The Assessment of Knowledge, in Theory and in Practice*

Jean-Claude Falmagne
University of California, Irvine

Eric Cosyn
ALEKS Corporation

Jean-Paul Doignon
Free University of Brussels

Nicolas Thiéry
ALEKS Corporation

Abstract

This paper is adapted from a book and many scholarly articles. It reviews the main ideas of a novel theory for the assessment of a student's knowledge in a topic and gives details on a practical implementation in the form of a software system available on the Internet. The basic concept of the theory is the 'knowledge state,' which is the complete set of problems that an individual is capable of solving in a particular topic, such as Arithmetic or Elementary Algebra. The task of the assessor—which is always a computer—consists in uncovering the particular state of the student being assessed, among all the feasible states. Even though the number of knowledge states for a topic may exceed several hundred thousand, these large numbers are well within the capacity of current home or school computers. The result of an assessment consists in two short lists of problems which may be labelled: 'WHAT THE STUDENT CAN DO' and 'WHAT THE STUDENT IS READY TO LEARN.' In the most important applications of the theory, these two lists specify the exact knowledge state of the individual being assessed. This work is presented against the contrasting background of common methods of assessing human competence through standardized tests providing numerical scores. The philosophy of these methods, and their scientific origin in nineteenth century physics, are briefly examined.

The assessment of human competence, as it is still performed today by many specialists in the schools and in the workplace, is almost systematically based on the numerical evaluation of some 'aptitude.' Its philosophy owes much to nineteenth century physics, whose methods were regarded as exemplary. The success of classical physics was certainly grounded in its use of a number of fundamental numerical scales, such as mass, time, or length, to describe basic aspects of objects or phenomena. In time, 'measurement' came to represent the *sine qua non* for precision and the essence of the scientific method, and physics the model for other sciences to imitate. In other words, for an academic endeavor to be called a 'science,' it had to resemble physics in critical ways. In particular, its basic observations had to be quantified in terms of measurement scales in the exact sense of classical physics.

*Send correspondence to: Jean-Claude Falmagne, Dept. of Cognitive Sciences, University of California, Irvine, CA 92697. Phone: (949) 824 4880; FAX: (949) 824 1670; e-mail: jcf@uci.edu. We wish to thank Chris Doble, Dina Falmagne, and Lin Nutile for their reactions to earlier drafts of this article.

Prominent advocates of this view were Francis Galton, Karl Pearson and William Thomson Kelvin. Because that position is still influential today, with a detrimental effect on fields such as 'psychological measurement,' which is relevant to our subject, it is worth quoting some opinions in detail. In Pearson's biography of Galton (Pearson [1924, Vol. II, p. 345]), we find the following definition:

> "**Anthropometry**, *or the art of measuring the physical and mental faculties of human beings, enables a shorthand description of any individual by measuring a small sample of his dimensions and qualities. This will sufficiently define his bodily proportions, his massiveness, strength, agility, keenness of senses, energy, health, intellectual capacity and mental character, and will constitute concise and exact* **numerical**[1] *values for verbose and disputable estimates*[2]."

For scientists of that era, it was hard to imagine a non-numerical approach to precise study of an empirical phenomenon. Karl Pearson himself, for instance—commenting on a piece critical of Galton's methods by the editor of the *Spectator*[3]—, wrote

> *"There might be difficulty in ranking Gladstone and Disraeli for 'candour,' but few would question John Morley's position relative to both of them in this quality. It would require an intellect their equal to rank truly in scholarship Henry Bradshaw, Robertson Smith and Lord Acton, but most judges would place all three above Sir John Seeley, as they would place Seeley above Oscar Browning. After all, there are such things as brackets, which only makes the statistical theory of ranking slightly less simple in the handling."* (Pearson [1924, Vol. II, p. 345].)

In other words, measuring a psychical attribute such as 'candor' only requires fudging a little around the edges of the order relation of the real numbers[4]. The point here is that real numbers are still used to represent 'quantity of attribute.'

As for Kelvin, his position on the subject is well known, and often represented in the form: "*If you cannot measure it, then it is not science.*" The full quotation is:

> *"When you can measure what you are speaking about, and express it in numbers, you know something about it; but when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meager and unsatisfactory kind: it may be the beginning of knowledge, but you are scarcely, in your thoughts, advanced to the stage of* **science***, whatever the matter may be."* (Kelvin [1889].)

[1] Our emphasis.

[2] This excerpt is from an address "Anthropometry at Schools" given in 1905 by Galton at the London Congress of the Royal Institute for Preventive Medicine. The text was published in the *Journal for Preventive Medicine*, Vol. XIV, p. 93-98, London, 1906.

[3] The *Spectator*, May 23, 1874. The editor was taking Galton to task for his method of ranking applied to psychical character. He used 'candour' and 'power of repartee' as examples.

[4] Making such a relation a 'weak order' or perhaps a 'semiorder' (in the current terminology of combinatorics). A binary relation $\precsim$ on a finite or countable set S is a *weak order* if there is a real valued function f defined on S such that $x \precsim y \Leftrightarrow f(x) \leq f(y)$ for all objects x and y in the set S. The relation $\precsim$ is a *semiorder* if the representation has the form: $x \precsim y \Leftrightarrow f(x)+1 \leq f(y)$. For these concepts, see e.g. Roberts [1979] or Trotter [1992].

