CHAPTER 21

Assessment and Evaluation in Social Studies Classrooms

A Question of Balance

John Myers

What Is the State of Assessment and Evaluation in Social Studies Today?

Since public education began in Canada 150 years ago, assessment has identified and ranked students using tests, essays, and quizzes. This limited role for assessment has broadened over the past twenty years for a number of reasons: First, the goals of education in general, and social studies in particular, have greatly expanded as we attempt to help our children acquire the knowledge, skills, attitudes, and behaviours required for living in the twenty-first century. As social studies educator Joseph Kirman argues, social studies is not about the accumulation of selected facts for later recall but should aim to:

> Produce a responsible person able to cope with change, capable of making reasonable decisions, who is an intelligent consumer and controller of science and technology, able to live with and appreciate human diversity, and support and defend human dignity. Such a person should be able to settle differences honorably, avoid the use of violence, be cognizant of, and active in, the stewardship of our planet, and have the skills necessary to maintain a functional economic system and democratic government.[1]

This expanded view of the purposes of social studies education is reflected across the curriculum and requires a much more sophisticated approach to assessment and evaluation.[2] Assessment reform has also been driven by increased knowledge about learning. After decades of viewing learning as a relatively passive accumulation of facts and discrete skills, we now recognize learning as a more active process in which learners construct knowledge and make connections among facts and concepts. Innovations in recent decades, such as curriculum integration, authentic assessment, and cooperative learning, reflect these current theories.[3]

A third factor driving assessment and evaluation reform comes from the widespread view that traditional ways of reporting student achievement—the evaluation side of things—are not telling us what we really need to know. It is not enough to want students to learn and to express our best intentions by setting course, unit, and lesson objectives. These objectives need to be achieved. We need results: clear evidence students have met the outcomes set for them. Many provincial ministries/departments of education are responding to this concern by trying to determine standards of achievement in various subject areas, including social studies, to help make grading more accurate.

Concurrent with these trends to broaden the scope of what and how to assess student progress has been a rise in large-scale, standardized assessments of various types including provincial, national, and international achievement tests. The term "standardized" refers to the fact that the results of such tests, if

administered according to their instructions, can be interpreted the same way in different circumstances. The standardization occurs as a result of field-testing with thousands of students. They are administered system-wide or province-wide. The instructions to students, the test conditions, the timing, and scoring are the same in all places. They are generally norm-referenced and provide comparative data on students. They usually test for what the test authors determine to be basic skills and knowledge, and are considered to be objective; i.e., free of teacher bias in scoring. Commercial standardized tests are supposed to be content-free; that is, they are not designed to fit any particular district or provincial curriculum. In Canada, students are usually given standardized tests in language and math. Alberta is the only province that currently gives standardized tests in social studies at a number of grade levels. Provincial examinations in all subjects including social studies contain many if not most elements of standardized tests.[4]

Whatever benefits such tests have in determining student acquisition of basic facts and simple skills, they do not set a standard as to what students should be able to do. They do not define clear expectations, other than to get the highest score possible. They seldom assess complex thinking and behaviours. Even the advocates of standardized tests express concerns for their overuse. Most importantly for the purposes of this chapter, standardized tests are incapable of providing teachers with the day-to-day, moment-to-moment information they need for making instructional decisions. What follows will therefore focus on classroom-based forms of assessment and evaluation.

Even at the school and classroom levels there are competing trends in assessment and evaluation. This chapter tries to make sense of these as they influence the assessment and evaluation of student achievement in social studies. The key focus is to promote "assessment literacy"; i.e., the capacity to better match assessment methods and the information desired about student achievement, make sense of the student achievement data we collect, communicate assessment results effectively, use assessment to maximize student motivation and learning, and identify current trends and issues in the field for further investigation and research so that we can be more proactive and better able to hold our own in debates about assessment, evaluation, standards, and accountability. As assessment expert Ruth Childs put it in the title of a recent address to a group of student teachers, "Critical Consumers Needed."[5]

What Is the Vocabulary of Assessment and Evaluation?

One of the confusing factors in the field of assessment is that professionals often use the same terms to mean different things. It might be helpful then to define some important terms and see how they relate to each other.[6]

The terms "assessment" and "evaluation" are often used interchangeably but they do not always mean the same thing. For many, assessment is the process of collecting data on student performance that is then used to evaluate (make judgements about) the attainment of certain expectations, objectives, or outcomes. Because we speak of judgements, evaluation is never neutral. It can only be objective in that assessments provide us with a quantity and quality of information sufficient to make judgements that are fair to the strengths and needs of all learners; precise in that students and others gain clear insights into how they are learning and what they can do to improve performance; and true in that our verdicts are valid, consistent, and accurate. It is important to remember that all assessments and evaluations are limited in that they do not tell us everything there is to know about the learning of particular individuals or groups.

