Normative Evaluation of Mathematics, Science and Technology Curricula
The case of the modeling curriculum

Ibrahim A. Halloun
Science & Mathematics Education Department, Lebanese University, Lebanon
halloun@inco.com.lb; http://www.inco.com/halloun

Evaluation of a given curriculum is not an end by itself, and it does not mark the end of curriculum development. It is an ongoing process carried out throughout curriculum development and implementation for continuous regulation. Regulation at any point may extend from minor refinements of some curriculum components to a radical reform of the curriculum on entirely new grounds. For reliable outcomes and sustainable impact, curriculum evaluation needs to be normative. Normative evaluation is conducted according to well-defined taxonomy of conceptions, processes and trends that students are expected to develop under the given curriculum. It comes out with quantitative indicators ascertaining in terms of pre-established criteria to what extent students have actually met original expectations. Similar taxonomy and criteria are also set for teacher practice in the classroom, the impact of which is assessed on student achievement. Normative evaluation is illustrated in the case of the modeling curriculum in secondary school physics. Data are presented and discussed in the context of Newtonian mechanics courses so as to illustrate with practical examples how normative curriculum evaluation may be conducted in a way to contribute to sustainable reform and meaningful and equitable learning in mathematics, science and technology.

Curriculum development in mathematics, science or technology (MST) is an involved process that goes beyond the conception of various curriculum components (pedagogical framework, objectives, discipline program, means and modes of learning and instruction, assessment, all along with teacher training programs and various support systems). The process extends to the evaluation of individual components in order not only to determine their value, but, more importantly, the extent to which they need to be regulated, i.e., refined or even reconceived altogether. As such, evaluation of a given curriculum is not an end by itself, and it does not end as long as the curriculum is being implemented. A curriculum is continuously being put to the test not only during its conception phase but also throughout its implementation. This is true for any educational field. It is more so for mathematics, science and technology education (MSTE) in an era where virtually all aspects of human life are dominated by these fields.

Viability of pedagogical framework and feasibility of objectives are determined when these components are deployed in setting out and implementing practical aspects of the curriculum, mainly programs, means and modes of learning and instruction (assessment included). The same is true for teacher training and support systems. Educational programs, and means and modes of learning and instruction are evaluated, in their turn, through their impact on student learning. More specifically, evaluation of student learning in relation to teacher practice in the classroom provides significant indices of the overall curriculum merits and reliable indicators of the direction in which refinement or reform (regulation) should be headed for various curriculum aspects. That is why we concentrate our discussion in this paper on the interplay between learning and instruction in the educational enterprise. Our discussion is illustrated with practical examples and actual data on how to evaluate student learning in relation to teacher practice and to deploy the outcomes in curriculum regulation, all in the context of modern educational theory.

Around the turn of the twentieth century, many a country around the world has embarked on curriculum reform in MST at the pre-college levels. Reform movements have been sparked by the colossal developments in science and technology, but most importantly by research in MSTE that has been constantly showing that existing curricula are not promoting meaningful understanding of MST in student minds. Local and international organizations that were behind the most influential reform movements have come to a certain common understanding on the direction in which reform should be headed for significant and sustainable impact. Some of these common grounds highlight the importance of *continuous* curriculum evaluation, and the necessity to conduct *normative* evaluation, i.e., evaluation in terms of clearly defined, "reasonable" standards, and to ensure that *all* students meet in this respect a "minimal set of standards" (Bibeau, Halloun, May & Reddy, 2002).

This paper discusses how such an evaluation may be conducted and what it may lead to when put to practice. The paper comes in six sections. In the first section, general procedures are outlined for the development of MST curricula. Discussion concentrates on the fact that evaluation should be meant for continuous curriculum regulation. The entire process would be carried out within the framework of a conveniently chosen educational theory, and this in a way to benefit from potentials and account for constraints that one may faces when things are put to practice in schools. Modeling theory, a synopsis of which is presented in the second section of the paper, is advocated as a viable theory for MST. Materials from a physics curriculum developed within the framework of modeling theory serve for illustration in subsequent sections. In the third section is a discussion of the kind of evaluation that MST curriculum developers need to undertake in order to tell whether things are working in the right direction, i.e., whether they are actually contributing to meaningful, equitable and sustainable reform. This is normative evaluation whereby the value of MST curricula is ascertained with particular instruments developed following pre-established, and conveniently graded, taxonomy and criteria. Taxonomy and criteria pertain to "reasonable" aspects of respectively what instructional modes teachers can follow in the classroom and what conceptions, processes, and trends (habits, attitudes and views) about knowing and learning MST students may come out with. Procedures of normative evaluation as implemented in the evaluation of a modeling curriculum are outlined in the fourth section of the paper. Specific outcomes are presented and discussed in the following section so as to illustrate how one can go about evaluating MST materials in a way to regulate a new curriculum and steer it in the right direction. Implications for sustainable reform are then discussed in the last section of the paper.

1. Curriculum development: A basic modus operandi

The state of education, and especially of MST, has reached such an alarming level by the end of the past century that radical reforms in educational policy and curricula have been called for around the world. Actors in education and those mainly in science and technology (S&T) kept heading in virtually opposite directions for the best part of the twentieth century. Educators kept relying until late that century on rules of thumb for conceiving and implementing various educational curricula. Meanwhile, scientists and engineers were developing S&T following systematic norms and rules stemming from well-defined and corroborated scientific theory. Action-research was also becoming the norm in S&T but not in education. This is evaluative research intended to continuously enhance productivity (as well as equity) in many respects. Educators have finally come to realize that, in order to turn things around in MSTE, and education in general, they need to gain insights from successful practice in scientific research and technological developments. In particular, rose the conviction that for meaningful and sustainable reform, MSTE should be *theory-laden* and steered by *action-research*.

In the first respect, educators became more and more convinced (gradually, through practice) that development of MST curricula, or of any other discipline to that matter, has to be grounded in a well-defined educational theory. Educational theory sets the general framework (i.e., pedagogical underpinnings, norms and guidelines) for making decisions about various aspects of a given curriculum. This includes, in the case of MST, the choice of whether to conceive mathematics, science and technology in integrated or separate curricula. Such a choice, like the choice of curriculum objectives, is though governed by a number of factors that may be beyond the scope of the chosen theory and/or the control of curriculum developers. These factors include national and local policies about MST, and especially the feasibility of implementing a particular approach or targeting a specific objective in a given society. Other theory-laden decisions pertain to the kind of profile students should develop under a given curriculum, and to the ways they should go about developing such a profile (i.e., they pertain to *what* and *how* should students learn).

A *student profile* includes all sorts of *conceptions* (including theories, models and concepts) and *processes* (including related tools, schemes and rules) that any student willing and capable of investing necessary efforts should develop following the completion of a given curriculum. As a consequence, student general *trends* should evolve in specific directions. Under general "trends" we hereby include things like *habits* (including learning styles), *attitudes* toward MST and MSTE, and *views* (including views about the nature of MST). Components of a student profile from the perspective of modeling theory are presented in the next section along with a synopsis of the theory in question.

As mentioned above, curriculum developers have not only to ground their work in a well-defined educational theory. They also have to take into account realities on the ground that affect their work (Figure 1). Concerned actors need to consider these realities in their *strengths* and *weaknesses*, take advantage of the *opportunities* they provide and circumvent *threats* they may impose. This is often referred to as the *SWOT* rule in business and industry; the rule applies as much to education as to any other enterprise. The *SWOT* rule applies to all *stakeholders* in the educational enterprise. These include policymakers, curriculum experts, developers of learning material, publishers, teachers, students, administrators, parents, and other concerned groups and communities. The rule also applies to *systemic structures* that are involved in the process and that include support systems and programs, and to logistics of all levels for developing, field-testing and implementing curriculum components at various scales, and subsequently refining (or reconceiving) such components.