Such a position, which equates precision with the use of numbers, was not on the whole beneficial to the development of mature sciences outside of physics. It certainly had a costly impact on the assessment of mental traits. For instance, for the sake of scientific precision, the assessment of mathematical knowledge was superseded in the U.S. by the measurement of mathematical aptitude using instruments directly inspired from Galton via Alfred Binet in France. They are still used today in such forms as the S.A.T.[5], the G.R.E. (*Graduate Record Examination*), and other similar tests. The ubiquitous I.Q. test is of course part of the list. In the minds of those nineteenth century scientists and their followers, the numerical measurement of mental traits was to be a prelude to the establishment of sound, predictive scientific theories in the spirit of those used so successfully in classical physics. The planned constructions, however, never went much beyond the measurement stage[6].

The limitations of a purely numerical description of some phenomena can be illustrated by an analogy with sports. It is true that the success of an athlete in a particular sport is often described by a set of impressive numbers. So, imagine that some committee of experts has carefully designed an 'Athletic Quotient' or 'A.Q.' test, intended to measure athletic prowess. Suppose that three exceptional athletes have taken the test, say Michael Jordan, Tiger Woods and Pete Sampras. Conceivably, all three of them would get outstanding A.Q.'s. But these high scores equating them would completely misrepresent how essentially different from each other they are. One may be tempted to salvage the numerical representation and argue that the asssessment, in this case, should be multidimensional. However, adding a few numerical dimensions capable of differentiating Jordan, Woods and Sampras would only be the first step in a sequence. Including Greg Louganis or Pele to the evaluated lot would require more dimensions, and there is no satisfactory end in sight. Besides, assuming that one would settle for a representation in n dimensions, for some small n equal 3, 4 or 5 say, the numerical vectors representing these athletes would be poor, misleading expressions of the exquisite combination of skills making each of them a champion in his own specialty. Evidently, the same shortcomings of a numerical description also apply in mathematics education. Numerical test results may be appropriate to decide who is winning a race. As an evaluative prelude to college, intended to assess the students' readiness for further learning, they are very imprecise indeed. The conclusion should be that a different descriptive language is needed.

More generally, in many scientific areas, from chemistry to biology and especially the behavioral sciences, theories must often be built on a very different footing than that of classical physics. Evidently, the standard physical scales such as length, time, mass or energy, must be used in measuring aspects of phenomena. But the substrate proper to these other sciences may very well be, in most cases, of a fundamentally different nature.

Of course, we are enjoying the benefits of hindsight. In all fairness, there were important mitigating circumstances affecting those who upheld the cause of numerical measurement as a prerequisite to science. For one thing, the appropriate mathematical tools were not

[5]Note that the meaning of the acronym S.A.T. has recently been changed by Education Testing Service from '*Scholastic Aptitude Test*' to '*Scholastic Assessment Test*,' suggesting that a different philosophy on the part of the test makers may be under development.

[6]Sophisticated theories can certainly be found in some areas of the behavioral sciences, for example, but they do not usually rely on measurement scales intrinsic to these sciences. One prominent exception in economics is the *money* scale.

yet available to support different conceptions. Combinatorics, for example, was yet to be born as a mathematical topic. More importantly, the 'Analytical Engine' of Charles Babbage was still a dream, and close to another century had to pass before the appearance of computing machines capable of handling the symbolic manipulations that would be required for another approach.

The theory reviewed here represents a sharp departure from other approaches to the assessment of knowledge. Its mathematics is in the spirit of current research in combinatorics. No attempt is made to obtain a numerical representation. We start from the concept of a possibly large but essentially discrete set of 'units of knowledge.' In the case of Elementary Algebra, for instance, one such unit might be a particular type of algebra problem. The full domain for High School Algebra may contain a couple of hundred such problems. Our two key concepts are the 'knowledge state,' a particular set of problems that some individual is capable of solving correctly, and the 'knowledge structure,' which is a distinguished collection of knowledge states. For High School Algebra, we shall see that a useful knowledge structure may contain several hundred thousand feasible knowledge states. Thus, precision is achieved by the intricacy of the representing structure.

Knowledge structures: main concepts

The precedence relation. A natural starting point for an assessment theory stems from the observation that some pieces of knowledge normally precede, in time, other pieces of knowledge. In our context, some algebra problem may be solvable by a student only if some other problems have already been mastered by that student. This may be because some prerequisites are required to master a problem, but may also be due to historical or other circumstances. For example, in a given environment, some concepts are always taught in a particular order, even though there may be no logical or pedagogical reason to do so. Whatever its genesis may be, this precedence relation may be used to design an efficient assessment mechanism.

A simple example of a precedence relation between problems is illustrated by Fig. 1, which displays a plausible *precedence diagram* pertaining to the six types of algebra problems illustrated in Table 1. Note in passing that we distinguish between a **type** of problem and an **instance** of that type. Thus, a type of problem is an abstract formulation subsuming a possibly large class of instances. For the rest of this article, 'problem' is almost always intended to mean 'problem type.' The exceptions will be apparent from the context.

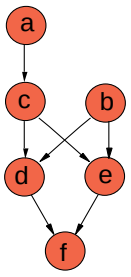

Fig. 1. Precedence diagram for the six types of algebra problems illustrated in Table 2.

The precedence relation between problems is symbolized by the downward arrows. For example, Problem (e) is preceded by Problems (b), (c) and (a). In other words, the mastery of Problem (e) implies that of (b), (c) and (a). In the case of these six problems, the precedence relation proposed by the diagram of Fig. 1 is a credible one. For example, if a student responds correctly to an instance of Problem (f), it is highly plausible that the same student has also mastered the other five problems. Note that this particular precedence relation is part of a much bigger one, representing a comprehensive coverage of all of Beginning Algebra, starting with the solution of simple linear equations and ending with problem types such as (f) in Table 1. An example of such a larger precedence relation is represented by the diagram of Fig. 2. (The diagram has 88 vertices, for the 88 problems used for the assessment. The full Beginning Algebra curriculum is slightly larger.) This larger precedence diagram is itself part of a still larger one, displayed in Fig. 3, and comprising Arithmetic, Beginning Algebra, Intermediate Algebra, and Pre-Calculus.