Assessment and evaluation are continuous activities in the classroom and they can be both informal and formal. *Informal assessment* occurs when a teacher collects information to use for the purposes of shaping ongoing instruction. At the end of a segment of a lesson, for example, the teacher might orally ask a random sample of students around the room to respond to questions about what was just taught. The responses to these questions will give

the teacher a sense of how well the class has learned the information, concept, or skill in question and whether they are ready to move on. Teachers might also carry out informal assessment of individual students by observing the way they are completing tasks assigned in class. This can tell them which students might need corrective feedback or more instruction in a particular area.

While informal assessment is a continuous and sometimes almost unconscious process, *formal assessment* is normally more overt and systematically planned. We are all familiar with typical manifestations of formal assessment: quizzes, tests, essays, and projects. But formal assessment may take a wide range of forms, including teacher observation of student performance. For example, a typical outcome for social studies is that students become more open- minded. One could imagine assessing progress towards that goal by having students write position papers on issues from various points of view, but observing students interacting in class with people expressing views different from theirs might also accomplish this. What makes this kind of observation different from an informal assessment is the structured nature of the data collection, perhaps using a checklist or taking notes, and sharing the information gathered with the students involved and perhaps others to provide *feedback*.

A key to effective assessment and evaluation is clarity about purpose. Broadly speaking in terms of purpose, there are two types of assessment: *formative* and *summative*. The purpose of formative assessment is to provide teacher and student with feedback that can direct future teaching and learning. For example, following some instruction on the skill of comparing and contrasting, the teacher might assign students the task of writing a short essay comparing and contrasting Napoleon's and Hitler's invasions of Russia. When the teacher examines the essays she finds that the majority of students tend to describe both invasions but do not actually develop comparisons and contrasts. This information tells her that there are areas of the skill that need further work. She can then develop a lesson showing students both the strengths and weaknesses of their work in terms of the skill, and provide further direction in developing the skill. Most of the assessment and evaluation we do will be formative, and the feedback loop that results (see Figure 1) is very important in the teaching and learning process.

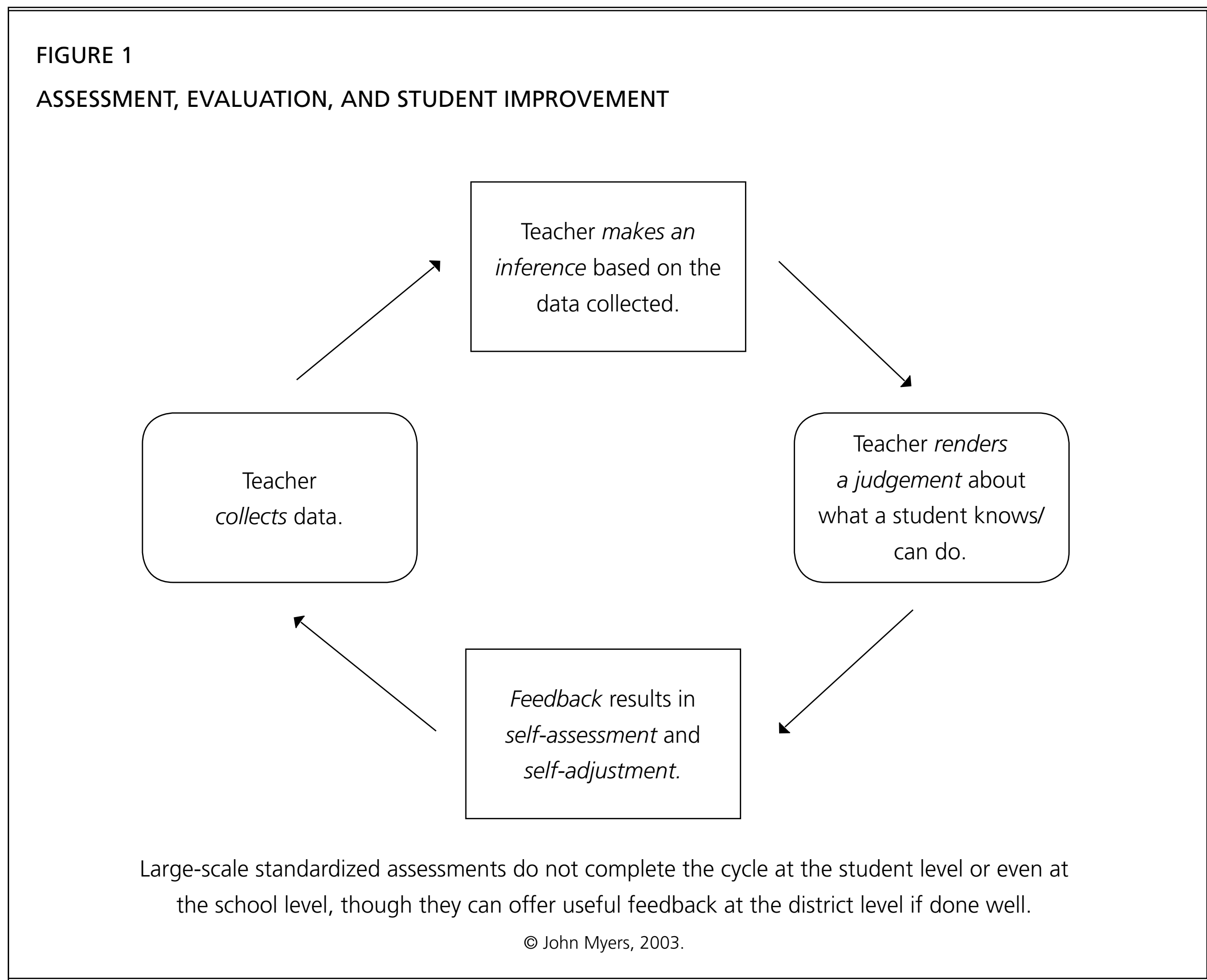
Central to formative assessment is the provision of quality feedback. Feedback consists of information that tells us how we are doing and what we need to do next, in the light of our intentions and goals. Feedback is not the same as praise and blame, rather it is precise information about where someone is in relation to the goals they are trying to attain and what they might do in the future to make progress towards those goals.

