JOURNAL OF RESEARCH IN SCIENCE TEACHING VOL. 40, NO. 1, PP. 34–52 (2003)

# From Nationwide Standardized Testing to School-Based Alternative Embedded Assessment in Israel: Students' Performance in the Matriculation 2000 Project

Yehudit J. Dori[1,2]

[1]*Department of Education in Technology and Science, Technion, Israel Institute of Technology, Haifa 32000, Israel*

[2]*Center for Educational Computing Initiatives, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139-4307*

*Received 10 December 2001; Accepted 13 August 2002*

Abstract: Matriculation 2000 was a 5-year project aimed at moving from the nationwide traditional examination system in Israel to a school-based alternative embedded assessment. Encompassing 22 high schools from various communities in the country, the Project aimed at fostering deep understanding, higher-order thinking skills, and students' engagement in learning through alternative teaching and embedded assessment methods. This article describes research conducted during the fifth year of the Project at 2 experimental and 2 control schools. The research objective was to investigate students' learning outcomes in chemistry and biology in the Matriculation 2000 Project. The assumption was that alternative embedded assessment has some effect on students' performance. The experimental students scored significantly higher than their control group peers on low-level assignments and more so on assignments that required higher-order thinking skills. The findings indicate that given adequate support and teachers' consent and collaboration, schools can transfer from nationwide or statewide standardized testing to school-based alternative embedded assessment. © 2003 Wiley Periodicals, Inc. J Res Sci Teach 40: 34–52, 2003

Educators' awareness of the need to modify the traditional testing system in high schools has increased since the mid-1980s throughout the western world (Black, 1995a, 1995b). In the mid-1990s, multiple choice items and standardized test scores have been supplemented with new methods, such as portfolios, hands-on, performance assessment, and self-assessment (Bexter, Shavelson, Goldman, & Pine, 1992; Ruiz-Primo & Shavelson, 1996; Tamir, 1998). Nowadays, the

*Correspondence to*: Y.J. Dori; E-mail: yjdori@tx.technion.ac.il
DOI 10.1002/tea.10059
Published online in Wiley InterScience (www.interscience.wiley.com).

researchers are investigating the effect of alternative assessment methods on various groups of students (Flores & Comfort, 1997; Klein et al., 1997; Lawrenz, Huffman, & Welch, 2001). Other studies investigate how teaching and learning in science can benefit from embedded assessment (Treagust, Jacobowitz, Gallagher, & Parker, 2001).

In Israel, most high schools employ a nationwide, standard, traditional battery of tests known as the matriculation examinations. This practice was instituted 50 years ago and has basically remained the same. The results of these tests constitute a major factor in determining students' prospects of being admitted to institutions of higher education. In 1995, the Israeli Ministry of Education launched a 5-year project called Matriculation 2000. Aimed at moving from the traditional and national examination system to a school-based alternative embedded assessment, this project encompassed a select group of 22 high schools from various communities in the country. The major objectives of the Matriculation 2000 Project were to foster deep understanding, higher-order thinking skills, and students' engagement in learning through alternative teaching and assessment methods. This article describes research conducted during the fifth year of the Project at 2 experimental schools and 2 control schools. The research was aimed at investigating students' learning outcomes in chemistry and biology in the Matriculation 2000 Project. Students' and teachers' attitudes toward the alternative teaching and assessment methods employed in the Project were also studied, but this is beyond the scope of this report.

## Theoretical Framework

> Many politicians, and most of the general public, have a narrow view of testing and assessment. The only mode which they know and understand is the conventional test, which is seen as a reliable and cheap way of comparing schools and assessing individuals (Black, 1995b, p. 462)

Until the last decade, standardized testing and authentic assessment were perceived as two different cultures. Standardized testing was traditional and quantitative, whereas authentic assessment was considered to be qualitative and often alternative (Kleinsasser, Horsch, & Tastad, 1993; Wolf, Bixby, Glenn, & Gardner, 1991). The standardized testing culture was based on the premise that the measurement experts were authorized to construct a test and interpret its results, with little or no relation to the teaching that occurred in the classroom. The authentic assessment culture developed as a result of teachers' and educators' dissatisfaction from the standardized tests, which were disconnected from the teaching (Baker & Herman, 1983; Birenbaum & Shaw, 1985; Linn, 1983). Another claim against standardized tests was that because teachers were teaching to the tests, they neglected developing their students' higher-order thinking skills (Frederiksen, 1984). Testing which is at the same time standardized and authentic has started to emerge recently.

Tamir (1998) defined student assessment as a collection of information on students' outcomes while learning is taking place (formative assessment) or after the completion of the learning task (summative assessment). According to Lewy (1996), assessment should be based on a set of tasks, including giving oral responses, writing essays, performing data manipulations with technology-enhanced equipment and selecting an alternative from a list of possible options. Madaus and Kellaghan (1993) defined assessment using the ''three P's'': performance, portfolio, and product.

In recent years, two concepts related to assessment have been receiving researchers' attention. One is alternative assessment and the other is embedded assessment. Nevo (1995) noted that when alternative assessment is applied, students are evaluated on the basis of their active

performance in using knowledge in a creative way to solve worthy problems. The problems have to be authentic, nonroutine, and multifaceted with no obvious solutions. Embedded assessment is an ongoing process that emphasizes the integration of assessment into teaching. Teachers can use embedded assessment to guide instructional decisions for making adjustments to teaching plans in response to the level of students' conceptual understanding (Treagust et al., 2001). In this research the two concepts—alternative assessment and embedded assessment—were combined, as the Matriculation 2000 Project employed embedded alternative assessment as an integral part of the teaching process from 10th grade throughout 12th grade.

High-level administrators and experts make many decisions in education. However, schools and teachers should be more involved in new developments in assessment methods (Nevo, 1995). Indeed, the National Science Education Standards (NRC, 1996) indicated that teachers are in the best position to use assessment data to improve classroom practice, plan curricula, develop self-directed learners, report students' progress, and research teaching practices. According to Treagust et al. (2001), the change from a testing culture, which is the common assessment practice, to an assessment culture, be it embedded or alternative, is a systemic change. Such a profound reform mandates that teachers, experts from educational institutions, and testing agencies rethink the educational agenda and the role of assessment.

As participants in authentic evaluation, researchers cannot set aside their individual beliefs and viewpoints, through which they observe and analyze the data they gathered (Guba & Lincoln, 1989). To attenuate the bias such individual beliefs cause, evaluation of educational projects should include opinions of the various stakeholders as part of the data. With this in mind, students were considered major stakeholders in the reform and their learning outcomes in the Matriculation 2000 Project were assessed accordingly. Other important stakeholders include the teachers and principals who participated in the Project. Although these participants were part of the research, elaborating on findings that reflect their perspectives warrants a separate paper.