Table 1: Six types of problems in Elementary Algebra

	Name of problem type	Example of instance
(a)	*Word problem on proportions (Type 1)*	A car travels on the freeway at an average speed of 52 miles per hour. How many miles does it travel in 5 hours and 30 minutes?
(b)	*Plotting a point in the coordinate plane*	Using the pencil, mark the point at the coordinates $(1, 3)$.
(c)	*Multiplication of monomials*	Perform the following multiplication: $4x^4y^4 \cdot 2x \cdot 5y^2$ and simplify your answer as much as possible.
(d)	*Greatest common factor of two monomials*	Find the greatest common factor of the expressions $14t^6y$ and $4tu^5y^8$. Simplify your answer as much as possible.
(e)	*Graphing the line through a given point with a given slope*	Graph the line with slope -7 passing through the point $(-3, -2)$.
(f)	*Writing the equation of the line through a given point and perpendicular to a given line*	Write an equation for the line that passes through the point $(-5, 3)$ and is perpendicular to the line $8x + 5y = 11$.

For concreteness, we consider a particular situation in which the assessment is computer driven and the problems are presented on a monitor, via the Internet. All the virtual tools needed for providing the answers to the test—pencil, ruler, graphical displays, calculators of various kinds when deemed necessary—, are part of the interface. In the course of a tutorial, the testees have been familiarized with these tools. In Problems (b) and (e), a coordinate plane is displayed on the computer monitor as part of the question, and the pencil and, for Problem (e), also the ruler, are provided. In this problem, the student must graph the line using the virtual pencil and ruler. We also suppose that all the problems have open responses (i.e. no multiple choice), and that 'lucky guesses' are unlikely. (Careless errors are always possible, of course, and a clever assessment procedure has to guard against them.)

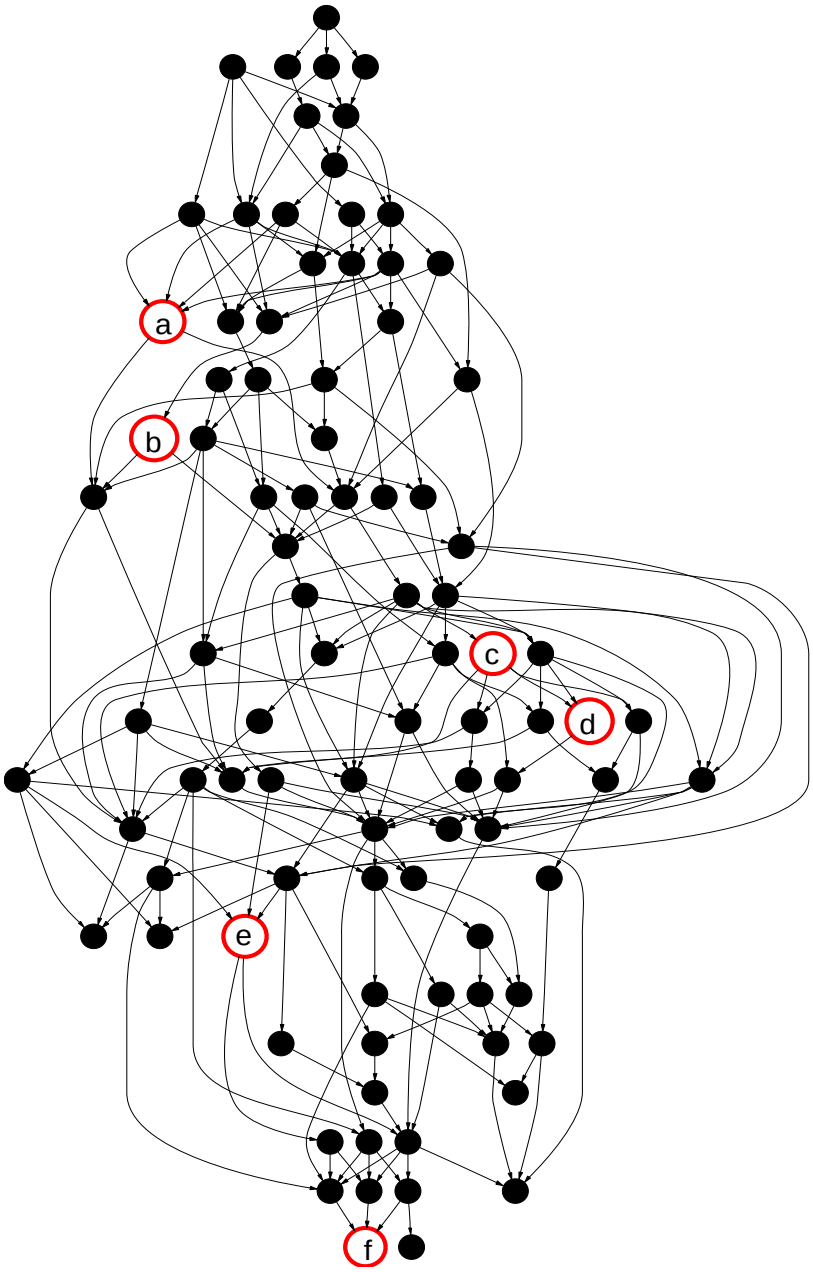

Fig. 2. Diagram of the precedence relation for Beginning Algebra. The vertices marked a-f refer to Problems (a)-(f) of Fig. 1, whose diagram may be inferred from the one above.