A research synthesis by Marzano, Pickering, and Pollock noted the following principles of quality feedback:[7] First, feedback should be timely (the longer we wait the less effect it has on achievement). Feedback should be specific (criterion-referenced) and "corrective" in order to show what went well, what needs improvement, and how to improve (all three components needed for maximum achievement). Feedback can be verbal and written and can come from teachers, peers, or the student him/herself. It should ask students to interpret data and self-assess in the light of their goals and intentions, rather than ask them to react to our interpretation. Finally, feedback should allow students to make decisions as to the nature of the improvements and adjustments that need to be made. The great inventor Thomas Edison had his own way of describing the importance of feedback: "I've never made a mistake. I've only learned from experience."

Summative assessment provides an accounting of student progress at a particular point in time. It is normally a measurement that describes where the student stands in regard to some sort of standard such as curriculum outcomes. An end-of-unit test, for example, is designed to let students know how well they have accomplished the goals of the unit in terms of knowledge and skill acquisition. The most familiar summative assessment is the report card that communicates to students and their parents the degree to which students are meeting expectations with regard to the curriculum. Of course summative assessments can be used in a formative manner—a report card might help students focus on particular

FIGURE 1

ASSESSMENT, EVALUATION, AND STUDENT IMPROVEMENT



Large-scale standardized assessments do not complete the cycle at the student level or even at the school level, though they can offer useful feedback at the district level if done well.

© John Myers, 2003.

areas where they need extra work—but that is not their primary intent.

In order to overcome the limits and minimize the errors in our assessments and evaluations, the assessment tools we use must be valid and reliable. *Validity* and *reliability* are terms usually associated with standardized testing, but the underlying ideas are important to assessment and evaluation more generally.

Validity simply means that the data collected is truly related to the outcomes we intend to measure. For example, social studies curricula across Canada call for students to develop conceptual and procedural understanding related to the disciplines of social studies including history. The new curriculum for Alberta calls for students to develop "historical thinking" which, among other things, involves "the sequencing of events, the analysis of patterns and the placement of events in context."[8] Much assessment in history, however (including surveys and tests that have garnered wide media attention), focuses on collecting information on discrete facts such as names and dates. Sam Wineburg's research demonstrates that knowing a set of discrete facts is not the same as understanding the concepts of the discipline at a sophisticated level or the procedures used in an inquiry for making sense of contradictory accounts.[9] He found that high-achieving senior high students studying the American Revolution knew more facts measured in a test and could outscore some history professors whose specialties were elsewhere. Yet the professors were superior in analyzing primary sources depicting the period, since they had a much deeper understanding of the procedures for making

valid historical claims about the past. Assessment instruments that only collect data on discrete historical facts, then, are not valid in assessing curriculum outcomes related to developing historical thinking.

Reliable assessment instruments are ones that will produce the same (or very similar) results in different situations. There are two components to be concerned about in producing reliable instruments. First, the activity itself should produce clear, consistent evidence of student achievement in the desired area. For example, ambiguous test questions that can be read and answered in many different ways are not particularly reliable because they might produce very different responses from students of similar ability in the same class. They cannot be relied upon to provide a relatively objective sense of student achievement. Second, the evidence should be interpreted the same way by independent observers. In the case of an essay, for example, reliability is demonstrated when qualified independent markers reach similar conclusions about the quality of the work. This kind of inter-rater reliability is achieved when both the assignment and the criteria for success are clearly understood. In standardized testing raters are trained in evaluating student work so that grading is consistent and fair. This is not possible in classroom situations but it is important that those involved—teachers, students, parents, and administrators—know the criteria for evaluation and can see they are being correctly and consistently applied.