## Matriculation Examinations in Israel: Present and Future

Matriculation examinations in Israel have been the dominant summative assessment tool of high school graduates over the past half-century. The grades of the matriculation examinations, along with a psychometric test (analogous to SAT in the United States), are a critical factor in college and university admission requirements. This nationwide battery of tests is conducted centrally in 7 or 8 different courses, including mathematics, literature, history, English, and at least one of the sciences (physics, chemistry, and biology). The Ministry of Education determines the goals and contents of each course. A national committee appointed by the Ministry is charged with composing the corresponding tests and setting criteria for their grading. This leaves the schools and the teachers with little freedom to modify either the subject matter or learning objectives. However, students' final grade in the matriculation transcript for each course is the average of the school grade in the course and the pertinent matriculation examination grade.

The advantage of the matriculation examinations is the high standards that the Ministry of Education sets throughout the country. This results in a high level of confidence on the part of the universities in Israel in the matriculation transcript grades, enabling these institutions to admit students without entry examinations. However, both pedagogical and sociocultural aspects of the matriculation system have been criticized. Pedagogically, this system forces teachers to emphasize teaching topics that will maximize their students' likelihood of success in the examinations. This takes away from teachers' efforts to ensure meaningful learning and the development of students' higher-order thinking skills. Moreover, learning to pass a battery of tests stresses students and hinders their ability to perform at their best.

Claims have also been made that the tests cater to populations of upper-class socioeconomic and cultural communities. As a result of this bias, the percentage of students in higher education from lower socioeconomic communities and minorities is significantly smaller than their percentage in the entire population.

A national committee headed by Ben-Peretz (1994) examined the issue of the matriculation examinations from two aspects: (a) pedagogical—quality of teaching, learning and assessment, and (b) sciocultural—the number and distribution of students from diverse communities eligible for the Matriculation Diploma.

Addressing the sociocultural aspect, several researchers (Gallard, Viggiano, Graham, Stewart, & Vigiliano, 1998; Sweeney & Tobin, 2000) have claimed that educational equity goes beyond the notion of equal opportunity and freedom of choice. How learning is fostered should be examined to verify whether students are allowed to use all the intellectual tools that they bring with them to the classrooms.

The Ben-Peretz Committee indicated that in their current format, the matriculation examinations do not reflect the depth of learning that takes place in many schools nor do they measure students' creativity. The Committee's recommendations focused, among other issues, on providing high schools with increased autonomy to apply alternative embedded assessment methods instead of the nationwide matriculation examination. The school-based assessment would combine traditional examinations with alternative embedded assessment methods in a continuous fashion throughout high school, from 10th through 12th grade. The proposed assessment methods included projects, portfolios, laboratory research, and assignments involving teamwork. The Committee called for nominating exemplary schools, which would be mentored and monitored by experts in 1, 2, or 3 courses in each school. The school grades in those courses would be recognized as the standard matriculation grades.

As a result of the Ben-Peretz Committee's recommendations, the Ministry of Education launched a 5-year project, titled Matriculation 2000. The Project aimed at developing deep understanding, higher-order thinking skills, and students' engagement in learning through changes in both teaching and assessment methods. During the period of 1995–1999, 22 schools from various communities participated in the Project. These schools represented a variety of communities, academic levels, and sectors, including urban, secular, religious, and Arab schools. The courses taught in these schools under the umbrella of the Matriculation 2000 Project were chemistry, biology, English, literature, history, social studies, Bible, and Jewish heritage. An expert group accompanied each school, providing the teachers with professional development programs, which included guidance in teamwork, school-based curriculum, and alternative embedded assessment methods. These expert groups were guided and managed by an overseeing committee headed by Ben-Elyahu (1995).

Tenth grade was the first year a student participates in the Project; 12th grade was the last one. All the students in the Project who studied chemistry and biology took the courses at the highest level of 5 units, which is comparable to Honors class in the U.S. high school system. Most of the students who studied liberal art courses took them at the basic level of 2 units, which is comparable to Curriculum II in the U.S. high school system. Understanding the need for a professional assessment of the Matriculation 2000 Project, the Ministry of Education's Chief Scientist issued a call for proposals to investigate the effect of the Project on students' performance as well as on the school system. Following a review process, the Ministry appointed an independent external academic research group, headed by this author, to study the Project outcomes and ramifications during the fifth year of the Project. This article focuses on the part of the research that addresses the effect of the Matriculation 2000 Project on students' performance in the science courses.

## Research Objective and Question

The research objective was to investigate students' learning outcomes in chemistry and biology in the Matriculation 2000 Project. The assumption was that alternative embedded assessment has some effect on students' performance. The research question addressed in this article was: How does the Project affect students' performance in chemistry and biology?

## Research Population and Settings

The research population included students ($N = 243$) from 12th grade in four heterogeneous high schools (Table 1). As noted, 22 exemplary schools participated in the Matriculation 2000 Project, but chemistry and biology were taught within the Project's framework only in Schools A and B. In Israel most of the students who elect to take the matriculation examination in sciences study their elected course at the highest (Honors) level, i.e., 5 units. Therefore, there was no school in the Project in which chemistry or biology was taught at a lower level.

School A, in which chemistry was taught within the Project's framework, is located in a high socioeconomic neighborhood. School B, in which biology was taught, is located in an intermediate socioeconomic neighborhood. Unfortunately, exemplary practice in science is more difficult to execute in schools with community constraints.

The two experimental schools that participated in the research are representative of the 22 schools as far as science is concerned. However, these 22 schools do not constitute a random sample of the high schools in Israel. Rather, they were chosen by the Committee on the basis of the quality of the proposals they submitted before the initiation of the Project, and the perceived probability that school teams would be able to carry out the Project throughout the entire 5-year period. Teachers and principals of these schools were more dedicated and experienced than an average school team. These schools can therefore be considered as exemplary.

As Table 1 shows, 140 students from 12th grade responded to achievement tests in chemistry (School A) and biology (School B). These students studied the relevant science course within the Project and therefore served as the experimental group for the purpose of assessing performance. Another 103 12th-grade students, who served as a control group, responded to identical achievement tests in chemistry and biology. These students were from two other high schools (labeled C and D) which did not participate in the Project but were at an academic level and socioeconomic background comparable to that of the experimental schools.

