Assessing Mathematical Knowledge in a Learning Space: Validity and/or Reliability*

Jean-Claude Falmagne† jcf@uci.edu

Eric Cosyn‡ ecosyn@aleks.com

Christopher Doble‡ Nicolas Thiéry‡ Hasan Uzun‡

{cdoble, nthiery, huzun}@aleks.com

Abstract

As implemented in the system discussed here, the assessment of knowledge in a learning space for a scholarly topic, such as beginning algebra, is comprehensive by design, in that the types of problems that can be asked in any assessment come from a large collection encompassing the full curriculum for the topic. The product of an assessment is a knowledge state gathering all the types of problems that the student is capable of solving. Typically, the number of feasible knowledge states is large, on the order of 10^7. The duration of an assessment is nevertheless tolerable, ranging around $30-35$ problems. We summarize the basic concepts underlying learning spaces and report the results of a large scale study (210,102 assessments) investigating whether such an assessment is predictive of the subject's responses to problems that are not part of the assessment. In each assessment, an additional question was asked, the response to which is predictable from the assessed state. The mean correlation between predicted and observed responses (correct or false) was around .67, and the mean log odds ratio 2.75. This type of analysis resembles the standard item-test correlations computed for the evaluation of psychometric instruments. The essential technical and philosophical differences between the two approaches are discussed.

Keywords: knowledge representation, adaptive assessment, artificial intelligence, human-computer interaction, mathematical modelling, learning, statistics

*We are grateful to Brian Junker for his comments to a previous draft of this paper. We also thank Jonathan Nguyen for extending the application of a routine for the computation of the tetrachoric coefficient to several matrices.

†Corresponding author. Dept. of Cognitive Sciences, University of California, Irvine, CA 92697. Phone: (949) 433 2735; e-mail: jcf@uci.edu.

‡ALEKS Corporation.

A. Background and Introduction

A learning space is a particular kind of 'knowledge space.' In essence, it is a family of sets of problems, called *knowledge states*. The structure of this family is such that learning can proceed smoothly and consistently, one type of problem at a time. The exact definition is recalled in Section B, in the form of two pedagogically compelling axioms labelled [K1] and [K2]. Many aspects of these structures have been investigated in detail, and the results were reported in various publications (see in particular, Doignon and Falmagne, 1985; Falmagne and Doignon, 1988; Albert and Lukas, 1999; Doignon and Falmagne, 1999; Falmagne et al., 2006)[1]. A learning space is known in the combinatorics literature as an antimatroid, a structure introduced by Edelman and Jamison (1985) (cf. also Welsh, 1995; Björner et al., 1999).

Assessing knowledge in a learning space contrasts with applying a standardized test, whose aim is to obtain a numerical score indicative of a degree of competence in some broad topic according to the principles of psychometric theory (c.f., for example, Nunnally and Bernstein, 1994). In the standardized test case, the issues of the reliability and, especially, the validity of the measurement, are legitimate concerns. We have a different situation in the case of a learning space because the set of all the questions potentially used in any assessment is in principle designed to represent a fully comprehensive coverage of a particular curriculum (such as beginning algebra, or possibly all of K-12 mathematics). Evidently, one may question the comprehensiveness of the domain of the learning space, and object to the presence or absence of some problem types. The domain can be verified by experts, and corrected if need be. One may also want to evaluate the extent to which the knowledge state resulting from the assessment is predictive of the student's responses to problems not used in the assessment. The subject of this paper is just such an evaluation. We go back to the critical issue of measuring the validity of an assessment in a learning space in our discussion section.

To avoid any misunderstanding, a specification regarding our vocabulary is useful. In the sequel, we use *'problem'* to mean 'problem type', and we call *'instance'* a particular case of a problem, obtained by randomly choosing the numbers involved in the problem, and also the 'story line' in the case of a word problem. For example, the problem entitled WORD PROBLEM WITH DECIMAL OPERATIONS: TYPE 1 currently has five story lines, one of them yielding an instance such as:

[1] An extensive database on knowledge spaces is maintained by Cord Hockemeyer at the University of Graz: http://wundt.uni-graz.at/kst.php.

Abdul works mowing lawns and rakings. He earns $5.40 an hour for mowing and $4.40 an hour for raking. How much will he earn for 5 hours of mowing and 1 hour of raking?

