Paris and the Birth of Modern Culture, 1850-1900” in the summer of 2008. This course was offered as part of the Indiana University Intensive Freshman Seminar Program, which provides all first-year students with the opportunity to take a three-week course before the fall semester begins. My thirteen students were highly motivated and very focused on the course, and they began with a wide range of historical skills. Thus, it provided me with a good opportunity to test the Decoding the Disciplines approach and to see if it would yield clear evidence of learning. But, since the nature of this course made it somewhat atypical, I will supplement this discussion with data from a larger course taught in the regular semester in the spring of 2009.

The difficulty many students have in employing evidence to support a historical interpretation was visible from the introductory essay that I asked students to write before they

arrived on campus. They were provided with excerpts from guidebooks to Paris written in the 1850s and 1860s and asked to write a two-page paper discussing how the city was presented to foreign visitors. Some of the students were able to advance a coherent thesis about the representation of Paris and to provide relevant evidence to support this argument. Others seemed to throw facts randomly at the question, hoping that some of them would stick. A careful reading of the latter provided a clearer understanding of the nature of this bottleneck.

One paper, for example, began with what appeared to be a promising thesis: "Paris is a dare. A dare to all other society's cultures, and countries to prove itself..." Here, I thought for a moment, was an adventurous and original thinker, who was focusing on the ways in which Paris was presented as a challenge to other cultures. But the paragraphs that followed rambled through unrelated details borrowed, seemingly without direction, from the readings. A typical bit of "evidence" informed the reader that "As previously stated Paris has an image to uphold, which explains the destruction of numerous buildings such as the building in which Duc de Berri was stabbed on the rue Richelieu and the portico where the Emperor and Empress were assassinated in 1858." This statement was factually incorrect, since the attempt on the ruling family was not successful. But, more importantly, the destruction of buildings was not a particularly good example of the notion of cultural challenge and its link to the larger issue was never spelled out clearly.

There was ample evidence from my own classes and from interviews with other historians that we videotaped as part of the History Learning Project that large numbers of the students taking courses in our department were prevented from fully succeeding by this kind of inability to select and to justify evidence. Following the decoding process, I now needed to describe as precisely as possible the things that I, as a professional historian, would do to get past this potential obstacle. To make this concrete, I focused on what I, myself, would do automatically when faced with a question from one of the web-based assignments for the course. Here students were asked to imagine that they were writing an essay defending the thesis that the activities of the Baron Haussmann, Prefect of Paris in the 1850s and 1860s, had a positive effect on the development of Paris. They did not actually write the paper, but they were asked to select a passage from the readings that could be use to support this interpretation and to explain briefly what about the passage made it useful in defending the thesis.

I then sought to define the kind of operations that would be necessary to successfully complete this task. In the following list I have defined some of these elements and indicated (*in italics*) how each abstract principle would be realized in the context of this particular assignment. Thus, to succeed at this task, students must:

1. Understand that there is more than one plausible explanation of a historical phenomenon. (*Understand that it might be reasonable to say either that Haussmann did or did not improve Paris.*)
2. Understand that, for an explanation to have plausibility, evidence must be presented that makes it seem more likely than competing explanations. (*Understand that it may be possible to discriminate between better and worse explanations of Haussmann's impact on the basis of evidence.*)
3. Define the basic terms used in the thesis. (*Define criteria for "positive effect" and "improved."*)
4. Uncover the propositions, implicit in the interpretation, that must be defended. (*Recognize that to support the thesis one would have to demonstrate that Haussmann's actions had an effect on Paris and that the effect of these actions was positive.*)