To enable comparison between the experimental and control groups, two aspects were investigated: (a) The academic level of the experimental and control groups (Schools A–D) was determined by the matriculation scores of students over 1991–1996, and (b) the grades teachers

Table 1
*Research population and research instrument administered*

| Research Group | School | Course for Which Achievement Test Was Administered | *N* | Achievement Test Administered to Grade |
|---|---|---|---|---|
| Experimental | A | Chemistry | 59 | 12th |
| | B | Biology | 81 | 12th |
| | Total | | 140 | |
| Control | C | Chemistry | 38 | 12th |
| | C | Biology | 35 | 12th |
| | D | Biology | 30 | 12th |
| | Total | | 103 | |

gave to the students at the end of 11th grade in the participating schools were collected and analyzed. The grades in chemistry and biology of the experimental students were: $\bar{x}_{chemistry} = 84.9$ [standard deviation (SD) = 9.9] and $\bar{x}_{biology} = 84.1$ (SD = 12.1). For the control students, the analogous grades were: $\bar{x}_{chemistry} = 84.5$ (SD = 9.1) and $\bar{x}_{biology} = 81.7$ (SD = 12.8).

Because these two measurements showed no significant differences, the experimental and control groups were considered identical.

The alternative embedded assessment methods applied in the experimental schools included portfolios, individual projects, projects done in teams, written and oral tests, class and homework assignments, self-assessments, field trips, inquiry laboratory activities, concept maps, scientific article reviews, and project presentations. These embedded assessment methods were integrated into the teaching throughout the school year. The most prevalent methods, as reported by teachers and principals, were written tests, class and homework assignments, individual or group projects, and scientific article reviews. In chemistry, the group effort was a miniresearch project that spanned over half a year. Students were required to raise a research question, design an experiment to investigate the question, carry it out, and draw conclusions from its outcomes. In biology, the students presented individual projects to their peers in class and expert visitors in an exhibition.

To gain deeper insight into the Project setting, the researchers visited School A and School B and met with the Project teams there (Dori, Barnea & Kaberman, 1999). Based on these visits, we subsequently describe how teachers uniquely embedded the spirit of the Matriculation 2000 Project in these schools.

During the middle of 10th grade, students in School A were given the opportunity to decide whether they wanted to elect 5 units of chemistry (which is analogous to the Honors level in the United States). Students who chose this option studied in groups of 20 per class for 8 hours per week throughout 11th and 12th grades. These students focused on 80% of the topics that were included in the national standard 5-unit chemistry matriculation examination (which they did not take because they were part of the Matriculation 2000 Project). They were also exposed to many more laboratory activities, as well as to reading and reviewing scientific articles. Because the current standardized chemistry matriculation examination is based on a paper and pencil test without a laboratory component, teachers in traditional chemistry 12th-grade classes avoid spending time on laboratory activities.

Alternative assessment was embedded throughout the curriculum. The teachers' teamwork included a weekly 2-hour meeting for designing assessment tools and setting assessment criteria. Teachers graded group projects and scientific article reviews according to topic rather than class affiliation. This, they claimed, increased the level of reliability and objectivity of the grades. The chemistry teachers in this school testified that their students enjoyed studying chemistry within the Project more than their peers from previous years who studied in the traditional mode.

In School B, where biology was taught in the Project, the traditional assessment elements included two tests (one in genetics and the other in photosynthesis), which accounted for 20% of the final grade, and a quiz (10%). The alternative assessment means were field trips, within which biological projects (known as ''biotops'') were conducted (35%), a portfolio summarizing the inquiry laboratories (25%), class involvement (5%), and self-assessment (5%). The teachers set clear and precise criteria to validate the scoring in the various assessment tools. Describing the assessment tools, teachers reported that their teaching objectives were commensurable with the learning outcomes, as assessed by these various tools. 12th-grade students worked in teams and carried out individual projects, which were displayed and presented to peers and the Project researchers in the middle of the school year.

Discussing the Project's environment and future, teachers of both School A and School B indicated that they had undergone a conceptual change in teaching and assessment methods, and

felt that they passed a point of no return. Although they felt great satisfaction, teachers noted that the amount of time required to carry out all these teaching and assessment activities far exceeded what they had been used to in traditional teaching. Indeed, the Ministry of Education compensated the teachers for these efforts in the Project by getting 2 extra hours for teacher team meetings, but they expressed their concern as to what would happen when the Project's funding was over.

Overall, the science courses taught in the experimental schools included more laboratory experiments and research projects, scientific article reading, and authentic assignments that fostered higher-order thinking skills. Both the experimental and the control teachers were experts in their field, but the teachers who prepared their students for the external matriculation examinations were dedicated more to covering material than developing a variety of thinking skills.

## Research Instruments

The effect of the Project on students' performance in chemistry and biology was measured through achievement tests that were administered to the experimental and control 12th-grade students. Three science education experts constructed each test and set predetermined criteria for its grading. Four other senior science teachers (2 chemistry and 2 biology teachers) validated the contents and difficulty level of the tests. One aspect of the tests' reliability was calculated. The internal consistency Cronbach $\alpha$ was .76 for the chemistry test and .65 for biology.

The science teachers were on sabbatical that year and did not teach a course. Therefore, they could grade each test independently and objectively. The final test grade was computed as the average of the scores assigned by two graders. In $<5\%$ of the cases, the difference between the grades each senior teacher assigned was $> 10$ (out of 100) points. In such cases, one of the experts who participated in constructing the test and the criteria also evaluated the test independently. This expert, who took in account the three grades, determined the final grade.

The assignments in these tests were categorized into low-level and high-level ones. Resnick (1987) stated that although it is difficult for researchers to define higher-order thinking skills, these skills could be recognized when they occur. Based on Costa (1985), Dillon (1990), and Shepardson and Pizzini (1991), and using TIMSS (Third International Mathematics and Science Study) (Shorrocks-Taylor & Jenkins, 2000) taxonomy, the two assignment types were designed.

Low-level assignments required the students to recall knowledge and understand concepts. Typical assignments at this level were:

- Draw Louis structures or geometrical structures of molecules whose formulae are provided.
- Provide two examples of peptides generated from three given amino acids.
- Tabulate given data regarding skin cancer patients in a township in Australia; researchers believe skin cancer is caused by depletion of the ozone layer.

A low-level assignment is usually characterized as having a definite, clear, correct response, so it is relatively easy to assess and grade it, and the assessment is, for the most part, on the objective and neutral side. The opposite is true for high-level assignments, discussed below, in which the variability and range of possible and acceptable responses are far greater and there is not just one school solution. By nature, assessing such assignments is more demanding and challenging, as the assessors need to be more open to different viewpoints and accept novel ideas or original, creative responses that the teachers had not thought of before.