Teachers have traditionally been the most marginalized stakeholders in the development of curricula and even of learning materials. They have often been treated as mere traders of finished products or as passive reporters about, or disseminators of, canned information and goods. Instead, teachers should take an active role at different stages of curriculum and materials development. Perhaps the most important role teachers can and should play in this respect is in *action-research*. As such, teachers should consider themselves, as individuals and groups, as most concerned about *continuous evaluation* of their curriculum and its *regulation* (refinement or even reconception) so that

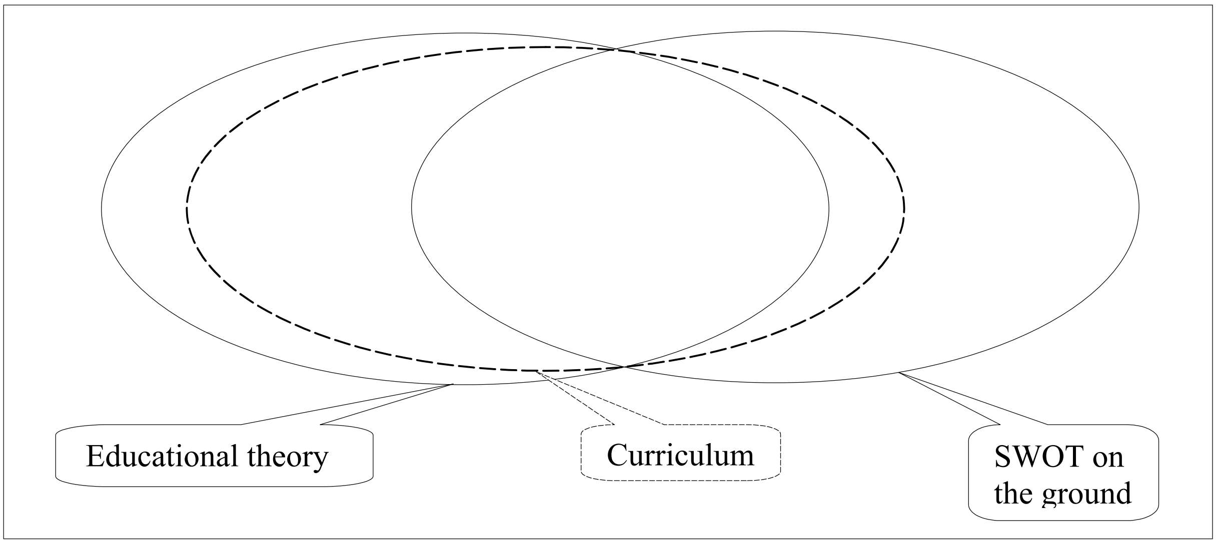

Figure 1. Interplay between educational theory and the SWOT rule in curriculum development.

Curriculum developers cannot adhere entirely to their educational theory and account for all aspects (mostly constraints) on the ground meant in the *SWOT* rule. They need to compromise between the two, but in a manner that favors more their theory than ground constraints.

they ensure *meaningful* and *equitable* learning experience to all participating students, and thus *sustainable* curriculum changes and/or reform. For action-research to succeed in this direction, systemic structures should be conveniently put in place in order to support teachers throughout their endeavor.

Curriculum evaluation is thus not an end by itself, and it does not mark the end of a process (curriculum development). Such an evaluation is meant to *gauge* and direct a curriculum in the direction of meaningful and equitable learning, and of sustainable reform. It is a continuous process that lasts as long as the curriculum is being implemented. Furthermore, the evaluation in question is not confined theoretically to the theoretical framework of the curriculum, and is not restricted in practice to a limited group of outside evaluators nominated by certain authority. Evaluation norms and criteria need to account for *SWOT* on the ground, though not necessarily to the same extent as to the originally adapted theoretical framework (Fig. 1). Evaluating bodies should span from external evaluators not originally involved in the conception of the curriculum to various stakeholders in the process and especially concerned teachers. Because they are the most accessible and testable components of a curriculum, and because they are supposed to concretize other curriculum components, means and modes of learning and instruction (assessment included) are paid a special attention in the *evaluation for regulation* process. Their evaluation should answer at least the following *triad* of *A*-questions:

1. What means and modes have actually been made *available*, and how do they fair between original intentions of curriculum developers (as stemming from the adapted educational theory), on the one hand, and realities on the ground (SWOT) on the other?

2. What is then *anticipated*, of teachers and of students, to accomplish?

3. What has actually been *achieved* on the ground? Who and what made that possible? How do we know?

2. Modeling theory in science education

Modeling theory is a theory for science education that this author has been working on for the last two decades (Halloun, in press). The theory that can be readily deployed into mathematics and technology education has been originally conceived for, and deployed in the development of, physics curricula at the secondary school and college (university) levels. Its recent implementation in other scientific disciplines and at lower educational levels suggests its viability as a generic theory for MSTE. Modeling theory is the object of a companion paper at this SAARMSTE conference (Halloun, 2004). Below is a quick rundown of some aspects of the theory that bear directly on the current paper.

Modeling theory is grounded in the philosophy of science and cognitive and educational research. It calls for science courses to empower students to achieve a *paradigmatic evolution* whereby they transcend their naïve realism (or common sense) and evolve into the realm of science. The evolution is *student- centered*, *teacher-mediated*. It is promoted in a learning environment that is structured enough (by the teacher) to be in line with scientific theory and inquiry, yet flexible enough to account for differences in students' initial knowledge state, and to allow for their self-evaluation and self-regulation.

A person's thought and behavior in a given situation are always governed by a particular paradigm or set of paradigms that constitute a paradigmatic profile. A *paradigm* is, for us, an intellectual system that: (a) determines the conditions of a person's conscious experience, and that (b) sets forth standards, rules and guidelines that govern the choice and the course of all actions necessary for the reification and continuous evaluation of the experience in question. A *paradigmatic profile* of an ordinary person consists of a set of paradigms, of the same scope or of closely related scopes, some of which may be in accordance with scientific paradigms while others at odds with such paradigms.

Educational research shows that paradigmatic profiles that secondary school and university students deploy for dealing with real world systems and phenomena are often dominated by *naïve realism,* and are thus at odds with science. Research also shows that student paradigms are not significantly affected by conventional instruction of lecture and demonstration. Many science educators have argued that this situation cannot be reversed unless science teachers bring into their classrooms whatever inquiry practices scientists resort to in their research, and especially *modeling* practices. Modeling theory amplifies this argument and calls for engaging students in insightful modeling inquiry that help them regulate their paradigmatic profiles in a way that significantly tunes down naïve realism in favor of scientific realism. A special attention is paid in the process for students to develop any scientific theory that is the object of a science course around a set of basic models.

A *scientific theory* consists, for us, of: (a) a set of models or families of models, and (b) a set of particular rules and theoretical statements that govern model construction and deployment and that relate models to one another and to specific patterns in the real world. A *scientific model* is a conceptual system mapped onto a specific *pattern* in the structure and/or behavior of some physical systems so as to allow us: (a) describe, explain, and predict (or postdict) the pattern in question, and, eventually (b) control or change physical realities exhibiting the pattern, and (c) reify the pattern in new realities. Among the models of a theory is a subset of basic models the understanding of which is critical for meaningful understanding of the theory in question. A *basic model* is one that is simple enough to facilitate student understanding of fundamental tenets and conceptions (concepts, laws, etc.) of the respective theory, and development of fundamental tools, skills and habits of scientific inquiry. Yet it is generic enough to serve in the construction of more complex models in the theory.

Modeling processes are pivotal in scientific inquiry. These processes include: (a) model construction and validation in the context of real world situations, and (b) model adduction and analysis for solving empirical or rational problems. In their endeavor, scientists rely, often implicitly, on various tools, the most important of which are modeling schemata.