5. Identify what kinds of evidence would support or undercut each specific proposition. (*Think: If Haussmann did improve the city, what evidence of this improvement might still be available.*)
6. Find evidence that would have meet the criteria in #4, above. (*Go back through the criteria established in #4 to see if any of these signs are present in the information that they have about the period and what followed.*)
7. Evaluate the quality of the sources of the available evidence. (*Evaluate the validity of the sources containing each relevant bit of information about Haussmann and Paris to determine which are the most dependable*)
8. Demonstrate to a reader how the existence of this evidence would make the argument in question more likely to be true. (*Demonstrate in writing steps #4, #5, and #6 in a manner what will be clear and convincing to an intelligent reader.*)

Breaking up the process of using evidence historically allowed me, first, to model these steps individually for my students and then to assess their ability to perform particular ones. This deconstruction process also made clearer the strategic choices I faced both in teaching and in assessing learning. Eight processes were too many to teach or assess in a single course. Therefore, I had to made choices about which were most important in the context of this course.

The first two steps, which involve the kinds of issues dealt with in William Perry's (1970) classic study of students' intellectual and moral reasoning, seemed to not be problematic for this group of students. They seemed to understand that historical knowledge is based on weighing of evidence, rather than the discovery of some absolute truth. Six basic processes were still a little too much to teach and assess in a three-week period. Therefore, I decided to deemphasize steps 3 (defining terms) and 7 (evaluating the quality of sources). These are very important, but, given the time constraints, I had to hope that they would be reinforced in the later courses that the students took.

Therefore, I devoted class time to modeling steps 4, 5, 6, and 8, and I gave my students opportunities to practice and receive feedback separately on each of them through in-class team exercises and daily on-line assignments, inspired by Gregor Novak's (1999) Just-in-Time Teaching warm-ups. (These assignments and exercises, along with some of the strategies I used in modeling these operations, may be found at <http://www.iub.edu/~hlp/supporting materials/Assessment in History>.)

Yet, the question remained – had my students really mastered these skills? Had the Decoding the Disciplines process given my students the tools that they would need to overcome similar bottlenecks in future courses. There was, as I have argued above, no way that I could absolutely prove this, any more than historians doing research on a historical problem can be sure that they are establishing the validity of a particular explanation without potential controversy. But I was convinced that I could amass evidence that would strongly suggest whether the students had mastered these operations.

I began by analyzing student responses to one of the on-line assignments near the end of the course. As part of this task they had to 1) generate a thesis about patterns of gender in late 19th century Paris, 2) identify three propositions that had to be true for the thesis to be valid, 3) find a bit of evidence that would support each proposition, and 4) finally explain what about the evidence should convince a reader that each proposition was credible. This promised information concerning students' mastery of operations 4 (define the propositions that would need to be supported to defend a thesis), 5 (identify the kinds of evidence relevant to each proposition), 6

(find examples of such evidence), and 8 (demonstrate the relevance of the evidence to the reader).

I created a rubric in which each relevant part of the assignment was associated with one or two of these operations. I then reread their work, assigning a point or some fraction thereof to each question and entered the results in a spreadsheet. Five of the thirteen students got a perfect score, receiving full credit for all three examples of propositions, supporting evidence, and justifications for the choices of evidence. As a group, they averaged 93 out of a possible 100 on the entire assignment.

These results were very promising, but they did not, in themselves, demonstrate that my students' ability to select and justify evidence had increased over the semester. However, I had also administered pre- and post-tests at the beginning and end of the course. The same questions were asked on both occasions, and I chose subject matter that had not been covered in the course to assure that I was measuring changes in students' ability to process historical material, not in their content knowledge. After the course was over, I had the tests coded and randomized, and, without knowing which of the tests had been taken at the beginning of the semester, I gave points to each answer, based on a rubric I had created. The pre- and post-scores were then separated and the differences compared on a spreadsheet.

In the section of the assessment most relevant to the issue of using evidence, students were presented with a passage from a standard textbook describing an early 19th century American entrepreneur and a brief interpretation of the factors leading to the Industrial Revolution in the United States. Students were asked to find evidence in the passage that would be useful in supporting or contradicting the interpretation and to explain what about the evidence made it useful for this purpose. I identified two types of evidence in the passage which could be used to answer the essay effectively (material dealing with new technologies and with entrepreneurship) and counted the number of times students were able to identify each.