High-level assignments were open-ended and required various combinations of application, analysis, synthesis, inquiry, and transfer skills. Open-ended assignments promote different types

of student learning and demonstrate that different types of knowledge are valued (Resnick & Resnick, 1992; Wiggins, 1989). Assignments at this level were based on biology/chemistry-related case studies. The case study method, also referred to as problem-based method, was chosen as a means to foster and assess higher-order thinking skills (Herried, 1994, 1997; Dori & Herscovitz, 1999; Dori & Tal, 2000). An example of a case study, which was included in the chemistry test, is presented below. It involves the research of the three 1998 Nobel laureates in physiology or medicine "for their discoveries concerning nitric oxide as a signaling molecule in the cardiovascular system" (http://www.nobel.se/medicine/laureates/1998/press.html).

> It was a sensation that this simple, common air pollutant, NO, which is formed when nitrogen burns, for instance in automobile exhaust fumes, could exert important functions in the organism. It was particularly surprising since NO is totally different from any other known signal molecule and so unstable that it is converted to nitrate and nitrite within 10 seconds. NO was known to be produced in bacteria but this simple molecule was not expected to be important in higher animals such as mammals.

Having read this case study, the students were requested to respond to high-level assignments (problems) such as the following: (a) Pose questions that refer to a given case study and suggest a solution to one of the questions. (b) Design an experiment or research study. Describe the research question, the variables, and the research settings. (c) Propose a creative way to present to your peers the main ideas in the case study.

As noted, grading the open-ended responses to the high-level assignments was based on a detailed set of criteria. It enabled the senior teachers to determine whether the response was excellent (which scored 14 of the 14 possible points), adequate (10–12), partially admissible (6–9 points), or inadequate (1–5 points), and grade it accordingly. Examples for various responses and how they were graded are provided in the Findings section.

Table 2 lists the number of assignments (problems) in the test of each course. For example, the test in chemistry consisted of 10 assignments, of which students had to choose 7. Because 3 of the 10 assignments were at a high level, a student could choose as many as 6 low-level and just 1 high-level assignment, or as few as 4 low-level assignments and 3 high-level ones. The test score for each student in a course was computed separately for low-level (knowledge) assignments and high-level (high-order thinking skills) assignments. Two types of scores were calculated for each student: absolute score and relative score. The absolute score was computed as the sum of points he or she scored for all the high-level (low-level) assignments divided by the sum of maximum possible points for all the high-level (low-level) assignments in the test, multiplied by 100. The relative score was computed as the sum of points he or she scored for all the high-level (low-level) assignments divided by the sum of maximum possible points only for the high-level (low-level) assignments he or she chose, multiplied by 100.

Table 2
*Distribution of low- and high-level assignments*

| Course | Total No. of Assignments in Test | No. of Low-Level Assignments in Test | No. of High-Level Assignments in Test* | No. of Required Assignments in Test |
|---|---|---|---|---|
| Chemistry | 10 | 7 | 3 | 7 |
| Biology | 14 | 10 | 4 | 11 |

*A student was required to respond to at least one high-level assignment.

For example, the test in chemistry consisted of three high-level assignments, each worth 14 points. If a student responded to two high-level assignments, the maximum points for these assignments was $2 \times 14 = 28$. If she scored 10 and 14 points for these two assignments, her relative score for high-level assignments was $100 \times 24/28 = 85.7$, whereas her absolute score was $100 \times 24/(3 \times 14) = 57.1$. The quantity of high-level assignments a student selected served as another indication of the student's higher-order thinking skills, and enabled us to compare the two research groups.

## Findings

The effect of the Project on students' performance in chemistry and biology was measured using the achievement tests described above. Each test included a different unseen case study followed by a set of assignments. Examples include: (a) a case study on homocysteine, high levels of which mark cardiovascular risk; (b) interactions between fungi and pine plants; and (c) the function of NO in medicine, which has been elaborated upon in the Research Instruments section.

### *Sample Responses for High-Level Assignments*

To illustrate the variety of responses to the high-level assignments and how they were scored, sample responses for three case studies in chemistry and biology are provided below.

- Assignment: If you were required to present to your peers the main ideas in the case study, what creative way would you choose to do it?

> Excellent response: First, I would prepare myself by reading and analyzing the case study [interactions between fungi and pine plants] and then I would summarize the main ideas and respond to all the assignments. Later, I would go to the library and search for additional resources to gain deeper understanding of the subject and be able to respond to questions my peers might ask concerning the case study, but for which answers cannot be found in the case study itself.
>
> Second, I will prepare three teaching aids to enhance understanding. (a) I will present transparencies or posters with the main ideas and several basic problems for my peers to solve to initiate a discussion in class. (b) I will divide the class into groups and provide them with complex questions on note cards. Each group will be asked to relate the case study to another topic we studied before, such as genetics, ecology, or microbiology. After the group discussions, a representative will present a summary and conclusion to the whole class. (c) Finally, I will design a game or a crossword puzzle that will show the relationship of various organisms to pine plant.

[Score: 14 points of 14; the student showed a structured approach to learning, suggested a variety of teaching aids, and successfully linked the topic to the students' prior knowledge.]

> Adequate response: I will present the main ideas by building a model which I will use to demonstrate the effect of NO passing through the cell membranes to muscle cells that wrap the arteries. This effect causes the relaxation of the cells, yielding dilation of the arteries that decreases blood pressure. I will do this by using a ring-shaped balloon. When air is released slowly through a valve, the inner loop of the ring expands. This will help demonstrate the phenomenon to my friends.

[Score: 12 points of 14; only one teaching aid (model) is suggested with a creative explanation.]

Partially admissible response: I will bring an experiment that shows decomposition of homocysteine with the help of vitamins. Then I will show how homocysteine precipitates and then I will present a model of a sclerotic artery.

[Score: 9 points of 14; two teaching aids (experiment and a model) are suggested, but the response lacks sufficient scientific explanation.]

Inadequate response: I will distribute the case study on papers and ask the students to solve the problems. However, this is not going to create an interest among my friends because it is not part of the curriculum.

[Score: 3 points of 14; only one traditional teaching aid (paper) with no creativity or any scientific explanation.]

- Assignment: Pose questions that refer to a given case study and suggest a solution to one of the questions.

Excellent response: (a) Is it possible to slow or even heal the sclerosis that lack of vitamins cause by treating the patients with these necessary vitamins? (b) Why are high blood pressure and smoking factors that accelerate blood vessels diseases? (c) How can we prevent homocysteine from contracting blood vessels and/or generating blood clots?

Solution to Question 3: To find out how to neutralize the adverse effects of homocysteine I will isolate it in an identical environment (in blood solution) and investigate to which site in the blood cells this amino acid links. Once this site is found, I will apply genetic engineering techniques to alter that site to prevent homocysteine from linking to the blood cell.