It is important to remember that no single instrument, no matter how carefully constructed, can collect all the information needed for a comprehensive evaluation of student progress or be completely valid and reliable. The evaluation of student progress is a very complex process and good teachers build a wide repertoire of approaches to both collecting information and making sense of it.

What Are Some Particular Assessment and Evaluation Challenges for Social Studies Teachers?

Designing and implementing appropriate mechanisms for tracking student progress and providing feedback is a complex endeavour for all teachers, but social studies presents some unique challenges for assessment and evaluation. Social studies includes many ways of thinking and knowing: from the narrative of history and the mathematical and statistical reasoning of economics, to the visual and graphic components of geographic study and the deep understanding of differing perspectives and beliefs underlying the study of public issues and current affairs. Even within a single social studies discipline, the kinds of learnings to be assessed cover a wide range.[10]

Many key social studies outcomes such as critical thinking, social responsibility, and informed decision-making are hard to define compared to outcomes from other subjects. Furthermore, some of these complex goals such as the development of responsible citizenship, may not be evident until after students have left school and engaged in tasks such as informed voting, social action, and other forms of civic participation.

As a result of these varied and contested outcomes, the field of social studies has had great difficulty reaching consensus on its key concepts and purposes, including what constitutes sound assessment and evaluation. Because social studies is concerned with affairs in the real world, it has always been subject to pressures from that world (the political dimension noted earlier). This has certainly been true in the area of assessment. Tests or quizzes purporting to test student knowledge of history, geography, or civics are routinely published in the media, with subsequent blaming and hand-wringing about our wayward youth.[11] These challenges lead to our next question.

How Do I Plan for Sound Assessment and Evaluation?

Sound assessment begins with the learner in mind. This “backwards design” or “design down” process requires we: (1) determine what the learner needs to know and or do as a result of the learning;

(2) identify clear evidence of learning; and (3) design instruction so that students will have opportunities to learn and demonstrate their learnings.[12] Good teachers will keep the following questions in mind when planning for assessment.

1. What is to be assessed? As mentioned above, assessment instruments should be valid; that is, they should provide information about specific, clearly defined elements of students' knowledge or skills. In order to design appropriate methods of assessment, teachers must be very clear about exactly what it is they want information on.
2. What is the purpose of the assessment? The means of collecting information and reporting on it will vary depending on whether the purpose of the assessment is formative—to provide information for the student and teacher to use to monitor or improve progress—or summative—to provide end-point information for students, parents, and/or administrators.
3. What assessment tools will best provide the information we need? Tools will vary depending on what is to be assessed and the purpose of the assessment. For example, an essay might not be the best way to assess students' ability to work with scale and symbols on a map, but an exercise requiring them to plan the best route between two points might work well for this purpose.
4. What form will the data take? This is obviously related to the previous question but there may be some variety within particular assessment tools. A teacher might, as suggested above, ask students to compare and contrast Napoleon's and Hitler's invasions of Russia, allowing students to present their report in written, oral, or graphic form. In the case of students who have difficulty writing, the latter two forms might provide better information about their facility with the skill being evaluated.
5. Who will collect the data: teachers, students, or outside judges? Involving students in self- or peer-assessment can be a very effective teaching and assessment technique. Asking students to use a checklist to rate their peers' performance in a debate by collecting data on several criteria (content, presentation, argument, etc) both directs the students to pay specific attention to these important criteria (teaching) and provides information on how well they understand them (assessment).
6. How often and when will the data be collected? As stated above, teachers assess students all the time. To paraphrase Thomas Jefferson, "The price of democracy is constant vigilance," and constant vigilance applies to assessment as well. Having said that, particular types of, and purposes for, assessment should guide our response to this question. Giving formal written tests to students every day will probably not prove effective, but research on skills teaching indicates that in the early stages of learning a skill, it is very important to provide timely and frequent corrective feedback.
7. What will be done with the data making up the assessment? As stated above, we assess for different purposes and, depending on the purpose, different things will be done with the data. If the assessment is formative the data needs to be provided to the students involved in an understandable way so they can use it to monitor and improve their progress. This is the feedback described above. We have all had the experience of getting back a test or an assignment on which we have not done well, but if we receive little or no feedback, we don't know where we went wrong. This is not particularly useful for understanding how we are doing so far and what we might do to improve. On the other hand, if the assessment is summative, information needs to be provided to the relevant people about student achievement. In the case of large-scale international assessments, for example, the information gathered is of little use to individual students who have probably moved on to the next grade or graduated before the results are even available. It is valuable, however, for curriculum planners, administrators, and teachers in

understanding the degree to which particular groups of students are meeting the objectives assessed.