From a scientific perspective, a *modeling schema* serves, on the one hand, as an organizational template for structuring models or related conceptions in a compact and coherent way. It offers, on the other, a set of well-defined rules for evaluating and deploying scientific models or their building blocks. From a pedagogical perspective, a modeling schema is indispensable for students to retrace scientists' steps meaningfully and efficiently. It also provides teachers with reliable means for planning instruction, and for assessing student learning and teaching practice.

Under modeling instruction, a student is thus supposed to develop a *profile* with the following characteristics:

- *Conceptions* consisting, in a given science course, of a particular scientific theory organized, in a middle-out structure, around a set of basic models in the context of which are developed all lower-level conceptions (laws, concepts, etc.).
- *Processes* of scientific inquiry, especially including *schemes* of model construction and deployment and of scientific discourse, along with associated *tools* (modeling schemata and various other tools) and *rules* (including tools' semantics and syntax).
- *Learning styles*, especially including processes of reflective thinking that help students regulate their own paradigmatic profiles in an insightful and meaningful way.
- *Emergent trends* (habits, attitudes and views) that are characteristic of the scientific community, especially habits and attitudes commonly referred to nowadays in calls for scientific literacy, and respective views pertaining to the nature and relevance of science (math and technology).

Students develop their profiles in five-phase *learning cycles* (exploration, model adduction, model formulation, model deployment, paradigmatic synthesis). Each cycle is devoted to the development of a specific model, along with its conceptual and practical requirements. All along a cycle, individual students ascertain their ideas in light of empirical and rational evidence at their disposal. They often negotiate these ideas with their peers under the *mediation* of the instructor, and this in a way that helps them detect and resolve any incommensurability with science that might emerge in the process.

In its most efficient form, the modeling physics curriculum is implemented at the secondary school and college levels in small classes of about thirty students each (scale-up projects are currently in place in some schools and universities). Students always work collaboratively in groups of three students each. The typical classroom includes eight to ten workstations. Each group is assigned a particular workstation where most, if not all, learning activities take place. A workstation is equipped with necessary laboratory equipment, preferably of MBL type (Microcomputer-Based-Laboratory). In the absence of appropriate modeling software that has not been developed yet, MBL can enhance learning logistics (reduce efforts and time invested by teachers and students), but not necessary the quality of conceptual understanding. Thus, in the current status of things, traditional laboratory equipment and most MBL may bring students up to the same quality of learning (i.e., the same level of paradigmatic evolution), but MBL does so more efficiently than the other.

Learning activities are designed with student participation. Manuals of modeling activities are also available for secondary school courses*. A manual is divided into units, each devoted to the construction and deployment of a specific model throughout a single learning cycle. Students actively construct a given model through appropriate experiments conducted on their workstations and/or outside the classroom in real world settings. Then they deploy it in various activities including traditional paper-and-pencil problems. All this takes place under teacher *mediation* in the manner described in the companion paper (Halloun, 2004).

3. Norms and guidelines of normative evaluation

Evaluation of MST curricula needs to be a *normative* process in order to contribute to meaningful and sustainable reform. As such, answers to the first question about available means and modes of learning and instruction (in the A-triad of § 1) must come by comparison to clearly defined norms and standards that are set in terms of the chosen educational theory and the SWOT rule on the ground. Similar norms and standards should be set in answer to the second triad question, all this *before* a curriculum is implemented and learning materials and instructional modes are developed, so that the third question be answered explicitly and objectively in terms of these pre-established benchmarks. In other words, it is imperative that benchmarks be clearly identified ahead of time, as a curriculum is being conceived and before practical aspects get developed and used, so that the evaluation process does not get biased by what goes on on the ground, especially in the classroom. For *evidence* (or counterevidence) to be reliable, it must be *criterial*, i.e., established against *objective criteria* that are not tainted by the special interests and idiosyncrasies of any of the stakeholders, especially end-of-the-line material users (teachers, students, administrators).

Two major steps need to be taken in normative evaluation before criterial evidence can be sought for MST curricula:

1. Establish a detailed *taxonomy* of conceptions, processes and trends (habits, attitudes and views) that would make up the profile that students are anticipated to develop following the completion of a curriculum.
2. Set *criteria* that establish whether individual students have actually developed each element of the anticipated profile, and to what extent they have done so.

* To inquire about instructional materials for modeling in physics, please visit: http://modeling.asu.edu/

In parallel, similar taxonomy and criteria need to be set for teachers' practice in the classroom in order to subsequently assess whether teachers have actually done everything that is necessary for students to develop the anticipated profile. Curriculum developers need then to specify which items in student profile each mean or mode of learning and instruction is supposed to foster and to what level. Furthermore, and in order for the evaluation process to focus on realistic aspects and result in meaningful outcomes, it is better that taxonomy and criteria be *graded*, i.e., that they be about ordered levels of student competency. At least two levels or *thresholds* need to be identified in this respect:

1. *Basic threshold*. This is the most fundamental level. It corresponds to the *minimum standards* of meaningful understanding that *any* student should meet, irrespective of the initial competence level and interests of the student.

2. *Mastery* or *critical threshold*. This is the highest threshold that students need to cross in order to master all fundamental conceptions and processes in a given MST course. In an ideal and truly equitable situation, all students willing to invest necessary efforts should be capable of reaching this threshold. In traditional classroom settings, critical threshold somewhat corresponds to the level of understanding of top grade students.

Students' failure to reach the basic threshold would be a *doom indicator* for the curriculum being evaluated (provided that means and modes have originally been considered as suitable, and that they have actually been deployed as anticipated). In science courses, this threshold corresponds to conceptions and processes that are most elementary in the scientific theory that is the object of a given course, and that serve in the construction of the most elementary models of the theory. For example, in classical (Newtonian) mechanics, and according to modeling theory, basic threshold corresponds to concepts and principles needed for the construction of the free particle model and the uniformly accelerated particle model. The first model corresponds to physical objects at rest or in linear uniform translation under no net force in a given inertial reference system. The second model corresponds to physical objects that are in linear or parabolic translation under a net constant force in such a reference system. Critical threshold in classical mechanics corresponds to conceptions and processes that are indispensable for the construction and deployment of the two models just mentioned, as well as of at least the model of a particle in uniform circular motion and the harmonic oscillator model. The four models are part of the set of *basic models* in Newtonian theory (Halloun, 2004).

Answers to all three questions in the A-triad (§ 1) are usually sought by internal and external panels of experts. *Internal panels* consist of people who have been directly involved with curriculum development and implementation, teachers, authors and publishers included. *External panels* consist of outside people who have not been involved in the process in any way. It is preferable that internal and external panels set their evaluation strategies separately and independently of one another before they get together to coordinate their efforts. It is even preferable, if the situation allows it, that each panel conducts its own evaluation on the ground independently of the other panel. In the end, the two panels would get together to compare and analyze their outcomes. In the entire process, both types of panels need to assess things in terms of the same taxonomy and criteria originally set by curriculum developers. However, by the end of the evaluation process, they may recommend some changes in these respects.

Evaluation panels usually rely entirely on their own judgment, based on qualitative evaluation, and without recourse to quantitative indicators, in order to answer the first two questions of the triad, the ones about the nature of available materials and anticipation. As such, panel judgment may be reliable and sufficient with respect to these two questions, provided that evaluation is made in terms of pre-established taxonomy and criteria. However, qualitative evaluation provides only circumstantial (and often anecdotal) evidence when it comes to achievement as addressed in the third question of the triad. Reliable evidence (or counterevidence) about achievement can only be sought through the use of appropriate instruments that measure teacher practice and student profile evolution as quantitatively as possible, and that provide outcomes that can be subject to all sorts of inferential statistics, including correlation between teacher practice and student achievement. A battery of instruments described below is often needed in this respect so that various aspects may be adequately assessed of student profile as well as of teacher practice. Outcomes would not only answer questions about teacher practice and student achievement (independently and in relation to one another). They would most importantly serve to reconsider various aspects of the curriculum, and thus refine (or even overhaul) the curriculum accordingly, and revisit the A-triad of questions with refined taxonomy and criteria. In this respect, the triad of questions constitutes a *closed cycle* that evaluators and other stakeholders need to go through periodically, over and over again, as long as the curriculum under evaluation is in effect (Figure 2).