[Score: 14 points of 14. The questions the student posed are related to issues that are beyond the scope of the case study. These questions stimulate higher-order thinking. The solution is a logical and potentially feasible experiment.]

- Assignment: You were asked to join a research team that investigates homocysteine (or NO, in the other case study). Design an experiment or a research project that you would like to carry out. Describe the research question, the variables, and the research settings.

Adequate response:

Research question: How does the concentration of homocysteine in the body affect blood vessel diseases among population with high blood pressure?

Research variables: concentration of homocysteine; severity of blood vessel diseases.

Experiment: Injection of homocysteine or preventing its neutralization in mice with the aforementioned problems and follow-up.

Research setting: Lab with mice, which would be monitored with proper instrumentation to detect changes in their blood vessels.

[Score: 12 points of 14; the student responded to all requested items. Two points were taken off because the student did not indicate which is the dependent variable and which is the independent one.]

Inadequate response:

Experiment: I will expose them to different levels of NO and check their reaction.

Research question: How do different NO levels affect cancer cells in comparison to the effect on blood cells?

[Score: 4 points of 14; the student did not respond to all the requested items and did not define the research variables and experimental setting.]

*Students' Performance in Low- and High-Level Assignments*

The absolute and relative average scores were examined in low- and high-level assignments by Project course—chemistry and biology—and by research group. As explained in the Research Instruments section, the absolute scores are computed with respect to the maximum total points that could be scored. The relative average scores are adjusted to the actual number of points a student scored relative to the maximum number of points he or she could have scored in the assignments at the low or high level, which he or she chose.

Comparing the absolute average scores of experiment vs. control students for the low-level assignments, as shown in Table 3, significant differences were found in chemistry in favor of the experimental students. In biology the difference between the two research groups was also in favor of the experimental students, but the significance was borderline.

For each course, the differences for the low-level assignments between the two research groups are more distinct in the relative average scores than in the absolute average scores. This is because students who chose more high-level assignments were penalized in the absolute scores for choosing less low-level assignments.

To determine the variables affecting the average scores of the high-level assignments, a GLM (General Linear Model) procedure (Tabachnick & Fidell, 1996) was carried out for each course studied within the Project. The aim of the analysis was to test whether the research group, version of achievement test, or interaction between the test version and the research group could explain the results for the high-level assignments. For chemistry, $F = 37.5$; $p < .0001$, the only explaining

Table 3
*Average scores of low-level assignments by Project course and research group*

| Course | Research Group | $N$ | Low-Level Assignments | | |
|---|---|---|---|---|---|
| | | | $\bar{X}_{Absolute}$ | $t$ | $\bar{X}_{Relative}$ |
| Chemistry | Experimental | 59 | 62.9 | 3.26* | 80.1 |
| | Control | 38 | 50.4 | | 57.4 |
| Biology | Experimental | 81 | 67.5 | 1.93** | 96.4 |
| | Control | 65 | 62.8 | | 83.7 |

*$p < 0.01$.
**$p < 0.1$. This is considered as insignificant but borderline case.

Table 4
*Average scores of high-level assignments by project course and research group*

| Course | Research Group | $N$ | High-Level Assignments | | |
|---|---|---|---|---|---|
| | | | $\bar{X}_{Absolute}$ | $t$ | $\bar{X}_{Relative}$ |
| Chemistry | Experimental | 59 | 41.2 | 6.14* | 82.7 |
| | Control | 38 | 18.1 | | 63.5 |
| Biology | Experimental | 81 | 52.6 | 4.87* | 64.0 |
| | Control | 65 | 38.3 | | 55.7 |

*$p < 0.0001$.

variable was found to be the research group. For biology, both the research group, $F = 21.7$; $p < .0001$, and test version, $F = 51.1$; $p < .0001$ were the explaining variables.

Examinations in all the courses were constructed such that at least one high-level assignment was mandatory. Hence, a high relative score in the high-level assignments indicates that the student is capable of performing one or more high-level assignment. However, it does not necessarily mean that the student would perform high in a variety of high-level assignment types. High absolute score in the high-level assignments is an indication of a student's ability to apply a variety of thinking skills and justified his or her self-perception of being able to cope with such assignments.

Table 4 shows the absolute and relative average scores in high-level assignments by Project course and by research group. Here, significant differences were found in the absolute average scores in favor of the experimental group. In high-level assignments, the largest gap between experimental and control groups was found in chemistry. Comparing the relative to absolute average scores of the high-level assignments, as presented in Table 4, one can see that the differences in relative scores between the two research groups are less distinct.

### *Students' Choice of High-Level Assignments*

One index that may be related to students' ability to cope with high-level assignments is the relative number of such assignments that a student chose. Students from the two research groups behaved differently when presented with the option of choosing between low- and high-level assignments (knowledge type vs. higher-order thinking skills type).

Figure 1 shows that in chemistry, more experimental group students than control group students chose high-level assignments. This is true for each of the three high-level assignments. The assignment regarding posing a question and suggesting a solution is labeled in Figure 1 as ''Question posing.'' The ratio between the number of students in the experimental and control groups choosing this assignment was 2.2 ($p < .05$). The analogous ratio for the assignment regarding calling for the design of an experiment or conducting a research project (''Designing experiment'') was 4.8 ($p < .05$). Chosen by 84% of the experimental students and 62% of the control students, presentation to peers was the most popular assignment amongst both research groups. Question posing followed, with 40% experimental and 19% control students choosing this assignment. The least favorable assignments for both groups was designing an experiment or conducting a research project, which only 25% of the experimental and 5% of the control group students chose.

Choice of assignments in biology for the experimental and control groups is shown in Figure 2. The ratios between the number of students in the experimental and control groups who chose the four high-level assignments were 1.2 ($p < .05$) for the ''Relation to other topics''

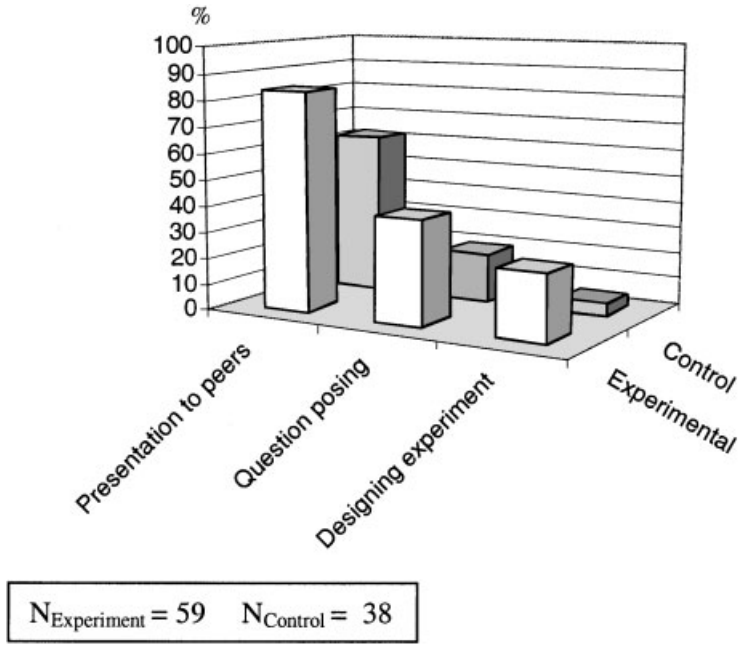


*Figure 1.* Percentage of chemistry students from the two research groups choosing high-level assignments.