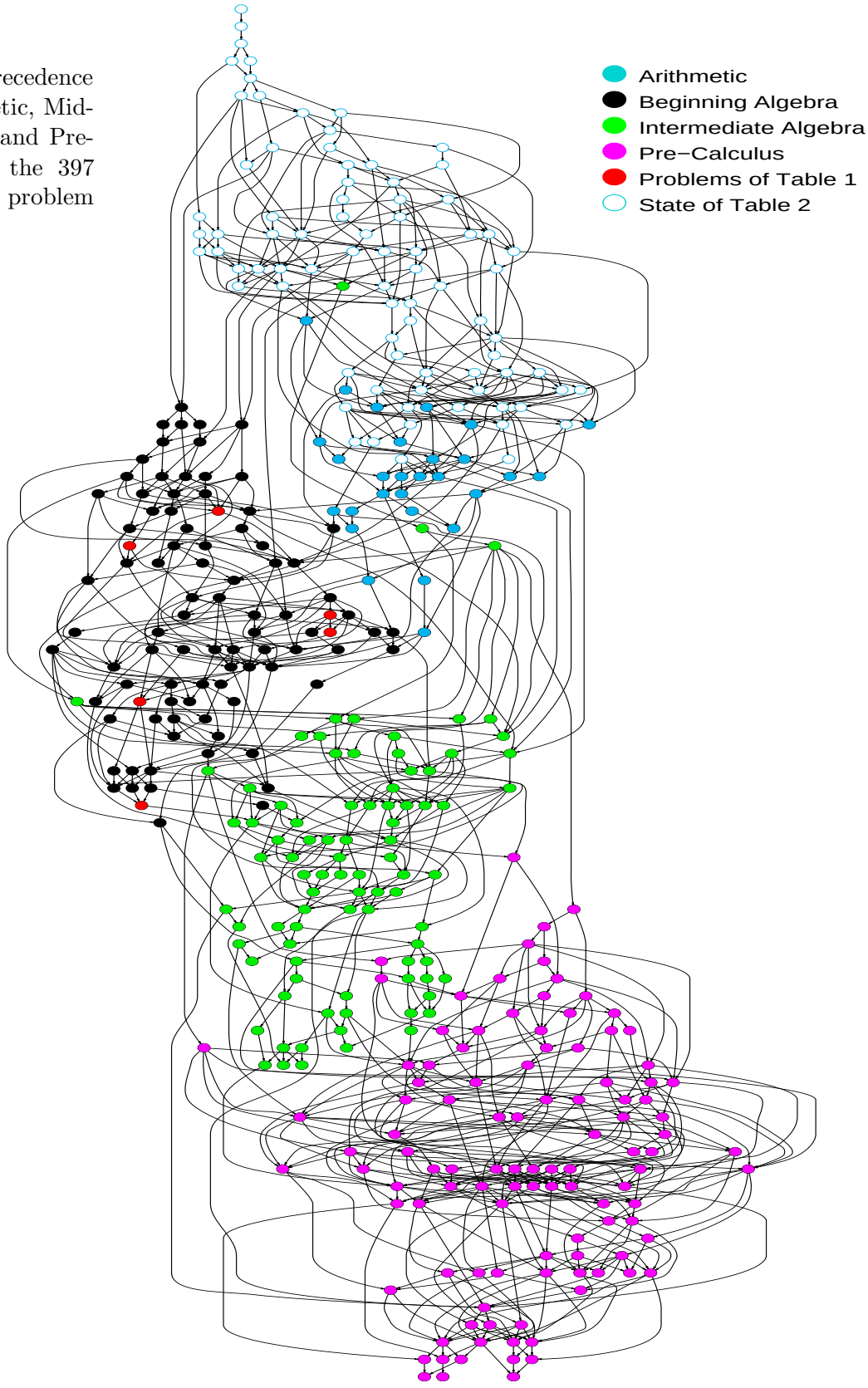

Fig. 3. Combined precedence diagram for Arithmetic, Middle School Algebra, and Pre-Calculus. Each of the 397 points represents a problem type.

We postpone for the moment the discussion of how to construct a valid precedence diagram for a realistically large problem set. (For example, how were the precedence diagrams of Figs. 2 or 3 obtained?) This question and other critical ones are considered later on in this article. For the time being, we focus on the miniature example of Table 1 which we use to introduce and illustrate the basic ideas.

The knowledge states. The precedence diagram of Fig. 1 completely specifies the feasible knowledge states. The respondent can certainly have mastered just Problem **a**: having mastered **a** does not imply knowing anything else. But if he or she knows **e**, for example, then **a**, **b** and **c** must also have been mastered, forming a knowledge state which we represent as the set of problems $\{\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{e}\}$ or more compactly **abce**. Analyzing carefully the precedence diagram of Fig. 1, we see that there are exactly 10 knowledge states consistent with it, forming the set

$$\mathcal{K} = \{\emptyset, \mathbf{a}, \mathbf{b}, \mathbf{ab}, \mathbf{ac}, \mathbf{abc}, \mathbf{abcd}, \mathbf{abce}, \mathbf{abcde}, \mathbf{abcdef}\},$$

where $\emptyset$ symbolizes the empty state: the respondent is unable to solve any of the 6 problems. The set $\mathcal{K}$ is our basic concept, and is called the *knowledge structure*. Note that a useful knowledge structure is not necessarily representable by a precedence diagram such as those of Figs. 1, 2 or 3 and may simply be specified by the collection of knowledge states.

The learning paths. This knowledge structure allows several *learning paths*. Starting from the naive state $\emptyset$, the full mastery of state **abcdef** can be achieved by mastering first **a**, and then successively the other problems in the order $\mathbf{b} \mapsto \mathbf{c} \mapsto \mathbf{d} \mapsto \mathbf{e} \mapsto \mathbf{f}$. But there are other possible ways to learn. All in all, there are 6 possible learning paths consistent with the knowledge structure $\mathcal{K}$, which are displayed in Fig. 4.