The battery of evaluation instruments includes:

1. For assessing teacher practice and viewpoints:

- *Surveys*, including paper-and-pencil surveys and interviews with teachers implementing the curriculum under evaluation.
- *Observation grids* prepared in accordance with the taxonomy and criteria set for teacher practice. Evaluators visit classes armed with the grids in which they note their observations about what takes place in the classroom. Class visits may sometimes be supplemented or even substituted with video-tapes of class sessions.
- *Teaching logs*, whereby teachers record regularly and systematically their own activities and comments about their own practice.

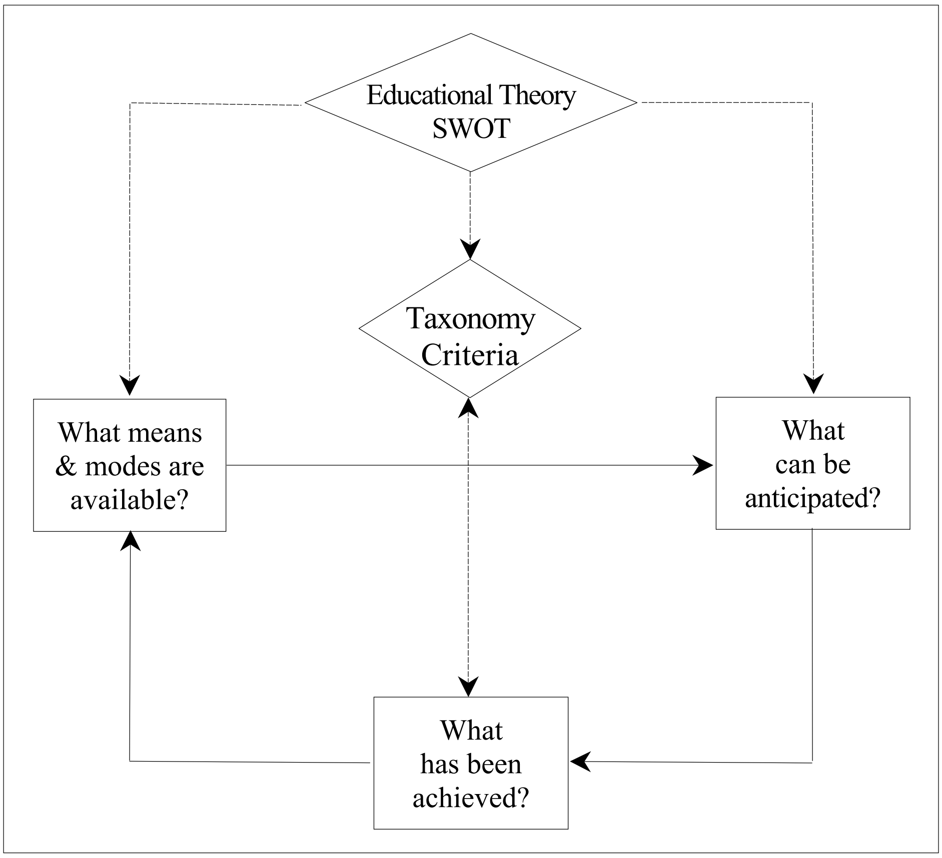

Figure 2. The triad of evaluation questions constitutes a closed cycle that evaluators keep going through periodically as long as the curriculum under study is in effect.

Educational theory and the SWOT rule are accounted for as means and modes of learning and instruction are being developed and in setting all sorts of expectations about teachers and students. Taxonomy and criteria that govern the process may be refined as a consequence of the evaluation exercise.

- *Listservs* maintained for groups of teachers sharing similar interests, and whereby peer discussions take place about what works and what not in the classroom. Such a tool has many potentials and capabilities including immediate feedback by peers and support staff (including material developers), which helps concerned teachers reconsider matters of interest in a timely manner.

2. For assessing student profiles:

- *Conceptual surveys*, including paper-and-pencil surveys and interviews with concerned students. Examples of such surveys are discussed in the following section.
- *Observation grids* similar to the ones used for teacher practice and covering all sorts of student activities in the classroom. Different grids may be used for different activities and student products.
- *Student logs*, whereby teachers record regularly and systematically what every student and/or groups of students achieve of the respective taxonomy in every teaching unit.
- *Formal course assessment*, including notes and grades on homeworks, exams and all sorts of assignment.

Conceptual surveys administered to students are perhaps the most reliable among all these instruments, especially when these surveys are developed and standardized by outside evaluators, and when they allow quantitative assessment of student achievement. Evaluators though need to ensure that such instruments are developed in accordance with a taxonomy that matches the one they are concerned with, and that instruments' criteria match the targeted benchmarks. *Standardized instruments* offer a special advantage. Corresponding criteria are set not only intrinsically in terms of the scientific theory the understanding of which is under investigation, but, more importantly, following inferential analysis of actual outcomes obtained outside the context of the curriculum under evaluation. This would facilitate grading the criteria as mentioned above (say, between basic and critical thresholds) and setting thresholds at objective and realistic levels.

4. Normative evaluation of the modeling curriculum

Normative evaluation as outlined above was implemented in the development of modeling theory, and particularly in the development of physics curricula for secondary school and college levels within the framework of the theory in question. Major aspects of the evaluation process are described in this section so as to illustrate how one can go about conducting such an evaluation in the direction of continuous curriculum regulation.

Development of modeling theory practically began in 1981 as part of this author's doctoral dissertation at Arizona State University. The theory was first developed in the context of introductory college physics courses, then it was gradually deployed into secondary school physics courses. Now that the theory has been duly corroborated for physics curricula, efforts are under way to deploy it into other disciplines of all educational levels (Halloun, 2004). The modeling curriculum in physics was first implemented in limited settings in USA and Lebanon. Now, it has become a national curriculum in USA where it was rated an *exemplary* curriculum by the Department of Education. Variants of the modeling curriculum are currently being implemented in many countries around the world.

Formal evaluation of the modeling curriculum takes place in accordance with norms and guidelines described in the previous section. Individual teachers implementing the curriculum are asked to assess the effectiveness of their own instruction and participate in action-research leading to continuously more meaningful and equitable learning experience across the board. To these ends, teachers and evaluators make use of batteries of instruments similar to the ones described above for assessing teacher practice and student performance (§ 3). However, standardized instruments of conceptual understanding are the most relied upon for evaluating the modeling curriculum in terms of student achievement, and particularly for assessing whether students have achieved to an acceptable level the paradigmatic evolution promoted by the curriculum. Let us illustrate the process in the case of classical (Newtonian) mechanics.