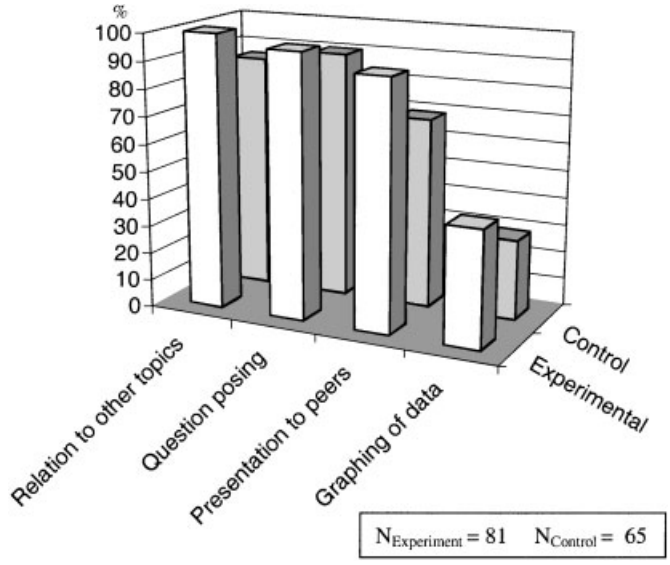


*Figure 2.* Percentage of biology students from the two research groups choosing high-level assignments.

assignment, 1.1 for ''Question posing,'' 1.3 ($p < .05$) for ''Presentation to peers,'' and 1.4 ($p < .05$) for the ''Graphing data'' assignment.

## Discussion and Further Implications

Gallagher et al. (1999) and Treagust et al. (2001) emphasized the importance of having assessment embedded in the learning process. They supported the notion that the integration of teaching with assessment leads to the improvement of learning in science. Such assessment is open-ended and sensitive enough to individual differences to reflect students' deep and broad understanding.

The research described in this article may serve as evidence that when assessment is integrated into the learning process and embedded in it, meaningful educational goals are achieved. Given an adequate school- and system-wide supporting environment, students develop higher-order thinking skills and their learning is more meaningful than the learning that takes place with traditional learning and assessment methods. Because the research was conducted on chemistry and biology courses, it is probably safe to claim that improvement is likely to occur in science courses. Moreover, when asked to express their attitudes toward the Project, students were in favor of extending the Project to additional courses despite their awareness of the increased demands on their time and intellectual effort (Dori, 2003). The teachers cooperated fully in applying the reform because they were involved in the Project from the outset by writing the proposal and developing the assessment methods. In discussing the alteration of power relationships between teachers and educational policy makers, Sarason (1990) used the definition of power as ''the ability to act or produce an effect'' (p. 51). He recommended that teachers be part of the educational decision process because they can really contribute. Enlarging teachers' role increases the likelihood of obtaining their commitment in implementing educational changes. In the Matriculation 2000 Project, teachers were deeply involved not only in teaching in a new setting but also in defining what the framework of the school-based alternative embedded assessment would be. Indeed, this was an important factor in the success of the Project. These findings and their implications are important to education policy administrators, the teaching community, and educational assessment professionals at a national level.

Black (1995a, 1998) argued that the potential of formative assessment is greater than that usually experienced in schools. In most cases, he claimed, formative assessment affects the learning in a positive way. Alternative embedded assessment has a significant formative element. Our findings are in line with Mitchell (1992), who pointed out the contribution of authentic assessment to learning processes and the advantages of engaging students, teachers, and schools in the assessment processes.

Traditionally, Israeli high school students are assessed at the end of 12th grade through a series of external matriculation examinations (counting for 50% of their matriculation transcript grade) and by their teachers (counting for the other 50%). The Project provided the experimental teachers with complete autonomy in determining 100% of the students' matriculation transcript grades in the pertinent courses. This grade was the culmination of a long 3-year process of alternative assessment that was embedded in and integrated into the teaching of these courses. Our findings show that overall, the experimental students achieved significantly higher scores than their control group peers on assignments that required knowledge as well as in assignments that required higher-order thinking skills. These differences between the two research groups were more significant and the gap was wider in the high-level assignment scores. This is a strong indication that the Project has indeed attained one of its major objectives: namely, fostering higher-order thinking skills. This is in line with Treagust et al. (2001), who reported a case

of a middle school teacher who used embedded assessment to endow her students with deep understanding of science while fulfilling their expectations of how they felt they best learned.

Most Israeli high school students who major in chemistry are assessed externally through paper and pencil examinations. Because of concerns regarding hazards and lack of appropriate safe experiments and resources, only a few schools carry out chemistry laboratory examinations. Biology examinations differ in that in addition to the paper and pencil examination, they comprise a practical laboratory test and an oral test involving a biological/ecological project, which is done individually or in pairs (Tamir, 1998). The difference in variability of testing methods between chemistry and biology may explain why the gap between the experimental and control groups in chemistry was larger than the corresponding gap in biology.

The students' ability to cope with high-level assignments was analyzed, in part, as the number of such assignments a student chose relative to the low-level assignments. When presented with the option of choosing between assignments requiring knowledge vs. ones requiring higher-order thinking skills, the experimental group students chose more high-level assignments than the control group students. Presentation to peers was found to be a popular selection for both chemistry and biology courses. This indicates that within the experimental group the Project has encouraged teamwork and knowledge sharing among peers.