I found that at the end of the three-week course students were 14% more apt to recognize new technologies as potential evidence and were 28% more apt to mention entrepreneurship. My evaluations of the quality of the justifications for their choice of evidence were 34% higher at the end of the course. (All thirteen students in the course took both the pre- and the post-test.)

The ultimate test of any pedagogical strategy must involve the integration of specific skills in a finished piece of work. Therefore, as a final assessment I compared the use of evidence in student papers at the beginning and the end of the course. In this essay students were asked to discuss factors that contributed to Parisian culture becoming more free and experimental in the second half of the 19th century.

Here is a paragraph from the final paper of the student whose pre-class paper was quoted above.

The social classes were transferring, the economy was revolutionized by the industrial revolution, and so too, the art world made a change. Prior to this point in art history, the Academy had favored the classics of paintings, sculptures, and the like, that portrayed Romanesque figures and a higher nobility of unattainable perfection, but were now forced to reckon with a new ideal in the art world. In this time period that academic art was replaced by impressionist art and romantic ideals portrayed in art. "There were, of course, conservative critics who mourned the decline of the grand tradition; but the greater danger was the invasion of the whole art world by the crude and tasteless standards of the hundreds of new middle-class purchasers". Romanticism and impressionism came about by the

rejection of the enlightenment rationality that was created decades previously, and resulted in conflicting ideas of nature and the exploration of human experience. It was through romanticism that the bohemian counter culture was generated. It had once been acceptable to be educated at the Ecole des Beaux Art, but now replaced by the advent of sharing their preferred form of art at cafes and local restaurants. Along with the romantic and impressionist art movement came a new system of patrons to buy this new depiction of art.

There are still many problems with this work – unfortunate word choices, incomplete understanding of the material, places where more appropriate evidence might have been selected, inadequate explanation of the significance of a quotation. But, unlike the student's initial work, this reads like a historical argument. Here the argument that the art world was changing was accompanied with concrete evidence supporting this claim. She mentions the decline of the influence of the French Academy and of the Ecole des Beaux Arts, the appearance of Impressionist painting, the role of cafes in promoting independent centers of culture, and changes in the economic and social systems supporting the arts. She definitely needs more work on learning to make clear the relevance of her evidence to the central argument, but the discerning reader would, nonetheless, be much more able to grasp the reasons she provided these details, than would have been the case with her initial paper. This represents significant improvement from her work just three weeks earlier.

Finally, I had an indirect means of determining whether students had internalized a sense of the importance of supporting arguments with evidence. On the last day of class each student wrote a short "letter," designed to give a hypothetical younger sibling or friend advice on how to succeed in my course. Instructions for the exercise provided earlier on the course web site listed "using evidence to support a position" as one of ten possible issues for consideration in the essay, but students had a limited amount of time and were generally only able to deal with a small number of the possible topics.

Nonetheless seven of the twelve "letters" made explicit reference to the importance of using evidence to support an argument in history courses. One offered this advice to a hypothetical friend or sibling: "Pay attention to quotes that really strike you in the reading. They will come in handy at some point in helping to back up an argument or support a point." Another clearly understood the use of evidence as part of a disciplinary procedure: "First, let me say what a history class in college is not. It is not a math class. There are no clear formulas that produce exacting answers. There is no one correct interpretation, or even two or three necessarily. Interpreting an event in history means compiling evidence and making some decisions." Others picked up on my use of the metaphor that historians must make and support an argument much as a trial lawyer does in court. One student advised her brother to "act like a lawyer: take a side, give evidence, and explain why that evidence supports your stance." And another quite clearly summed up the challenge facing students in college history course:

Think to yourself "lawyer, lawyer, lawyer!" when deciding your argument and the evidence that reinforces both the argument and the thesis. Evidence is very crucial in a history paper because it is an account of the past that no longer exists in the present reality. Make sure you use evidence that *actually* supports your argument and not end up with a paper where the defense of the conclusion is different from that of the introduction!