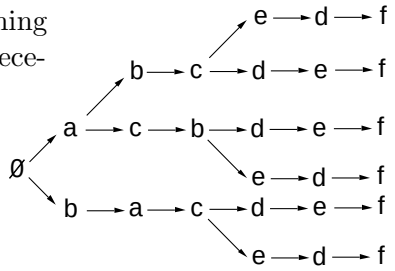

Fig. 4. The 6 possible learning paths consistent with the precedence diagram of Fig. 1.

In realistic knowledge structures such as those for Arithmetic or Elementary Algebra, the numbers of feasible knowledge states and of learning paths become very large. In the case of Beginning Algebra, whose precedence diagram was given in Fig. 2, there are around $60,000$ knowledge states and literally billions of feasible learning paths. These numbers may be puzzling. Where is the diversity coming from? After all, these mathematical subjects are highly structured and typically taught in roughly the same sequence. However, even though the school curriculum may be more or less standard, learning the material, and also forgetting it, follows their own haphazard course. Besides, $60,000$ states form but a minute fraction of the 2^{88} possible subsets of the set of 88 problems. In any event, it is clear that, even in a highly structured mathematical topic, an accurate assessment of knowledge involves sorting out a considerable array of possibilities.

The outer and inner fringes of a knowlegde state. As suggested by the precedence diagrams and by the learning paths of Fig. 4, the knowledge structures considered here have the property that learning can take place step by step, one problem type at a time. More precisely, each knowledge state (except the top one) has at least one *immediate successor* state, that is, a state containing all the same problems, plus exactly one. The knowledge state **abc** of $\mathcal{K}$, for instance, has the two states **abcd** and **abce** as immediate successors. Problems **d** and **e** form the 'outer fringe' of state **abc**. In general, the *outer fringe* of some knowledge state K is the set of all problems **p** such that adding **p** to K forms another knowledge state. The concept of outer fringe is critical because this is where progress is taking place: learning proceeds by mastering a new problem in the outer fringe, creating a new state, with its own outer fringe.

Conversely, each knowledge state (except the empty state) has at least one *predecessor state*, that is a state containing exactly the same problems, except one. The knowledge state **abc** that we just considered has two predecessor states, namely **ab** and **ac**. Problems **b** and **c** together form the *inner fringe* of state **abc**: removing either **b** or **c** from state **abc** creates other states in the structure, that is **ab** and **ac**. If for some reason a student experiences difficulties in mastering the problems in the outer fringe, reviewing previous material should normally take place in the inner fringe of a student's state. Figure 5 illustrates these concepts of fringes and others introduced so far. A state K is pictured with three problems in its outer fringe. Another state K' has two problems in its inner fringe.

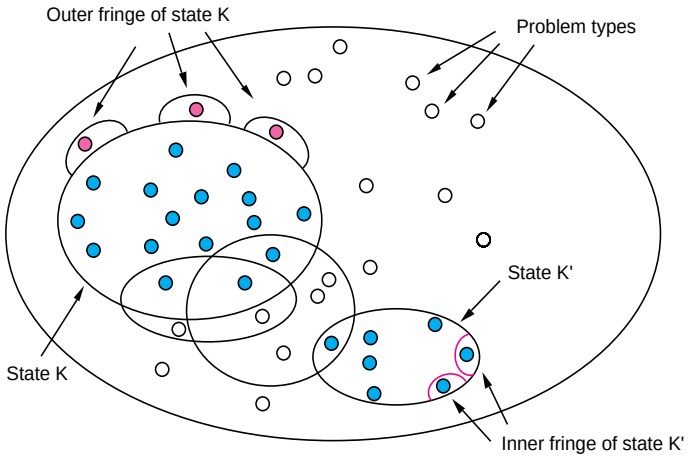

Fig. 5. The outer fringe of a state K and the inner fringe of another state K'.

Thus, we can use the two fringes as the main building blocks of the 'navigation tool' of the system, with the outer fringes directing the progress, and the inner fringes monitoring temporary retreats, and making them profitable.

Interestingly, the fringes also play a less obvious, but equally important role in summarizing the results of an assessment. A knowledge state is essentially a list of all the problems mastered by a student at the time of an assessment. Such a list will often be unwieldy and contain several dozen problem names, not a very convenient description. It can be shown

mathematically, however, that for the most useful kinds of knowledge structures, the two fringes suffice to specify the knowledge state completely. In other words, the result of an assessment can be given in the form of two short lists, one for the inner fringe (WHAT THE STUDENT CAN DO, which is understood here as the most sophisticated problems in the student's state), and one for the outer fringe (WHAT THE STUDENT IS READY TO LEARN). Experience with realistic knowledge structures in school mathematics has shown that these two lists together will contain on average 11 problems, enabling a very compact and faithful presentation of the result of an assessment.

Table 2 contains a typical example of the two fringes of a knowledge state, which is that of an actual student currently using the system in a middle school. Taken together, the two fringes amount to 9 problems, which together suffice to specify the 80 problems of that student's state which is represented in the top region of Fig. 3. The economy is startling.

Table 2: A knowledge state in Arithmetic specified by its two fringes

Inner fringe: WHAT THE STUDENT CAN DO
Double negation: $-(-12) - 7 =$
Arithmetic with absolute value: $\Big\lvert\, \lvert 9 - 12\rvert - \lvert 5\rvert \,\Big\rvert$
Word problem with clocks: A clock runs too fast and gains 6 minutes every 5 days. How many minutes and seconds will it have gained at the end of 9 days?
Word problem on percentage (Problem type 1): A pair of sneakers usually sells for \$45. Find the sale price after a 20% discount.
Mixed number multiplication: $3\frac{3}{4} \times 2\frac{4}{9} =$ (Write your answer as a mixed number in lowest terms.)