In the two courses, the average scores of the high-level assignments were lower than those of the low-level assignments. This is consistent with finding of other researchers (Harlen, 1990; Lawrenz et al., 2001) who showed that open-ended assignments are more difficult and demanding because they measure more higher-order thinking skills and because students are required to formulate original responses. Open-ended assignments, which are categorized in this report as high-level ones, provide important feedback that is fundamentally different in nature from what can be obtained from assignments that are defined as low-level ones. This particular finding is contrasted with what Lawrenz et al. (2001) found. Their results indicated that written open-ended items appear to provide less unique information compared with multiple choice questions. A possible explanation for this discrepancy is the combination of differences in the age group of the research population (9th grade in Lawrenz's study as opposed to 12th grade in our study) and the academic level at which the course is taught (basic level vs. honors class). The high-level assignments, developed as research instruments for this study, required a variety of higher-order thinking skills and can therefore serve as a unique diagnostic tool.

Following the recommendations of Gitomer and Duschl (1998) and Treagust et al. (2001), this study has shown that the Matriculation 2000 Project improved learning outcomes and shaped curriculum and instruction decisions at the school and classroom level through changing the assessment culture. The reform that took place in the 22 high schools is a prelude to a transition from a nationwide standardized testing system to a school-based alternative embedded assessment system. Moreover, teachers, principals, superintendents, and Ministry of Education officials, who were engaged in this Project, became involved in convincing others to extend the Project boundaries either to additional courses at the same school or to additional schools in the same district. Relevant stakeholders in the Israeli Ministry of Education recognized the significance of the research findings. However, a combination of changes in the political system, which puts more emphasis on values and culture at the expense of science, and the economical downturn has caused a delay in the extension of the Project. Meanwhile, as a result of this study the chemistry assessment is undergoing a change at the national level. Similar to biology, in addition to the paper and pencil examination, chemistry assessment is going to include a practical laboratory test and an unseen case study in the spirit of the tests developed in this research (Dori, Sasson, Kaberman, & Herscovitz, 2002).

Additional changes have taken place in science teacher education and professional development programs. Alternative and embedded assessment methods are implemented in several teaching methods courses and in-service trainings at the Technion. Annually, some 20 preservice science teachers graduate at the Technion's Department of Education in Technology and Science, whereas about 30 in-service teachers take professional development summer courses. Based on the findings of this research and the literature review, these current and future teachers are being taught to apply these methods; at the same time, they are being assessed using these alternative methods. Similar activities for in-service teachers take place at other academic institutions in Israel (Hofstein & Even, 2001; Eylon & Bagno, 1997).

### *Research Limitations*

Fourth Generation Evaluation (Guba & Lincoln, 1989) maintains that educational research findings should not be regarded as objective truths, but rather as a dynamic picture of what occurs in classrooms, schools, and larger systems related to education. Although not involved in the Project and considered an external academic evaluator, this author is indeed a stakeholder in the study described in this report. As an educator, this reform fits my system of beliefs and as much as I tried to set my opinions aside, the interpretations of the data might be influenced by these biases.

The two experimental schools that participated in the research were the only ones of the 22 schools where science was taught. A total of 6 of the 22 schools were investigated concerning the student attitudes (Dori, 2003), but this is beyond the scope of this article. It is worth repeating that the 22 schools do not constitute a random sample of the high schools in Israel. Rather, they can be viewed as exemplary schools and therefore the conclusion should be considered with this in mind. In the long term, investigation of the conditions and impacts of such a project on settings of constrained communities is needed to shed valuable light on how innovative programs falter in less than ideal settings.

Finally, one should bear in mind that even though the reform's goal was to change the assessment practices, in effect the project inevitably influenced not just the assessment but also the teaching in the experimental schools.

### *Further Research and Recommendation*

The research findings indicate that given adequate support in funding additional teachers' time and academic support for their professional development, as well as teachers' consent and collaboration, schools can transfer from nationwide or statewide standardized testing to school-based alternative embedded assessment. This study indicates that well-designed, challenging curriculum, which includes a variety of formative and summative assessment practices, may provide opportunities for students to engage in higher-order reasoning. These students can perform better in tasks congruent with the teaching strategies than students who have not engaged in such rich learning and assessment practices.

A prerequisite to concluding that the Project is a complete success and should be implemented nationwide is a larger-scale research effort, which will include all the high schools in a sample of districts. The objective of such research would be to validate the findings of the current research and examine their applicability on a large scale. In view of the fact that the schools that took part in the Project are not representative of the entire high school population in the country, the competence of an average school team to apply alternative embedded assessment successfully while maintaining an unbiased grading system should be examined.

In parallel, interim solutions can be contemplated. As the structure of the standard biology Matriculation examination at the national level has shown, improvement in the variety of assessment methods is not necessarily linked to the fact that the assessment is done locally. Following the biology example, incorporating additional testing means can be augmented to chemistry. Independently, schools can transfer to the new alternative embedded assessment method in a graduated, continuous mode, such that each year they would be authorized to increase the weight of the school-based assessment.

This research was funded by the Chief Scientist of the Israeli Ministry of Education. I would like to thank Nitza Barnea and Tzvia Kaberman for their contribution to the research described in this paper and Lori Breslow for her valuable comments on the draft of this paper.

## References

Baker, E.L. & Herman, L.J. (1983). Task structure design: Beyond linkage. Journal of Educational Measurement, 20, 149–164.

Ben-Elyahu, S. (1995). Summary of the feedback questionnaire of ''Matriculation 2000'' Project first year. Pedagogical Secretariat, Ministry of Education, Jerusalem, Israel [in Hebrew].

Ben-Peretz, M. (1994). Report of the Committee for Examining the Format of Israeli Matriculation Examination, Ministry of Education, Jerusalem, Israel [in Hebrew].

Bexter, G.P., Shavelson, R.J., Goldman, S.R., & Pine, J. (1992). Evaluation of procedure-based scoring for hands-on science assessment. Journal of Educational Measurement, 29, 1–17.

Birenbaum M. & Shaw, D.J. (1985). Task specification chart: A key to a better understanding of test results. Journal of Educational Measurement, 19, 259–266.

Black, P. (1995a). Assessment and feedback in science education. Studies in Educational Evaluation, 21, 257–279.

Black, P. (1995b). Curriculum and assessment in science education: The policy interface. International Journal of Science Education, 7, 453–469.

Black, P. (1998). Assessment by teachers and improvement of students' learning. In Fraser, B.J. & Tobin, K.G. (Eds.), International handbook of science education (pp. 811–822). Dordrecht, The Netherlands: Kluwer Academic.

Costa, A.L. (1985). Teacher behaviors that enable student thinking. In Costa, A.L. (Ed.), Developing minds: A resource book for teaching thinking. Alexandria, VA: Association for Supervision and Curriculum Development.

Dillon, J.T. (1990). The practice of questioning. London: Routledge.

Dori, Y.J. (2003). A framework for project-based assessment in science education. In Segers, M.S.R., Dochy, F.J.R.C., & Cascaallar, E.C. (Eds.), Optimizing new modes of assessment: In search of qualities and standards (Chap. 5). Dordrecht, The Netherlands: Kluwer. In press.