Meaningful paradigmatic evolution in Newtonian mechanics is set in terms of a set of basic models. The set includes four to five models, depending on the course level (Halloun, 2004). An anticipated student profile includes conceptions and processes required for the construction and deployment of these models, as well as generic trends about physics and its relevance in everyday life. A number of standardized instruments are used in order to assess to what extent a student has developed the anticipated profile. Among these instruments are the following:

- *Inventories of Basic Conceptions** (IBC). Each IBC is a paper-and-pencil instrument designed to assess students' understanding of fundamental *conceptions* required for the construction and deployment of basic models in a particular scientific theory. IBC-Mechanics is the one pertaining to Newtonian theory of mechanics. It supersedes the Mechanics Diagnostic Test or MDT (Halloun, 1986, 2001a; Halloun & Hestenes, 1985) and its offspring, the Force Concept Inventory or FCI (Halloun & Hestenes, 1995; Hestenes, Wells & Swackhamer, 1992). Conceptions assessed in IBC-Mechanics include the concepts of force, position, velocity and acceleration and their relations to one another as established by state laws and Newton's laws of dynamics. An item in these instruments consists of a multiple-choice question about a particular situation in everyday life. Answering a given question requires only qualitative understanding of assessed conceptions and no quantitative manipulation of any sort of mathematical representations.
- The *Mechanics Baseline Test* or MBT (Hestenes & Wells, 1992). This is a multiple choice, paper-and-pencil instrument that assesses student understanding of fundamental *processes** required for the construction and deployment of the models in questions. More specifically, MDT targets required mathematical representations and operations. Often MBT is supplemented or substituted with projects or paper-and-pencil problems that require particular modeling skills.
- The *Views About Science Survey** or VASS (Halloun, 2001a&b; Halloun & Hestenes, 1998). This is also a paper-and-pencil instrument. Its items are formulated following Halloun's *Contrasting Alternatives* rating scale. VASS is intended to assess student views about the structure and validity of scientific theory, scientific methodology, reflective thinking (learning styles), learnability and personal relevance of a particular discipline.

Each of these instruments has been administered to thousands of secondary school and university students around the world, students enrolled in physics courses (and other science courses for VASS) that are taught following a variety of instructional strategies. Strategies extend from conventional instruction of lecture and demonstration to all sorts of student-centered courses that follow modern educational theory. All three sets of instruments are duly validated and standardized following graded criteria as discussed above.

Student profile anticipated following a given course of classical mechanics is graded in taxonomy and criteria between two fundamental levels corresponding to the basic and the critical thresholds. As discussed in the previous section, and according to modeling theory, the *basic threshold* of paradigmatic evolution corresponds to conceptions and processes required for the construction of the free particle model and the uniformly accelerated particle model. The *critical threshold* corresponds to the four models mentioned in § 3. To each of these thresholds corresponds a particular set of items and a particular score on IBC-Mechanics (or its predecessor, FCI) and MBT. Similar thresholds and characteristic scores are associated with student views as assessed in VASS (Halloun, 2001b).

---

* IBC-Mechanics and IBC-DC circuits, as well as VASS forms for physics, chemistry and biology, are available upon request from this author. Other IBC forms are currently under development along with companion process and trend inventories.

Practice of teachers following modeling instruction is also assessed with an appropriate battery of instruments. Like in the case of students, the most relied upon instrument is a paper-and-pencil survey. Teachers fill out this Likert survey at the end of a semester or year. The instrument is intended to ascertain to what extent individual teachers implement particular aspects of modeling instruction in their classes. These aspects pertain to modeling tenets, model-centered course content, modeling/inquiry processes, student reflective and interactive engagement, scientific discourse, promoted learning styles, teacher mediation, and action-research. A quantitative score is associated with a teacher answer on a particular item, and a sum of scores is calculated for every teacher. Item lumping and score grading on the teacher survey are done so as to reflect overall teacher practice in two major respects. One pertains to course content and structure, with emphasis on models and related processes. The other pertains to mediated learning and various aspects of learning cycles.

Development of the modeling curriculum began in the eighties, in a limited number of college then high school mechanics classes, in USA, and then in Lebanon. At that time, instructors made use of available textbooks and laboratory equipment. Textbooks were supplemented with handouts making explicit the model-based structure of Newtonian theory, and traditional laboratory equipments were set up to suit the needs of modeling instruction to the extent that was possible. In the mid nineties, development of the modeling curriculum went on in USA on a larger scale. A broad band of high school and college teachers got gradually involved in the process, and high school and introductory college physics courses got gradually all covered in the process. A few years later, the modeling curriculum was widely disseminated within and outside the U.S. In the following section, we report on evaluation data pertaining to selected samples that participated in the development and dissemination of the modeling curriculum in classical mechanics. Data are presented and discussed so as to illustrate with actual examples how the instruments presented above were used in normative evaluation of the modeling curriculum.

5. Evaluation for regulation in action

Use of all student instruments and interpretation of respective data are governed by identical or similar norms and criteria. That is why we concentrate in what follows on only one student instrument, the FCI, and the way it has been used in the evaluation and development of the physics modeling curriculum at the high school level. IBC-Mechanics has been put to use in the current academic year. Its immediate precursor, the FCI, was previously the most relied upon for assessing student conceptions in mechanics. Examples of FCI data are provided and interpreted in relation to data pertaining to teacher practice at certain stages of curriculum development. The object is to illustrate how data can be collected, displayed, analyzed and interpreted so as to reveal as best as possible weaknesses and strengths in a new curriculum, and thus refine the curriculum accordingly.

Reported data pertain to two samples of secondary school classes. The first sample consists of classes of about fifty high school teachers who participated in the early broad dissemination of the modeling curriculum, and who were involved in the development of the first modeling materials (manuals and laboratory setups). The second sample consists of classes of ten secondary school teachers who joined in later in the process, at a time where the modeling curriculum began to take its current shape.

At an early stage of the modeling curriculum development, about fifty in-service high school teachers were trained at length to implement modeling theory in teaching classical mechanics at the high school level. During their first year of training, these teachers participated in the development and field-testing of a learning manual for classical mechanics. The manual includes guidelines for MBL and other activities devoted for the construction and deployment of basic models. Like all others participating in the development and implementation of the modeling curriculum, these teachers administered to their own students all three student instruments (FCI, MBT and VASS). Students were given each test as pretest and posttest, the year preceding their teacher training, and regularly afterwards, during and after training. In parallel, teachers themselves were administered a teacher survey at the end of every academic year in order to monitor changes in their teaching practice throughout training and some time afterwards.

Figure 3 shows FCI data for students of the fifty teachers collected at the beginning (pretest) and end (posttest) of the first year of training. Data are displayed in the form of three clouds or clusters (mostly to preserve confidentiality). The boundaries of each cluster actually consist of a best-fit envelope that can be wrapped around a particular cluster of data points. Each point originally represents the posttest average plotted against the corresponding pretest average of FCI scores of students taking classical mechanics with a particular teacher. The three clusters or groups are distinguished with the extent to which teachers implemented modeling instruction in their classes, as revealed with answers on an early version of the teacher survey. Group 1 refers to teachers who claimed to be most systematic in this respect (20% of participants), and group 3 to those (46%) who claimed to be least systematic (virtually sticking to conventional instruction). Group 2 (34% of participants) stands in the middle between the two extreme groups. Needless to say that by the first year of trial, none of participating teachers had already mastered the modeling approach, and modeling materials were just being piloted and were still far from their current state.