Underpinning all these questions, of course, is the principle that our assessments must align with curriculum outcomes. Constraining our use of the full range of techniques and tools is our ability to manage, assess, and teach at the same time. So manageability is a factor in our assessment planning decisions. No one can do it all, even the most skillful of teachers! So we should strive to be fair by taking advantage of the many informal opportunities for assessment, such as simply watching our students while they work and giving them opportunities to reflect on their own learning. Such opportunities are built into sound lesson planning. With good instruction we are well on our way to making those first steps towards assessment literacy.

Is sound assessment easy? No, but nothing of value is. Even if we have useful answers for all of the above questions, they will not be perfect answers for all students on all occasions. We will now turn to an examination of some of the specific assessment and evaluation tools that might be helpful in teaching social studies.

What Specific Forms of Assessment and Evaluation Can Be Used in Social Studies?

As stated above, assessing and evaluating student progress is a complex endeavour. Good teachers will develop a range of tools for helping with this task. Some will be traditional forms with which we are all familiar and some will be newer forms such as performance or authentic assessment, which have been developed in response to recent knowledge about teaching and learning. Some options are outlined below but these represent only a small part of what is possible. It is important to remember that each of these tools has strengths and weaknesses and teachers need to be able to select appropriate ones to serve their particular assessment and evaluation purposes.

Selected response tests

Selected response tests, often called "objective paper and pencil tests" consist of multiple choice, true-false, matching, and fill-in-the-blank items. The phrase "selected response" is most accurate since it reflects the procedure of selecting correct responses from a range of possibilities. We use them because we perceive that they are objective, because we experienced these types of tests when we were students, and because the ease of marking makes evaluation more manageable.[13]

In true-false tests, data is quickly collected on a range of knowledge targets such as generalizations or propositions, with little demand on reading ability. Among their limitations are the strong possibility of guessing and the difficulty of designing items beyond the factual knowledge level.

Fill-in tests work for outcomes such as vocabulary understanding and, unlike other forms of selective response, do not promote guessing. On the other hand, poorly designed items and students' poor spelling can result in ambiguous answers that are difficult to score. With matching items, data is collected speedily with fewer demands on reading ability. Matching is particularly useful for assessing student recognition of how ideas are associated. In social studies examples of associations are countries and capitals; definitions and terms; or historical figures and their accomplishments/discoveries. Matching items are also prone to student guessing, and unless designed carefully, are susceptible to the use of irrelevant clues or obvious answers when more than one topic for the matching set is used.

Multiple-choice items, usually consisting of a complete statement of the problem or question (stem/lead); construction of plausible distracters (wrong answers); random placing of the key (correct answer), are the most versatile form of selective response, ranging in use from recall of basic information to interpretation, analogies, and other complex outcomes. Furthermore, guessing is discouraged and well-designed items of the type used in standardized tests are high in reliability. However, they take time and effort to design well.[14] Reading ability may influence some scores, thus threatening the validity of the assessment. Moreover, while some complex

thinking can be assessed, prior knowledge possessed by students can easily turn multiple choice questions into recall items.

Selected response items have the advantage of being easy to administer and score and, if well constructed, are high in validity and reliability. Using selective response for assessment beyond routine knowledge or simple skills, however, is difficult. Furthermore, it is too easy to include irrelevant data in selective response items, allowing students to guess a correct answer without actually knowing anything.

Extended- or essay-response tests

Essays are items that require an extended or constructed written answer to a relatively open-ended question for which a variety of appropriate responses are possible. Along with short-answer questions, selective and essay responses constitute the traditional assessment options in social studies classes.

Essay responses can come in various forms such as writing a letter to the editor of the local newspaper about a current issue or writing a detailed analysis of a case similar to what a judge would write. Essay tests or assignments may be more valid than selected response items in that they are more likely to provide data on important outcomes such as the ability to identify an issue, organize relevant information into an argument, reveal cause-effect relationships, recognize human-environmental interactions, weigh evidence, and so on. The essay can range from a short paragraph to many pages.

Essays do, however, raise serious design and scoring issues. Without clear and precise scoring criteria, applied consistently by teachers, one person's "A" can be another person's "C." The unreliability of teacher scoring on essay items is one reason for the rise of standardized testing. This lack of reliability can be offset somewhat by the use of carefully developed examinations at the provincial level, as is the case in British Columbia and elsewhere. Committees of specialists, especially teachers, meet to design questions based on curriculum outcomes, to be assessed with clear criteria and reliable scoring systems. Where there are no carefully designed, administered, and scored provincial examinations, social studies teachers can work together using the above process to reach consensus on scoring.