Dori, Y.J., Barnea, N., & Kaberman, T. (1999). Assessment of 22 high schools in the ''BAGRUT 2000'' (Matriculation 2000) Project. Research report for the Chief Scientist, Israeli Ministry of Education, Department of Education in Technology and Science, Technion, Haifa, Israel [in Hebrew].

Dori, Y.J. & Herscovitz, O. (1999). Question posing capability as an alternative evaluation method: Analysis of an environmental case study. Journal of Research in Science Teaching, 36, 411–430.

Dori, Y.J., Sasson, I., Kaberman, T., & Herscovitz, O. (2002, August). Computerized chemistry laboratory. Paper presented at the 224th American Chemical Society National Meeting, Boston, MA.

Dori, Y.J. & Tal, R.T. (2000). Formal and informal collaborative projects: Engaging in industry with environmental awareness. Science Education, 84, 1–19.

Eylon, B. & Bagno, E. (1997). Professional development of physics teachers through long-term in-service program: The Israeli experience. In The changing role of physics department in modern universities: Proc. ICUPE—International Conference of University Physics Education.

Gallagher, J.J., Parker, J., & Ngwenya, J. (1999). Embedded assessment and reform in science teaching and learning. East Lansing, MI: Michigan State University.

Gallard, A.J., Viggiano, E., Graham, S., Stewart, G., & Vigiliano, M. (1998). The learning of vouluntary and involuntary minorities in science classrooms. In Fraser, B.J. & Tobin, K.G. (Eds.), International handbook of science education (pp. 941–953). Dordrecht, The Netherlands: Kluwer Academic.

Frederiksen, N. (1984). The real test bias: Influences of testing on teaching and learning. American Psychologist, 39, 193–202.

Gitomer & Duschl, R.A. (1998). Emerging issues and practices in science assessment. In Fraser, B.J. & Tobin, K.G. (Eds.), International handbook of science education (pp. 791–810). Dordrecht, The Netherlands: Kluwer Academic.

Guba, E.G. & Lincoln, Y.S. (1989). Fourth generation evaluation. London: Edward Arnola.

Herried, C.F. (1994). Case studies in science: A novel method of science education. Journal of College Science Teaching, 23, 221–229.

Herried, C.F. (1997). What is a case? Bringing to science education the established tool of law and medicine. Journal of College Science Teaching, 27, 92–95.

Hofstein, A. & Even, R. (2001). Developing chemistry and mathematics teacher leaders in Israel. In Nesbit, C.K., Wallace, J.D., Pugalee, D.K., Miller, A.C., & DiBiase, W.J. (Eds.), Developing teacher leaders in science and mathematics: Professional development in science and mathematics (pp. 189–208). Columbus, OH: ERIC Clearinghouse for Science, Mathematics and Environment Education.

Klein, S.P., Jovanovic, J., Stecher, B.M., MacCaffrey, D., Shavelson, R.J., Haertel, E., Solano-Flores, G., & Comfort, K. (1997). Gender and racial/ethnic differences on performances assessments in science. Educational Evaluation and Policy Analysis, 19, 83–97.

Kleinsasser, A., Horsch, E., & Tastad, S. (1993, April). Walking the talk: Moving from a testing culture to an assessment culture. Paper presented at the annual meeting of the American Educational Research Association. Atlanta, GA.

Kremer, B.K. & Walberg, H.J. (1981). A synthesis of social and psychological influences on science learning. Science Education, 65, 11–23.

Lawrenz, F., Huffman, D., & Welch, W. (2001). The science achievement of various subgroups on alternative assessment formats. Science Education, 85, 279–290.

Lewy, A. (1996). Postmodernism in the field of achievement testing. Studies in Educational Evaluation, 22, 223–244.

Linn, R.L. (1983). Testing and instruction: Links and distinctions. Journal of Educational Measurement, 20, 180–189.

Madaus, G.F. & Kellaghan, T. (1993). The British experience with ''authentic'' testing. Phi Delta Kappa, 74, 458–469.

Mitchell, R. (1992). Testing for learning: How new approaches to learning can improve American schools. New York: Free Press.

Nevo, D. (1995). School-based evaluation: A dialogue for school improvement. Oxford, UK: Elsevier Science.

National Research Council (1996). National science education standards. Washington, DC: National Academy Press.

Oliver, J.S. & Simpson, R.D. (1988). Influences of attitude toward science, achievement motivation, and science self conception on achievement in science: A longitudinal study. Science Education, 72, 143–155.

Resnick, L.B. (1987). Education and learning to think. Washington, DC: National Academy Press.

Resnick, L.B. & Resnick, D.P. (1992). Assessing the thinking curriculum: New tools for educational reform. In Gifford, B.R. & O'Connor, M.C. (Eds.), Changing assessments: Alternative views of aptitude, achievement and instruction (pp. 37–75). Boston: Kluwer Academic.

Ruiz-Primo, M.A. & Shavelson, R.J. (1996). Rhetoric and reality in science performance assessment: An update. Journal of Research in Science Teaching, 33, 1045–1063.

Sarason, S.B. (1990). The predictable failure of educational reform: Can we change course before it's too late? San Francisco: Jossey-Bass.

Shepardson, D.P. & Pizzini, E.L. (1991). Questioning levels of junior high school science textbooks and their implications for learning textual information. Science Education, 75, 673–682.

Shorrocks-Taylor, D. & Jenkins, E.W. (2000). Learning from others. Dordrecht, The Netherlands: Kluwer Academic.

Sweeney, A.E. & Tobin, K. (Eds.). (2000). Language, discourse, and learning in science: Improving professional practice through action research. Tallahassee, FL: SERVE.

Tamir, P. (1998). Assessment and evaluation in science education: Opportunities to learn and outcomes. In Fraser, B.J. & Tobin K.G. (Eds.), International handbook of science education (pp. 761–789). Dordrecht, The Netherlands: Kluwer Academic.

Tabachnick, B.G. & Fidell, L.S. (1996). Using multivariate statistics (3rd ed.) New York: Harper Collins.

Treagust, D.F., Jacobowitz, R., Gallagher, J.J., & Parker, J. (2001). Using assessment as a guide in teaching for understanding: A case study of a middle school science class learning about sound. Science Education, 85, 137–157.

Wiggins, G. (1989). A true test: Toward more authentic and equitable assessment. Phi Delta Kappan, 70, 703–713.

Wolf, D., Bixby, J., Glenn, J. III, & Gardner, H. (1991). To use their mind well: Investigating new forms of student assessment. Review of Research in Education, 17, 31–73.