Five critical lines are shown in Figure 3, three horizontal and two inclined. The bottom horizontal line (posttest average of 50%) corresponds to the *basic threshold* on the FCI, i.e., to the minimum

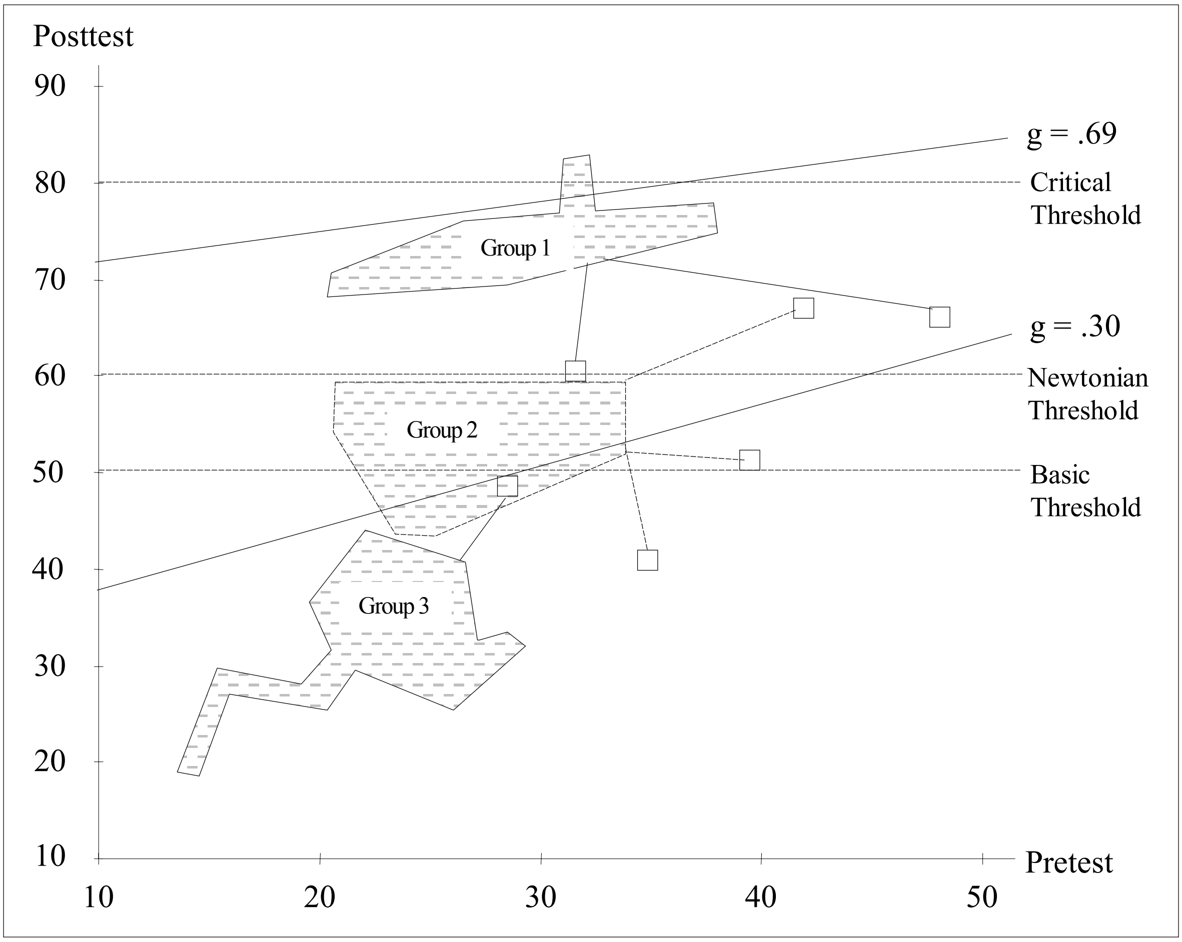

Figure 3. FCI pretest-posttest data (in score percentage) for high school students of three groups of teachers who took part in the early stages of development of the modeling curriculum.

The three groups are ranked from most (Group1) to least (Group 3) systematic in their implementation of various aspects of the modeling curriculum. Individual outliers in each group are represented, like in boxplots, with squares connected to corresponding clusters.

level of understanding required for students to be able to construct the two most elementary basic models of Newtonian mechanics (free particle and uniformly accelerated particle). The top line (80%) corresponds to the *critical threshold* or mastery level required for meaningful understanding of the entire set of basic models. The line between the two (60%) corresponds to what we have called *Newtonian threshold*. This is the *minimum* level students need to attain in order to *begin* understanding *all generic* aspects of the Newtonian theory within the context of basic models. The three thresholds have been set after long years of FCI administration and analysis of data pertaining to thousands of students around the world. The two slanting lines correspond to average pretest-posttest gains (g) on the FCI for two groups of students. The gain expression is given by:

$$g = \frac{< \text{Posttest\%} > - < \text{Pretest\%} >}{100 - < \text{Pretest\%} >}$$

The g = .30 line corresponds to the national threshold of interactive engagement (IE) courses in USA. These are high school physics courses where teams of students are actively engaged in reflective inquiry in the classroom. Hake (1998) showed that, average gains of IE classes fall all *above* this threshold, while average gains of classes following conventional instruction of lecture and demonstration fall all *below* the same threshold. He also revealed that the overall average gain for IE courses is about .48 (not shown in Figure 3), and that g = .69 is the maximum average gain ever reported in USA for IE courses.

The overall distribution of pretest means in Figure 3 is in line with the international distribution of FCI pretest scores. Pretest class means fall all below the basic threshold (50%) as one would expect. Except for one case in Group 1, pretest means are even all below 40%. Classes in Group 3 appear to be originally at a slight disadvantage with respect to the other two groups; their pretest means are all equal to or below 30%, and only in this group there are two classes whose pretest average is below 20%. Analysis of variance on pretest means revealed no significant differences among the three groups (at the .05 level).

Significant differences are though detected among the three groups on FCI posttest data. Among these differences we note the following:

1. For the same pretest means, classes in Group 1 achieve the highest gains on the posttest, followed by those in Group 2 then Group 3. For example, among classes who averaged around 27% on the pretest: (a) all three classes in Group 1 averaged 70% or more on the posttest, (b) the five classes in Group 2 averaged between 45% and 60%, and (c) all three classes in Group 3 averaged below 40%.

2. 80% of classes in Group 1 averaged 70% or more on the posttest (i.e., closer to the critical threshold than the Newtonian threshold), and had pretest-posttest gains averaging well above the IE mean of .48 and close to the .69 ceiling of IE courses. One class in this group reached a posttest average just above the 80% mark of the critical threshold, and had an average gain of .72. Average posttest score and gain in this class were at the time the highest ever reported on the FCI in a high school. Overall, and except for one class in Group 2, posttest means (and gain g for the upper 80% of classes) in Group 1 had no match in any of the other two groups.

3. About 80% of classes in Group 2 averaged between 50% (basic threshold) and 60% (Newtonian threshold) on the posttest. One class in this group had a posttest average of about 68%. Remaining classes averaged between 40% and 50%. Four classes in this group had an average pretest-posttest gain just below the IE threshold of .30, while all others (83% of Group 2) had gains averaging above this threshold.

4. All classes in Group 3 had a posttest average below the basic threshold of 50% and a gain below the IE threshold of .30.

As mentioned previously, participating teachers were in their first year of training on the modeling curriculum, and learning materials (as well as the teacher survey) were at the early stages of their development. This is what makes the chosen sample so interesting. It comes at the beginning of the large-scale experimental phase of the modeling curriculum, and it includes a wide array of teacher practice in the classroom extending from conventional instruction (including the use of significantly more conventional learning materials than modeling materials) to IE instruction that begins to come close to modeling instruction but that is not quite there yet. As such, the sample is quite heterogeneous in terms of what its members implemented of the modeling curriculum. It allows detailed analysis of what works and what not in this curriculum, with as close as possible to a good control of variables. A specific curriculum dimension can be isolated (to a certain extent), and its impact on student achievement can be conveniently analyzed almost independently of other dimensions.

Further analysis of results displayed in Figure 3, and especially of those listed in the four points above reveals the following:

1. By comparison to FCI data pertaining to the students of the same teachers a year before training (not reported in this paper), posttest means and gains are higher for classes of all teachers following their first year of training, even for those in Group 3 who, in all practical aspects, were still following conventional instruction. This already indicates that by and large the modeling curriculum seems to be on the right track.

2. Posttest and gain results improve gradually from Group 3 to Group 1. This indicates that the more comprehensively teachers follow the modeling approach, i.e., the more dimensions of the modeling curriculum they implement in their classes, the better student achievement.