Finally, the student, whom we have been following throughout this essay, indicated that she knew what she needed to do in the course, even if she had not yet completely mastered the steps that she needed to get there:

I have found through this class that history is a constant barrage of questions and challenging opinions, where you must research and collect evidence and test multiple theories in order to arrive at either a truth or a falsity. You must question what is valued by the author or the person making the argument. You must ask further questions to be able to answer the original observation. Finally, you must provide some sort of evidence to explain your argument and reasoning for answering the question.

One course — particularly one with only thirteen students — cannot by itself provide sufficient evidence to demonstrate the effectiveness of a particular teaching strategy. Therefore, I repeated the teaching strategies in a larger, upper division course and sought evidence on their impact on the students' ability to choose and justify evidence. Students generally began this course with a much higher mastery of basic skills, than those in the class discussed above. But, when I compared the work of 33 students on assignments early and later in the semester, using the approach described above, I detected a 7% increase in their ability to select relevant evidence (from an average of 86% to 92%) and a 16% increase (from an average of 72% to 92%) in their ability to explain the relevance of their evidence.

There was the possibility that these scores were affected by the differences in the subject matter dealt with in the two parts of the course. Therefore, I again gave pre- and posttests at the beginning and the end of the semester, had them coded, and used a rubric to evaluate without knowing which were done at the beginning or end of the course. The students were given a quotation from a 19th century British author and two interpretations of developments within British society in that period. They were asked which interpretation was most clearly supported by the evidence, to specify what would have to be demonstrated to "prove" the interpretation, and to indicate how the interpretation might be used in this demonstration. [The pre- and post-test and the statistical results may be viewed at <http://www.iub.edu/~hlp/supporting materials/Assessment in History>.]

I again evaluated these tests without knowing which came from the beginning or the end of the course, looking for their choice of interpretation, for certain elements from the passage that could be used to support it, and for the quality of their explanation of the relevance of the evidence. The results of the assessment were quite positive. I limited my analysis to the 43 students (of 67 in the class) who took both the pre- and post-test. The portion of students choosing the more appropriate interpretation, increased from 30 (71%) to 35 (80%), an improvement of 9%. In evaluating their success at understanding what was called for in the interpretation, I decided to limit the analysis to the 25 students who had chosen interpretation A on both tests.² On the second iteration of the test these students collectively did significantly better in all the categories considered. The greatest increase was in the recognition of the importance of issues of regulation and social control, where there was an improvement of 58% across the semester. On the second iteration of the test 30% more acknowledged the importance of the idea of transformation in the interpretation, 18% more mentioned the theme of competition and individuals, and 21% more students made explicit reference to the time period covered by

2 The 25 students who had taken both tests and who had chosen interpretation "A" accounted for 50 of the 86 tests available. I limited my analysis to these tests because of the difficulties in establishing a clear comparison of students' treatment of different interpretations.

the interpretation. There was also a dramatic 66% decrease in the number of students who framed the entire interpretation in ahistorical and moralistic terms, but the small numbers (3 and 1, respectively) makes this less than meaningful.

The differences in the pre- and post-tests reinforced the suggestions in the comparison of the early and late assignments that across the semester students became more able to define the claims that had to be defended when dealing with a particular interpretation, to find relevant evidence, and to explain the significance of this evidence for the interpretation. The presence of positive results in two very different classes supports the notion that the teaching strategies used in these courses made a contribution to student mastery of basic skills that are essential for success in history courses and that can be invaluable to students in future life.

However, these exercises in assessment raise some important questions about just how much improvement represents real progress. It is obvious that a student's entire way of operating cannot be transformed within a single semester, but just how much change is required for a teaching strategy to be judged a success? It would seem likely that any increase over 20% is a clear positive indicator of success. But what about 10%? Or 5%? Until systematic assessment of learning in history courses becomes more common, it will be difficult to know just what constitutes success.