Outer fringe: WHAT THE STUDENT IS READY TO LEARN
Decimal division: $5.2\overline{)7.54}$
Word problem on percentage (Problem type 2): A sofa is on sale for \$630 after a 28% discount. What was the price before discount?
Word problem with inverse proportion: If 4 machines are needed to complete a task in 21 days, how long will it take 7 machines to complete the same task?
Average of two numbers: What is the average value of 114 and 69?

The information provided by such a table is a more meaningful result of an assessment than a couple of numerical scores from a standardized test. It is also considerably more precise. An assessment involving all of high school mathematics, based on the knowledge states consistent with the precedence diagram of Fig. 3, would classify the students in hundreds of thousands of categories, each with its own unique table of inner and outer fringes. By contrast, a quantitative S.A.T. classifies the test taker into one of roughly 40 categories (from 400 to 800, in steps of 10).

Building a knowledge structure

We now turn to what is certainly the most demanding task in a realistic application of these ideas. It certainly makes sense to enroll experts, such as seasoned teachers or textbook writers, to find the knowledge states. This must be done at least for the first draft of a knowledge structure, which can then be refined by a painstaking analysis of student data. However, we cannot simply sit an expert in front of a computer terminal with the instruction: "provide a complete list of all the knowledge states in a given topic." Fortunately, an indirect approach is possible. An expert can reliably respond to questions such as these:

> **Q1.** Suppose that a student is not capable of solving problem $\mathbf{p}$. Could this student nevertheless solve problem $\mathbf{p}'$?

It can be proven that a knowledge structure represented by a precedence diagram such as the one of Fig. 2 can be inferred exactly from the responses to a complete collection of questions of the type **Q1**. (For a very large precedence diagram, such as the one of Fig. 3, several diagrams are first constructed by querying experts on each of the fields of knowledge, like Arithmetic, Beginning Algebra, *etc.* Those diagrams are then 'glued' together, relying again on experts' judgment.)

In the case of the precedence diagram of Fig 1, the mastery of problem **e**, for instance, implies that of a single minimum set of precedent problems, namely **a**, **b** and **c**. In other words, all learning paths in Fig. 4 progress through these three problems before reaching **e**. There are important cases, however, in which the mastery of a problem may be achieved via anyone of several distinct minimum sets of precedent problems. Such structures, which generalize those that can be represented by precedence diagrams, are called *knowledge spaces*. They are derived from the responses to the collection of more difficult questions of the following type:

> **Q2.** Suppose that a student has not mastered problems $\mathbf{p_1}, \mathbf{p_2}, \ldots, \mathbf{p_n}$. Could this student nevertheless solve problem $\mathbf{p}'$?

In practice, not all questions of type **Q1** or **Q2** must be asked because, in many cases, responses to some questions can be inferred from responses to other questions. For typical knowledge structures encountered in education, an expert may be required to respond to a few thousand questions to get a complete description of all the knowledge states.

By interviewing several experts and combining their answers, one can build a knowledge structure which reflects their consensual view of the field. This alone does not guarantee the validity of the knowledge structure, that is, the agreement between the states in the structure and the actual states in the student population. Actual student data are also needed. With an Internet based, largely distributed assessment system such as the one discussed here, data from several thousand users can be collected in the span of a year, providing a bounty of information. Such data can be used to refine a knowledge structure obtained from experts' judgments via the questions of type **Q1** or **Q2**. To begin with, states occurring rarely or not at all in the empirical applications can be deleted from the knowledge structure. More importantly, the accuracy of the structure can be evaluated by the following probe, and corrected if necessary. In most assessments, an extra problem $\mathbf{p}^*$

is added to the questioning, which is not used in the choice of the final knowledge state K representing the student. Using K, one can predict the student answer to $\mathbf{p}^*$ which should be correct if $\mathbf{p}^*$ is in K—except for careless errors—and false otherwise. In the knowledge structure for Beginning Algebra for example, as it is used by students today, the correlation between predicted and observed answers hovers between .7 and .8, depending on the sample of students. These high values actually *underestimate* the accuracy of the structure: a student having mastered some problem $\mathbf{p}^*$ contained in his or her knowledge state may nevertheless make a careless error in solving it. This correlation index is a powerful statistical tool continuously monitoring the validity of the knowledge structure, pointing to weaknesses, and evaluating the corrections prompted by some earlier analysis.

Uncovering a knowledge state in a knowledge structure

Suppose that a satisfactory knowledge structure has been obtained. The task of the assessment is to uncover, by efficient questioning, the knowledge state of a particular student under examination. The situation is similar to that of *adaptive testing*—i.e. the computerized forms of the S.A.T. and the like—with the critical difference that the outcome of the assessment here is a knowledge state, rather than a numerical estimate of a student's competence in the topic.

The assessment procedures available all pertain to the scheme outlined in Fig. 6.

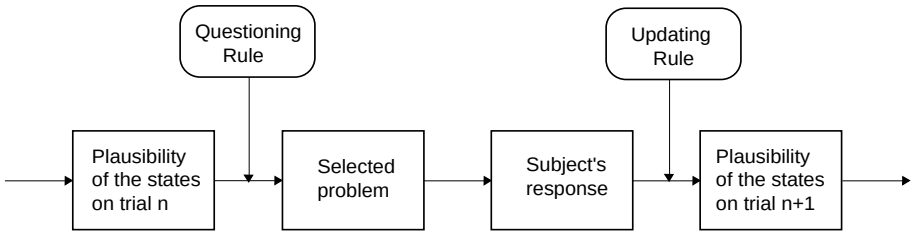

Fig. 6. Diagram of the transitions in an assessment procedure.