3. Preliminary analysis of individual teachers' practice revealed that teachers in Group 2 were more systematic in changing their own methodology in the classroom than in changing the content and structure of their courses. They focused on learning cycles, but in the direction of completing course content as available more in conventional textbooks than in the modeling manual that they were supposed to pilot. The same was true, though to a slightly lesser extent, for the two teachers in Group 1 who had the lowest FCI gains.

4. Analysis of Group 1 practice revealed that, and except for the two teachers just mentioned, teachers in this group were trying to balance between instructional methodology and restructuring course content following the modeling manual.

5. Analysis of differences in teacher's practice with respect to content structure revealed that the more systematic teachers were in helping students develop models of well-defined structure, the better student achievement. Specific modeling tools and rules (including MBL related) had proven to be more effective (and sometimes more realistic) than others in this direction.

6. Analysis of student answers on individual items on the FCI revealed a certain hierarchy in conceptual complexity and cognitive requirements across the respective taxonomy. It showed that student understanding of Newtonian theory develops gradually and almost in a specific order through certain elements in the taxonomy, and this irrespective of teacher practice. Basic threshold is crossed only after students develop a specific set of conceptions in Newtonian theory. Newtonian and critical thresholds are reached only after students develop increasingly more comprehensive sets of the theory's conceptions. It was thus revealed that certain clusters of items in the taxonomy are critical for reaching or passing a certain threshold, and that students need to develop conceptions in a certain order within each cluster so that they proceed in their paradigmatic evolution successfully and meaningfully. This, again, was true irrespective of teacher practice. However, the practice of teachers in Group 1 fostered the evolution in question more systematically within and across conceptual clusters than the practice of teachers in

Group 2. Teachers in Group 3 were not capable of promoting the evolution to any acceptable level or in any coherent way.

7. Analysis of teachers' implementation of learning cycles revealed that certain phases of a cycle were more critical than others, and that certain refinements were needed at this level. It also revealed the necessity to direct and manage teacher and peer feedback, as well as student discourse, more in certain directions than others.

FCI outcomes were consistent with those obtained with other evaluation instruments. Based on the evaluation exercise conducted with all instruments, the modeling curriculum was refined in many respects, including course content and manual, and MBL setups. The refinement process lasted for a number of years (and is still in process, though it is now more in the direction of exploring new tools than in any other respect). In later stages, teachers from all around the world began implementing the modeling curriculum in secondary school and college physics courses. Figure 4 shows evaluation outcomes for a small group of secondary school classes that joined in later in the process.

Figure 4 pertains to a group of ten secondary school teachers who decided at a certain point to pilot the modeling curriculum in their classical mechanics course. Like all other participants, these teachers administered a number of instruments, including FCI, to their students as pretest and posttest during the first year of their pilot program. Teachers themselves filled out the Modeling Practice Survey (MPS) at the end of the same year. MPS is an upgrade of the teacher practice survey used in Figure 3. It includes 100 Likert questions asking teachers about the extent to which they implement various aspects of the modeling curriculum. A perfect score on the MPS is of 200; it indicates that a teacher claims to implement the modeling curriculum to the letter. A point in Figure 4 shows the score of a particular teacher on the MPS against the average pretest-posttest gain g on the FCI of the

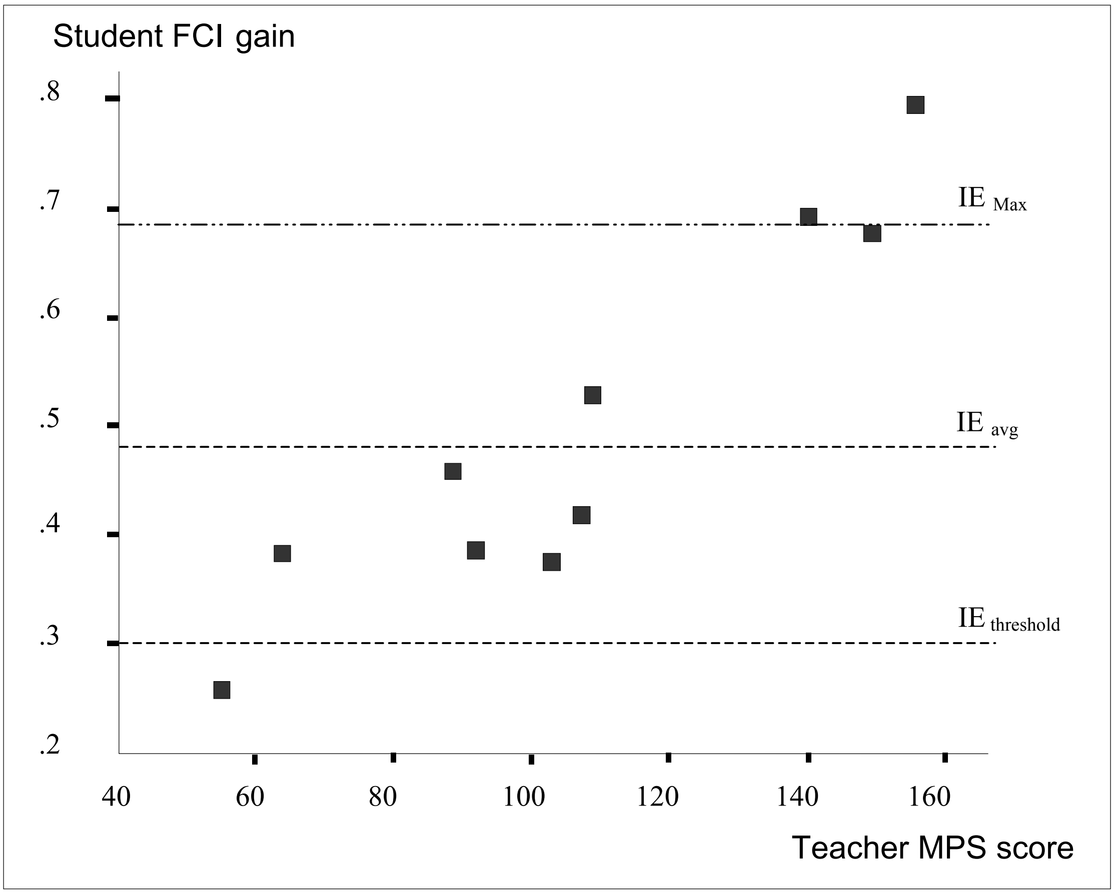

Figure 4. Student FCI gain versus teacher practice as assessed with the Modeling Practice Survey (MPS).

students of this particular teacher. The overall distribution of points is quite similar to results obtained elsewhere and at various stages of the development of the modeling curriculum. It indicates a consistent trend in the outcomes: a significant correlation between teacher practice and student achievement, achievement that gets significantly better as teachers account more and more systematically to various dimensions of the modeling curriculum.

Teachers in Figure 4 fall actually in three distinctive groups. The first group consists of the three teachers whose MPS score is equal to or above 140. We will refer to this group as Group 1. The second group consists of the five teachers whose MPS score is above 80 but below 120 (Group 2). The third group consists of the two teachers whose MPS score is below 80 (Group 3). Close comparison of groups of teachers implicated in Figures 3 and 4 reveals noticeable similarities between groups bearing the same number both in terms of teacher practice and student achievement. Group 1 in both figures consists of teachers who are most systematic in implementing the modeling curriculum. Group 3 in both places consists of teachers who are still favoring conventional instruction. Group 2 is somewhere in the middle between the two, and respective teachers in both Figures 3 and 4 concentrate more on instructional methodology than on model-centered content. The trend in both figures is typical of all outcomes obtained in the evaluation of the modeling curriculum. It has the following major characteristics:

1. Best results with the modeling curriculum are obtained when all dimensions of the curriculum are accounted for systematically and comprehensively. Selective implementation of some components but not others from each of the two broad dimensions (instructional methodology and course content) may enhance student achievement enough to cross the basic threshold or even the Newtonian threshold, but not enough to reach the critical threshold.