Moreover, there remains the question of whether the semester is the optimal unit for measuring increases in learning. It is quite possible that deep learning requires a longer time to sink in and that it may only be fully visible a year or even multiple years after the process has been initiated in a particular class. The History Learning Project has begun to explore this issue by taking "snapshots" of student abilities through short assessment exercises in multiple classes each semester. This will hopefully allow us to trace typical skills trajectories across the undergraduate curriculum and, perhaps, to trace the development of particular students, who take multiple history courses. However, the process of capturing an image of student abilities in a brief exercise is daunting, as is the effort to link success or failure to any particular teaching strategy.

It is important, however, to stress that assessment is not just about measuring change. It can also provide useful information about the level of learning with which students begin a course, the type of difficulty students have in mastering certain disciplinary skills, and the manner in which students go about solving problems. The assessments described above, for example, told me that I needed to focus even more on the issue of discovering evidence in the lower level course, but that in the upper level course I could focus more of my effort on helping students unpack the claims that had to be supported in an interpretation.

Other useful information can emerge from this kind of systematic analysis of student achievement. I learned, for example, that in the upper level course students' grade on the second weekly web assignment correlated very closely with the average final grade in the course (82.54% for the assignment versus 83.18% for the course as a whole). This suggests that it might be important to pay particular attention to supporting the learning of students who had difficulty with this assignment. And I noticed that in the later assignment the students in the larger course picked evidence from a wide range of possible primary sources from the web site, rather than hurriedly grabbing something from the source at the top of the page. This indirect evidence suggests that the level of their motivation remained relatively high, even in the harried thirteenth week of the semester.

It is, however, important to stress that the evaluation of learning in history can never be restricted to a single, externally imposed instrument of assessment. It is clear from both the nature of historical practice and the intellectual politics of the discipline that it would be very unfortunate to attempt to force a single form of assessment of student learning on the discipline from the top down. Different historians appropriately concentrate on different aspects of learning, and no single instrument can possibly capture this useful diversity. Moreover, in attempting to impose such a universal standard, there will always be the temptation to ignore the complexity of what really needs to happen in the history classroom and to focus, instead, on what is easiest to measure – students' memorization of factual knowledge. If historians do not develop their own criteria for evaluating student learning, such crude and inappropriate approaches to measuring student learning may, sadly, be imposed by forces outside academia. But it is difficult to imagine that this would have anything but a negative impact on student learning.

It is possible, however, to imagine a different path toward a loose consensus about how student learning might be evaluated. Individual historians might seek to define specific aspects of historical thinking and to develop means for systematically determining whether student mastery of these increases across a semester or across a student's college career. The results of such assessments could provide useful clues concerning what aspects of individual teaching seem to be yielding positive results and what strategies need to be reevaluated.

On a broader scale these approaches might help a department gain a better understanding of the skills that students bring into history classrooms at various levels in its curriculum. As the work of the History Learning Project suggests, (Diaz, Middendorf, Pace, & Shopkow, 2008) information derived from this work can help departments decide what skills should be introduced at different levels of the curriculum and what disciplinary ways of thinking can be assumed to be present at the beginning of courses. It is, thus, possible to imagine a future in which a faculty member could begin a semester with a much clearer notion of what it is reasonable to expect of students in a particular course and what basic disciplinary skills should be focused on to allow the maximum number of students to master the course material.

If they were made public through publications or websites, such local experiments in creative assessment might provide the basis for a broader discussion among historians about how we evaluate what students are or are not mastering in college history classrooms. Individual instructors could build on the work of others or explore aspects of historical reasoning that had been ignored in previous studies. It is even possible to imagine banks of questions made available to historians interested in determining the level of historical reasoning of the students entering his or her class or in evaluating the amount of change that occurred in these skills across a semester. This would truly create a scholarship of teaching and learning history that is, to return to Shulman's formulation, "accessible for *exchange and use* by other members of one's scholarly community."