Language joins reliability as a scoring issue in essays. Social studies requires a high degree of literacy. Throughout Canadian classrooms, there are many students who struggle with basic literacy skills or for whom the language of instruction is not their first language. Thus, in designing, administering and scoring essay-type questions we should not penalize students unduly for difficulties in initial comprehension of questions or writing of answers, if the goal of the essay question is to reveal other abilities.

We are increasingly aware of the need to support students' learning by providing supports or scaffolds. Scaffolding helps many learners write appropriate responses until they are able to respond more independently. The elements of the writing process in language arts classrooms, including the role of collaborative small group talk, can prepare students for doing higher-quality work than when left to sink or swim on their own.[15]

How can we modify an essay question to meet the diverse needs of our students? We can narrow or broaden the content focus by increasing or decreasing the factors or variables to consider. For example, an essay may challenge students to focus on a narrow historical period such as the 1920s. Other essays might have students look at a larger period such as the 1920s and 1930s. Still other essays could require students to examine (usually for purposes of comparison and contrast) the interactions among geographic and economic patterns in two regions of Canada or reduce the cognitive demands to an analysis of these interactions in one region.

We can also specify the cognitive demands. Some questions can provide additional support for learners by specifying what is required in an answer; for example, which events need to be examined or what factors need to be compared. Some questions offer a context that can serve as a thesis statement while other questions require students to develop their own thesis statements.[16] Vague instructions offer no support for students, as these instructions fail to define criteria for success.

Some questions personalize responses through role play to offer a more authentic context for writing;

for example, students are asked about an event or an idea using a newspaper story format, including an appropriate headline. Other essays present a quotation or provocative proposition and ask students to make an argument based on evidence to support or refute the validity of the statement.

Finally we can direct students to respond in structured ways using action verbs connected to Bloom's Taxonomy of Cognitive Objectives (i.e., describe, explain, analyze, synthesize, and evaluate).[17] This works if students know the meaning of the verbs and see examples of their use in sample responses.

Document-based questions

As outlined above, Alberta and most other provinces have identified procedural understanding as a key goal for social studies education. In other words, students are supposed to be able to use the concepts and procedures of the disciplines involved in social studies such as history, geography, and economics. It is not enough to know what the causes of World War I are; students should also understand something of the way historians arrived at those causes, including what evidence they considered and how they made sense of that evidence. In teaching history in primary grades, curricula across the country recommend the use of primary sources: pictures, artifacts, maps, and written and oral accounts. We have been less successful in using these in assessment. In North America, document-based questions (DBQs) used to be considered appropriate only for senior high students in International Baccalaureate or Advanced Placement programs, though the British have been using "sources" for decades. Now we can use them to bring more authenticity to instruction and assessment. The Begbie Contest in British Columbia has used document-based questions since 1994 for high-school students.[18] Some jurisdictions have used them in elementary grades as well. These questions are not designed to test student's recall of information but rather their abilities to critically use sources as the basis for constructing historical accounts.

PERFORMANCE ASSESSMENT

Performance assessment is considered new although it is as old as assessment itself. Unlike selected response, the assessor does not count correct responses in order to render a judgement. Instead she or he collects data on the process or makes a judgement about the quality of the final product as students actually do something. In some jurisdictions performance tasks serve as end-of-unit assessments or culminating activities instead of the traditional quiz or project. Performance assessment tasks are not add-ons, fillers, or breaks for the teacher, but opportunities to combine instruction with assessment. There are many examples of social studies performance tasks, such as the following:[19]

- Three prominent international conflicts are drawn from newspapers. Students select one of the three, write a summary of the conflict, and discuss the influence of climate, resources, and location on the conflict. As well, the students sketch from memory a map of the region of the world showing national boundaries, capitals, and salient landforms. A legend and compass rose are included.
- Students identify, then compare and contrast, a diverse set of examples of societies organized under, or attempting to organize under, the democratic ideal, with examples drawn from three continents.
- Students analyze a transcribed excerpt of a discussion of a recurring public issue, distinguishing among factual, definitional, and ethical issues, and judging the quality of each participant's contribution.
- The federal deficit is about $500 billion. Students translate this number into a form the average reader of the local paper can understand.

According to Bower, Lobdell, and Swenson, culminating performance activities or projects:[20]

- are central to the unit and its big question
- are known to students in advance

- require students to think deeply about important issues
- ask students to create a meaningful product or performance
- demand students use different learning styles and intelligences
- make clear to students the standards by which their work will be judged
- foster the habit of self-assessment
- allow teachers to act as coaches

AUTHENTIC ASSESSMENT

Authentic assessments are a particular form of performance assessment where students are required to perform a real-life activity and an assessment is made based on that activity. Students in a grade 3 class studying their community, for example, might be asked to do an assessment of the area around their school (sidewalks, parks, public places) for wheelchair accessibility and prepare a report for a relevant civic body such as a committee of city council. The activity does not necessarily have to relate to contemporary situations. Currently, some Acadian people are asking that the Queen apologize for the expulsion of their ancestors from the Maritimes. Students might be asked to prepare a recommendation for the Crown on whether or not, considering the historical evidence, such an apology would be justified.