2. When teachers cover comprehensively one broad dimension rather than the other, students are enabled to cross the basic or even Newtonian threshold but not the critical threshold. This is reflected in Figure 4 where the teacher in Group 3 with the higher MPS score concentrated, though erratically, more on restructuring content than on following the learning cycle, and where all teachers in Group 2 had the balance tipped in the opposite direction.

3. Between the two broad dimensions, the one pertaining to restructuring course content around basic models and emphasizing modeling processes has a significantly higher impact on student achievement than the dimension pertaining to instructional methodology. However, only when the two are conveniently accounted for, students can cross the critical threshold of a given course (and the IE ceiling).

4. Learning materials, other than course manual, and including MBL, have little impact on student achievement if not used in the manner advocated by modeling theory to develop a scientific theory that is the object of a science course around basic models, and to foster systematic development of modeling processes.

6. Toward meaningful and equitable learning of mathematics, science and technology

Our long experience with the modeling curriculum shows that normative evaluation as described in this paper is a necessary condition: (a) for science education to help students evolve into the scientific realm in meaningful ways and become adequately skilled for scientific practice, (b) for MSTE to become an equitable enterprise whereby all students willing to invest necessary efforts, and capable of doing so, are enabled to cross certain thresholds, and (c) for curriculum reform to be sustainable. More specifically, curriculum evaluation must be conducted according to pre-established taxonomy and criteria set primarily in terms of the educational theory within the framework of which curriculum was formulated. Furthermore, normative evaluation should not be an end by itself. It should not be carried out for the mere sake of ascertaining the intrinsic value of a given curriculum, but for the purpose of continuous regulation of various aspects of the curriculum. Concerned teachers must take an active role as action-researchers in this direction.

Various curriculum components cannot be developed and subsequently evaluated independently of one another. Efficacy of a particular mean or mode of learning and instruction is not an intrinsic factor that depends only on the nature of the component of concern. For example, the merits of learning materials depend on the way materials are being used in the learning process. Subsequently, it is indispensable that learning materials be evaluated in terms of their effectiveness on student achievement, and this in relation to instructional methodology and student learning styles inside and outside the classroom. This is how we proceeded with the modeling curriculum, as described in part in the previous section.

For the evaluation exercise to be effective, especially in the direction of continuous curriculum regulation, concerned teachers must get involved in systematic action research intended to get their practice increasingly more equitable and student understanding (and skills) more and more meaningful. Unless teachers act as such after they get adequately trained and motivated, curriculum and materials would fail to meet their ends no matter how well they may be originally articulated. This brings us to a critical issue, teacher training and motivation.

Everytime some novelty is introduced in a curriculum, teachers must be adequately trained so they know how to appropriately adjust their practice. This is true for both pre-service and in-service teachers. It is especially true for the latter, no matter how long an experience they might have. In fact, teaching experience, especially when of conventional type, may be a detriment in the process. Teachers who get used to specific method and materials often develop some inertia and even close-mindedness that prevent them from being capable or even open to adjust to certain novelties on their own. This was in fact the case with Groups 3 in Figures 3 and 4. Our experience suggests that the shift of in-service teachers from conventional to IE or modeling instruction requires long training and particular incentives. Contrary to what some policymakers and educational administrators around the world, and especially the developing world, actually believe and push for on the ground, innovations in MST curricula cannot be shoved to in-service teachers in a few days or even weeks of training (if any!).

Such innovations, especially when in line with modern educational theory like modeling theory, require that teachers spend at least one full academic year of special training, followed by years of monitoring and support on the ground. In parallel, teachers need to be offered particular incentives, including but not limited to reduction of contact hours at school and special fringe benefits. Support must be provided, almost on a daily basis in the early implementation of new materials, locally and/or remotely through appropriate support systems and programs. These systems and programs are typically run by educational experts, including master teachers (i.e., teachers who have developed some experience with the new curriculum), university professors, material developers and publishers, who are capable of providing timely and effective feedback to all sorts of inquiry put forth by teachers and administrators on the ground. In this respect, particular partnerships between local universities and schools are especially helpful. The object of such partnerships goes beyond formal training of in-service teachers, which must be reserved to universities and teacher training institutions. It extends to all nuts and bolts of a curriculum, and especially to action-research that is better carried out by teams of school teachers and university professors. Teachers bring in their classroom experience and daily concerns, and professors contribute their expertise in educational theory, scientific theory, modern technology, statistical analysis and various aspects of educational research.

All in all, normative evaluation for curriculum regulation is a tedious process that requires understanding and commitment on behalf of many actors, including, but not restricted to, policymakers, administrators, curriculum developers, publishers, teachers, and, of course, students. Long-term training and support systems and programs need to be instituted in partnership with local universities and teacher training institutions. Means and modes of learning and instruction need to be developed and evaluated in concert with concerned teachers and primary actors in support systems, and this within the framework of an appropriate educational theory that helps setting ahead of time explicit criteria of success and failure, criteria that account to certain extents to the SWOT rule on the ground. Failure to proceed in this direction would condemn a new curriculum to failure no matter how well it may be originally intended and formulated.

Acknowledgments

This paper is based in part on work presented at a regional UNESCO workshop (Halloun, 2003).

Evaluation of the modeling curriculum in the way reported in this paper could not have been possible without the collaboration of numerous colleagues and students around the world and the sponsorship of a number of organizations. They all deserve to be duly acknowledged. Special acknowledgment is though due to the collaboration of David Hestenes and the Modeling Research Team at Arizona State University, and to the sponsorship of the National Science Foundation in USA and Lebanese University in Beirut, Lebanon.

References

Bibeau, J. R., Halloun, I., May, J. & Reddy, V. (2002). *Report on the Evaluation of UNESCO's Science and Technology Education (STE) Programme*. Evaluation report presented to the Science and Technology Education Division, UNESCO. Paris, France.

Hake, R. R. (1998). Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. *American Journal of physics*, *66* (1), 64-74.

Halloun, I. (In press). *Modeling Theory in Science Education.* Dordrecht: Kluwer Academic Publishers.

Halloun, I. (2004). Modeling Theory for Paradigmatic Evolution. *Proceedings of the 12th Annual Meeting of the South African Association for Research in Mathematics, Science and Technology Education*.

Halloun, I. (May 2003). Evaluating science & technology learning materials. *Plenary paper presented at UNESCO regional workshop*. Beirut, Lebanon.

Halloun, I. (2001a). *Apprentissage par Modélisation : La Physique Intelligible.* Beirut, Lebanon: Phoenix series / Librairie du Liban Publishers.

Halloun, I. (2001b). *Student Views about Science: A comparative Survey.* (2001). Beirut, Lebanon: Phoenix series / Educational Research Center, Lebanese University.

Halloun, I. (1986). Le réalisme naif et l'apprentissage de la physique. *Recherches Pédagogiques*, 17, 23-47.

Halloun, I., & Hestenes, D. (1998). Interpreting VASS dimensions and profiles. *Science & Education,* 7 (6), 553-577.

Halloun, I., & Hestenes, D. (1995). Interpreting the Force Concept Inventory. *The Physics Teacher,* 33 (8), 502-506.

Halloun, I., & Hestenes, D. (1987). Modeling Instruction in Mechanics. *American Journal of Physics*, 55 (5), 455-462.

Halloun, I., & Hestenes, D. (1985). The initial knowledge state of college physics Students. *American Journal of Physics,* 53 (11), 1043-1055.

Hestenes, D., & Wells, M. (1992). A Mechanics Baseline Test. *The Physics Teacher,* 30 (3), 159-166.

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