References

Bain, K. (2004). *What the Best College Teachers Do*. Cambridge, MA: Harvard University Press.

Booth, A. (1996). Changing assessment to improve learning, in A. Booth and P. Hyland, (Eds.), *History in Higher Education: New Directions in Teaching and Learning*. Oxford: Blackwell Publishers.

Booth, A. (2004). Rethinking the scholarly: Developing the scholarship of teaching in History. *Arts and Humanities in Higher Education*, 3(3), 247-266.

Brawley, S., Kelly, T., & Timmins, G. (2009). SoTL and national differences: Musing from historians from three countries. *Arts and Humanities in Higher Education*, 8(1), 8-25.

Calder, L., Cutler, W. & Kelly, T. (2002). History lessons: Historians and the scholarship of teaching and learning in M.T. Huber and S.P. Morreale (Eds.), *Disciplinary Styles in the Scholarship of Teaching and Learning: Exploring Common Ground*. Washington, D.C.: American Association for Higher Education and The Carnegie Foundation for the Advancement of Teaching.

Diaz, A, Middendorf, J, Pace, D, & Shopkow, L, (2008). The History Learning Project: A department “decodes” its students. *Journal of American History*, 94 (4), 1211-1224.

Diaz, A, Middendorf, J, Pace, D, & Shopkow, L. (2007). Making thinking explicit: A history department decodes its discipline. *National Teaching and Learning Forum*, 16 (2), 1-4 .

Drake, F., & McBride, L. (1997). Frederick D. Drake; Lawrence W. McBride reinventing the teaching of history through alternative assessment. *The History Teacher*, 30, 145-173.

Hake, R. (1998). Interactive-engagement vs. traditional methods. *American Journal of Physics*, 66 (1), 64-74.

Hestenes, D., Wells, M., & Swackhamer, G. (1992). Force Concept Inventory. *The Physics Teacher*, 30(3), 141-158.

Karras, R. (1984). A multidimensional multiple-choice testing system. *Perspectives*, Reprinted in R. Blackey, (Ed.), *History Anew: Innovations in the Teaching of History Today*. Long Beach, California: California State University Press.

Novak, G. (1999) Just-in-time teaching: Blending active learning with web technology. Upper Saddle River, NJ: Prentice Hall.

Pace, D. (2008). Opening history’s ‘black boxes’: Decoding the disciplinary unconscious of historians. In C. Kreber, (Ed.), *Teaching and Learning Within and Beyond Disciplinary Boundaries*. London: Routledge.

Pace, D. (2004). The amateur in the operating room: History and the scholarship of teaching and learning. *American Historical Review*, 109(4), 1171-1192.

Pace, D., & Middendorf, J. (Eds.) (2004). *Decoding the Disciplines: Helping Students Learn Disciplinary Ways of Thinking*. Hoboken, NJ: John Wiley.

Pace, D.

Perry, W.G., Jr. (1970). *Form of Intellectual and Ethical Development in the College Years: A Scheme*. New York: Holt Rinehart, and Winston.

Scott, A. (1983). Life is a multiple choice question. *Perspectives*. Reprinted in R. Blackey, (Ed.), *History Anew: Innovations in the Teaching of History Today*. Long Beach, California: California State University Press.

Shulman, L. (1998). Course anatomy: The dissection and analysis of knowledge through teaching. In P.Hutchings, (Ed.), *The Course Portfolio: How Faculty Can Examine Their Teaching to Advance Practice and Improve Student Learning*. Washington, D.C.: American Association for Higher Education.

Wineburg, S. (2001). *Historical Thinking and Other Unnatural Acts: Charting the Future of Teaching about the Past*. Philadelphia: Temple University Press.