STRUCTURED OBSERVATION

Many performance and authentic assessments will include "hard" evidence of student progress such as written components, constructed models, and visual representations, but often much of the evidence of student progress will be gathered through watching them work. For example, many social studies outcomes are related to that important and hard-to-define term "thinking." However thinking is defined, conventional tests reveal only so much. One approach is to determine what thinking looks like when we see it. Such "intelligent behaviours" are "habits of mind": a repertoire of mindful strategies we use when faced with problems or decisions. If we are going to provide useful feedback, either of a formative or summative nature, about something like thinking, we need to say more than, "John has shown great improvement in critical thinking." We need to be able to be more specific about what we mean and that kind of specificity can be obtained by structured observation: watching for, and collecting evidence on, particular behaviours. In the area of thinking, for example, we might watch for: [21]

- Perseverance: Do students give up or back up and use a different strategy if the first ones do not work?
- Decreased impulsiveness: Do students blurt out answers and make many corrections in their written responses or do they pause before answering, make sure they understand the learning task, and consider the responses of others in building arguments?
- Flexible thinking: Do students use the same approaches for different problems or do they use and weigh the merits of alternative strategies, consider the approaches of others, and deal with more than one classification system simultaneously?
- Metacognition: Are students unaware of how they learn or do they describe and reflect on the processes they used in learning?
- Careful review: Do students hand in uncorrected or unedited work as soon as it is done or do they take time to review and edit?

These observations would be structured not only in the sense that the teacher looks for evidence for precise criteria like these but also in the sense that some form of record is kept of the observations such as checklists or anecdotal notes.

For all of these assessment tools other than limited response items a difficulty is designing consistent and fair ways to make evaluations based on them. A common way to deal with this challenge is

to design rubrics that describe clear criteria for making evaluative judgements. A rubric is an achievement scale: a set of scoring guidelines for evaluating student work. Rubrics answer questions such as: By what criteria should performance be judged? What should we look for when we judge performance tasks? What is quality? How can different levels of quality be described and distinguished from one another? The word "rubric" comes from the Latin word for "red." In the European Middle Ages it referred to the highlights of a legal decision as well as the directions for conducting religious services, found in the margins of liturgical books—both written in red.

When an activity requires a complex performance to assess complex outcomes scored at more than two levels, then a rubric is required. In one of our studies we asked grade 9 students from more than a dozen schools to write a personal letter at a specific point in historical time and place to a public figure in Canada. They were asked to write a letter in the first person and allowed to choose their role from either gender and from a range of occupations authentic to the time and place.

Before determining specific criteria, it is important for students to discuss what quality looks like. Fortunately, writing in role is a common activity in the Canadian history course in Ontario, so students could see and critique samples. Students also had access to real letters written during such periods as World War I and the Depression. If students have a role in determining criteria they are more likely to understand what is required and develop some sense of ownership and commitment to achieving the learning goal.

The teachers and student teachers scoring the hundreds of responses first did a scan of the letters to sort into three general categories: "good," "ok," and "poor." Then they discussed why letters were placed into each category. Veteran teachers' experience enabled them to better articulate and defend their reasons on the basis of some shared values about the nature of historic empathy and the role of evidence, two key concepts in historical understanding. From this task and discussion the following rubric was built based on what students could actually demonstrate.

Student work should demonstrate:

1. Historical Argument: Could students present historically accurate and relevant arguments well supported with evidence?
2. Context: Could students display a strong sense of historical empathy reflecting a broad understanding of historical circumstance, events, and relationships?
3. Role: Could students articulate a personal point of view within a clearly identified historical role?
4. Organization: Could students offer a well-organized and coherent letter in paragraph form?
5. Persuasiveness: Could students use persuasive language appropriate for the historical audience?

A word on weighting criteria for grading purposes: the issues around converting performance levels to marks or grades are complex.[22] Teachers should consider which criteria are absolutely critical, without which the task cannot be accomplished. Which are very important, without which the quality of the product is seriously flawed? Which criteria may not be critical to the product though they may affect the quality?

Does this sound subjective? Yes, but remember that evaluation using a grade or rubric represents a judgement about what is valued. If we take the mystery out of these judgements, students are more likely to meet our standards.

Conclusion

The trends in assessment and evaluation in social studies, as in other parts of the curriculum, are in a state of flux. The issues are complex; easy solutions are impossible. Assessing student progress is one of the most difficult and public things teachers do. It takes time and hard work to develop the range of procedures and instruments necessary to adequately provide feedback to both teacher and students (formative assessment) and end-point information

for students, parents, and others in the education system (summative assessment). Informed, reasoned discussion by researchers, practitioners, and policy-makers is a must. If this chapter offers us a direction for becoming more assessment literate, then it has done its job.