In this article, we focus on one particular assessment procedure in which the plausibility of a state is its current likelihood, based on all the information accumulated so far. At the outset of the assessment (trial 1 of the procedure), each of the knowledge states is assigned a certain *a priori* likelihood, which may depend upon the school year of the student if it is known, or some other information. The sum of these *a priori* likelihoods is equal to 1. They play no role in the final result of the assessment but may be helpful in shortening it. If no useful information is available, then all the states are assigned the same likelihood. The first problem $\mathbf{p_1}$ is chosen so as to be 'maximally informative.' This is interpreted to mean that, on the basis of the current likelihoods of the states, the student has about a 50% chance of knowing how to solve $\mathbf{p_1}$. In other words, the sum of the likelihoods of all the states containing $\mathbf{p_1}$ is as close to .5 as possible. If several problem types are equally informative (as may happen at the beginning of an assessment) one of them is chosen at random. The student is then asked to solve an instance of that problem, also picked randomly. The student's answer is then checked by the system, and the likelihood

of all the states are modified according to the following *updating rule*. If the student gave a correct answer to $\mathbf{p_1}$, the likelihoods of all the states containing $\mathbf{p_1}$ are increased and, correspondingly, the likelihoods of all the states *not* containing $\mathbf{p_1}$ are decreased (so that the overall likelihood, summed over all the states, remains equal to 1). A false response given by the student has the opposite effect: the likelihoods of all the states *not* containing $\mathbf{p_1}$ are increased, and that of the remaining states decreased. If the student does not know how to solve a problem, he or she can choose to answer "I don't know" instead of guessing. This results in a substantial increase in the likelihood of the states not containing $\mathbf{p_1}$, thereby decreasing the total number of questions required to uncover the student's state. Problem $\mathbf{p_2}$ is then chosen by a mechanism identical to that used for selecting $\mathbf{p_1}$, and the likelihood values are increased or decreased according to the student's answer via the same updating rule. Further problems are dealt with similarly. In the course of the assessment, the likelihood of some states gradually increases. The assessment procedure stops when two criteria are fulfilled: (1) the entropy of the likelihood distribution, which measures the uncertainty of the assessment system regarding the student's state, reaches a critical low level, and (2) there is no longer any useful question to be asked (all the problems have either a very high or a very low probability of being responded to correctly). At that moment, a few likely states remain and the system selects the most likely one among them. Note that, because of the stochastic nature of the assessment procedure, the final state may very well contain a problem to which the student gave a false response. Such a response is thus regarded as due to a careless error. On the other hand, because all the problems have open-ended responses (no multiple choice), with a large number of possible solutions, the probability of lucky guesses is negligible.

To illustrate the evolution of an assessment, we use a graphic representation in the guise of the *likelihood map* of a knowledge structure. In principle, each colored point in the oval shape of Fig. 7 represents one of the 57,147 states of the knowledge structure for Arithmetic. (Because of graphics limitations, some grouping of similar states into a single point was necessary. To simplify the exposition, we suppose in the sequel that each point of the map represents one state.) The precedence diagram of this structure was given in Fig. 3.

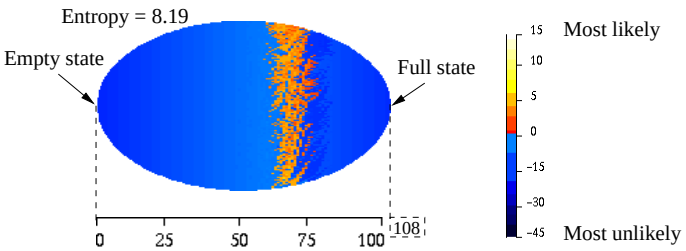

Fig. 7. Likelihood map of the Arithmetic knowledge structure whose precedence diagram was given in Fig. 3.

Knowledge states are sorted according to the number of problem types they contain, from 0 problems on the far left to 108 problems on the far right. The leftmost point stands for the empty knowledge state, which is that of a student knowing nothing at all in Arithmetic. The rightmost point represents the full knowledge state and corresponds to a student having mastered all the problems in Arithmetic. The points located on any vertical line within the oval represent knowledge states containing exactly the number of problems indicated on the abscissa. The oval shape is chosen for esthetic reasons and reflects the

fact that, by and large, there are many more states around the middle of the scale than around the edges. For instance, there are 1,668 states containing exactly 75 problems, but less than 100 states, in Arithmetic, containing more than 100 problems or less than 10 problems. The arrangement of the points on any vertical line is largely arbitrary.

The color of a point represents the likelihood of the corresponding state. A color coded logarithmic scale, pictured on the right of Fig. 7, is used to represent the likelihood values. Red, orange, and yellow-white indicate states that are more likely than average, with yellow-white marking the most likely states. Similarly, dark blue, blue, and light blue represent states that are less likely than average, with dark blue marking the least likely states.

Figure 8 displays a sequence of likelihood maps describing the evolution of an assessment in Arithmetic from the very beginning, before the first problem, to the end, after the response to the last problem is recorded by the system and acted upon to compute the last map. The full assessment took 24 questions which is close to the average for Arithmetic. The initial map results from preliminary information obtained from that student. The redish strip of that map represents the *a priori* relatively high likelihood of the knowledge states containing between 58 and 75 problems: as a six grader, this student can be assumed to have mastered about two thirds of the Arithmetic curriculum.