Endnotes

1 Joseph M. Kirman, *Elementary Social Studies* (Scarborough, ON: Prentice Hall, 1991), 11.

2 Overviews of this expanded role can be seen in Kieran Egan, "Testing What for What?" *Educational Leadership* 61, no. 3 (2003): 27–30; and Steve Alsop and Larry Bencze, "A Tale of Two Scientists: Professional Scientist/Citizen Scientist," *Orbit* 31, no. 3 (2000): 21–24.

3 Three collections of articles illustrating recent assessment reform can be found in *Orbit* 30, no. 4, (2000); *Canadian Social Studies* 34, no. 1 (1999); and Roland Case and Penney Clark, eds., *The Canadian Anthology of Social Studies: Issues and Strategies for Teachers* (Vancouver: Pacific Education Press, 1997).

4 Charles Ungerleider, *Failing Our Kids: How We Are Ruining Our Public Schools* (Toronto: McClelland & Stewart, 2003), 252–253 provides an overview (already out of date) of the changing provincial testing scene in Canada.

5 Ruth Childs, "The EQAO Tests and the OTQT: Critical Consumers Needed" (keynote address, OISE/UT Preservice Assessment and Evaluation conference, Toronto, February 11, 2004).

6 This section is an adaptation of two earlier publications by John Myers and Fiorella Finelli: *Canadian History: Patterns and Transformations, Teachers' Resource Guide* (Toronto: Irwin, 2003); and "Assessment and Evaluation", in *Civics Today: Teachers' Resource Guide*, ed. Jennifer Watt, et al. (Toronto: Irwin, 2000), 18–32.

7 Robert J. Marzano, Debra J. Pickering, and Jane E. Pollock, *Classroom Instruction that Works: Research-based Strategies for Increasing Student Achievement* (Alexandria, VA: Association for Supervision and Curriculum Development, 2001).

8 Alberta Learning, *Social Studies Kindergarten to Grade 12 Validation Draft* (Edmonton: Author, May 2003), 9.

9 Samuel S. Wineburg, "Historical Problem Solving: A Study of the Cognitive Processes Used for the Evaluation of Documentary and Pictorial Evidence," *Journal of Educational Psychology* 83, no. 1 (1991): 75–87.

10 See, for example, Canadian Council for Geographic Education, *Canadian National Standards for Geography: A Standards-based Guide to K–12 Geography* (Ottawa: Royal Canadian Geographic Society, 2001), for the range of learnings to be assessed in just one component of social studies.

11 Daniel Gardner, *Youth and History: Policy Paper and Survey Results* (Toronto: Dominion Institute, 1997) was the first of a series of results based on national quizzes of Canadian students' knowledge of Canadian history and citizenship.

12 Grant Wiggins and Jay McTighe, *Understanding by Design* (Alexandria, VA: Association for Supervision and Curriculum Development, 1998).

13 Richard J. Stiggins, *Student-involved Classroom Assessment*, 3rd ed. (Upper Saddle River, NJ: Merrill Prentice Hall, 2001).

14 John Myers et al., "Doin' the DBQ: A Project in Designing Useful Assessments" (in progress, Toronto, OISE/UT; first phase finished in 2003). This project, carried out with the assistance of fifty-eight student teachers, has demonstrated, among other things, both the importance and the challenge of designing useful multiple-choice questions.

15 The literature on the role of talk in small groups is vast. See, for example, the International Association for the Study of Cooperation in Education website, www.iasce.net, and its links.

16 See, for example, the essay questions in the Begbie Canadian History Contest, http://www.begbiecontestsociety.org/, which held its first contest in British Columbia in 1994 and is now a national contest.

17 R.J. Cornfield et al., *Making the Grade: Evaluating Student Progress* (Toronto: Prentice Hall, Canada, 1987).

18 The Begbie Canadian History Contest, http://www.begbiecontestsociety.org/.

19 Adapted from Walter C. Parker, *Renewing the Social Studies Curriculum* (Alexandria, VA: Association for Supervision and Curriculum Development, 1991).

20 Bert Bower, Jim Lobdell, and Lee Swenson, *History Alive! Engaging All Learners in the Diverse Classroom*, 2nd ed. (Palo Alto, CA: Teachers' Curriculum Institute, 1999).

21 Art L. Costa and Bena Kallick, *Habits of Mind: A Developmental Series* (Alexandria, VA: Association for Supervision and Curriculum Development, 2000).

22 See Ken O'Connor, "Grading—An Exercise in Professional Judgment," *Orbit* 30, no. 4 (2000): 40–42.