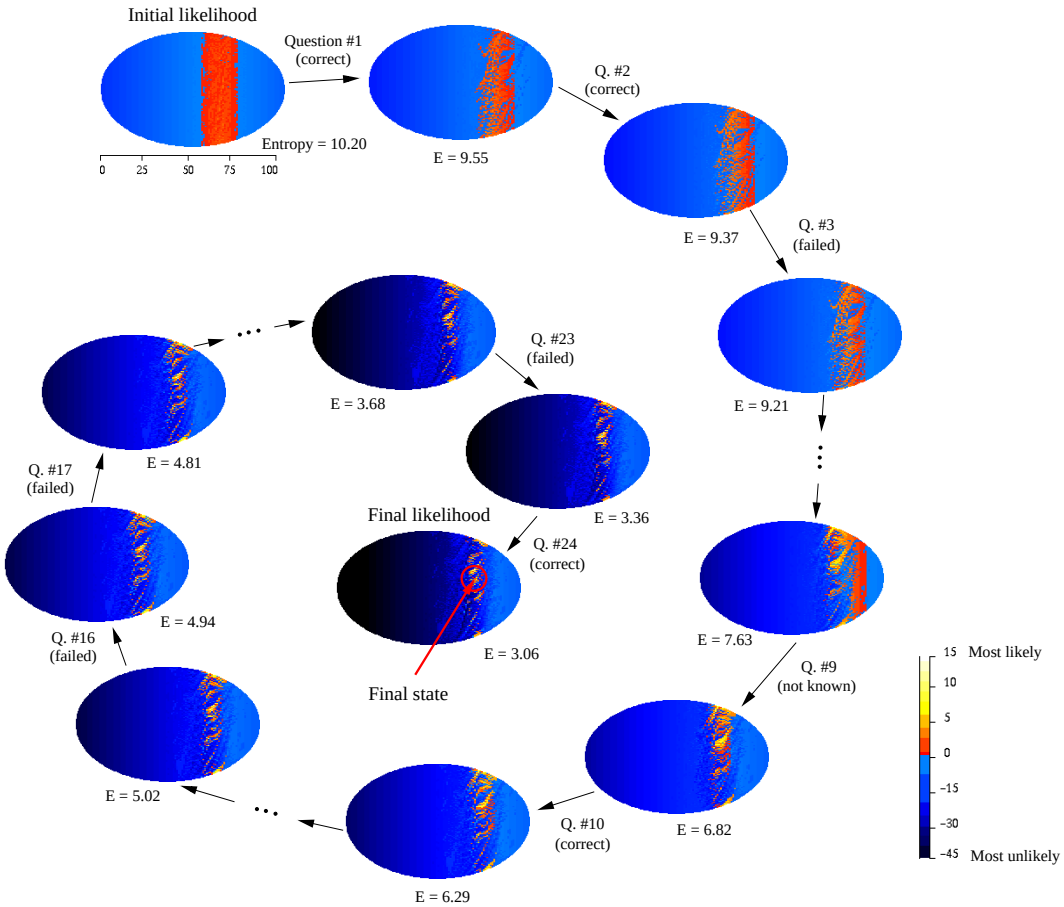

Fig. 8. Sequence of likelihood maps representing an assessment converging toward the student's knowledge state.

Next to each map in Fig. 8, we indicate the entropy of the corresponding likelihood distribution, and the student's response to the question (correct, false, or not known). Note that the initial entropy is 10.20, which is close to the theoretical maximum of 10.96 obtained for a uniform distribution on a set of 57,147 knowledge states. As more information is gathered by the system via the student's responses to the questions, the entropy decreases gradually. Eventually, after 24 questions have been answered a single very bright point remains among mostly dark blue points and a few bright points. This very bright point indicates the most likely knowledge state for that student, based on the answers to the problems. The assessment stops at that time because the entropy has reached a critical low level and the next 'best' problem to ask has only a 19% chance of being solved by the student, and so would not be very informative. In this particular case only 24 problems have sufficed to pinpoint the student's knowledge state among 57,147 possible ones. This striking efficiency is achieved by the numerous inferences implemented by the system in the course of the assessment.

The assessment procedure described in this article is the core engine of an Internet based, automated mathematics tutor which is used in several hundred colleges and school districts in the U.S. Numerous data indicate that learning is very efficient, which must be attributed to the precision of the assessment: teaching is always on target, in the outer fringe of a student's state. In the U.S., the extensive research leading to this system has been supported since 1983 by various grants, mostly from the National Science Foundation. The first paper on this research topic, which is named 'Knowledge Spaces,' was published in 1985 by J.-P. Doignon and J.-Cl. Falmagne, two of the authors of this article. Important results have also been obtained by researchers from other countries, such as D. Albert (Austria), C. Dowling (Germany) and M. Koppen (The Netherlands). Most of the results are presented in a monograph entitled *'Knowledge Spaces,'* by Doignon and Falmagne [1999].

References

J.-P. Doignon and J.-Cl. Falmagne. *Knowledge Spaces.* Springer, Berlin, 1999.

W.T. Kelvin. *Popular Lectures and Addresses (in 3 volumes)*, volume (Vol. 1: *Constitution of Matter*, Chapter *Electrical Units of Measurement.*). MacMillan, London, 1889.

K. Pearson. *The Life, Letters and Labours of Francis Galton*, volume (Vol. 2: *Researches of Middle Life.*). Cambridge University Press, London, 1924.

F.S. Roberts. Measurement theory with applications to decision making, utility and the social sciences. In G.-C. Rota, editor, *Encyclopedia of Mathematics and its Applications*, volume 7: Mathematics and the Social Sciences. Addison-Wesley, Reading, Ma., 1979.

W.T. Trotter. *Combinatorics and Partially Ordered Sets: Dimension Theory.* The Johns Hopkins University Press, Baltimore, Maryland, 1992.