The dollar amounts are multiples of $0.10 and are chosen in specified intervals possibly different for each of the story lines, integer amounts being excluded. Counting the five story lines and all the numerical cases for each of them, we end up with about 28,000 instances for this particular problem. Giving this problem to a subject in the course of an assessment results in the random choice of one among these 28,000 instances. Thus, an instance in our sense corresponds to an item in standardized testing.

The final product of an assessment is the knowledge state that contains all those problems, and only those problems, that the student is capable of solving, barring careless errors. (In the application of learning spaces considered here, there are no lucky guesses since all the questions have either open responses or multiple choices with a large number of possible responses.) Such a knowledge state is one among many that are feasible. Typically—say, in arithmetic or beginning algebra—the number of feasible knowledge states in a learning space is on the order of 10^7. The duration of an assessment is nevertheless tolerable, ranging around 30-35 problems in most cases. This paper describes the results of a large scale study, based on 210,102 assessments in a learning space for beginning algebra. In each of these assessments, an *additional problem* $\boldsymbol{p}$ was randomly selected in a uniform distribution on the set of all the problems, and an instance of that problem, also randomly chosen, was given to the subject at some point during the assessment. The response to that problem was not taken into account in gauging the knowledge state. However, as the knowledge state is a complete description of someone's mastery in the relevant topic, a prediction can be made of that person's mastery of any problem in the domain. So, the correlation between the actual response of the person to problem $\boldsymbol{p}$ and the prediction derived from the person's knowledge state obtained from the assessment can be estimated by standard indices. We report here the results of such computations for the case of a learning space for beginning algebra, which has a database containing 250 problems. Note that in some cases, another instance of the additional problem $\boldsymbol{p}$ may also be presented to the student as part of the assessment. All such cases have been discarded for this study.

The next section recalls and discusses the axioms of a learning space and some of their consequences. Section C summarizes and illustrates the basic mechanism of the assessment instrument, the mathematical underpinnings of which have been fully described elsewhere (see e.g. Falmagne and Doignon, 1988; Doignon and Falmagne, 1999). The correlation data

and the issue of the validity/reliability of the assessment are covered in Section D. The paper ends, in Section E, with a summary and a discussion.

B. Learning Spaces

Consider a broad topic in mathematics education, such as beginning algebra (sometimes called "Algebra 1"). From the standpoint of assessing the students' competence, this topic can be delineated by a finite set Q of problems that a student must master, together with the relevant concepts. The set Q is intended to be a fully representative coverage of the curriculum. An examination of a sample of textbooks indicates that, for beginning algebra, the set Q contains between two and three hundred problems. A pair $(Q, \mathcal{K})$ is a *knowledge structure* if $\mathcal{K}$ is a family of subsets of Q containing all the *knowledge states* that are feasible, that is, that could characterize some individual in a particular population under consideration. In other words, an individual whose knowledge state is K can, in principle, solve all the problems in K and would fail any problem not in K. It is assumed that the family $\mathcal{K}$ contains both the empty set and the total set Q of problems: we conceive that it is possible for someone to know nothing in Q, and for someone else to know everything in Q. Because algebra is a highly structured topic, $|\mathcal{K}|$ is considerably smaller than 2^{250}, the number of subsets in a set of size 250. Typically[2], $|\mathcal{K}|$ is on the order of 10^7, which is well within the capabilities of personal computers. Note that, in the sequel, we often say 'state' to mean 'knowledge state.'

Axioms. Further constraints are imposed on $\mathcal{K}$ in the form of the two axioms recalled below (cf. Falmagne et al., 2006; Cosyn and Uzun, 2006), making the pair $(Q, \mathcal{K})$ a *learning space*.

[K1] If $K \subset L$ are two knowledge states in $\mathcal{K}$, with $|L \setminus K| = n$, then there is a chain of states $K_0 = K \subset K_1 \subset \cdots \subset K_n = L$ such that $K_i = K_{i-1} \cup \{q_i\}$ with $q_i \in Q$ for $1 \leq i \leq n$.

In words, intuitively: *If the state K of the learner is included in some other state L then there is a sequence of problems $q_1, \ldots, q_n$ that are learnable one at a time, leading the individual from state K to state L.*

[K2] If $K \subset L$ are two knowledge states in $\mathcal{K}$, with $q \notin K$ and $K \cup \{q\} \in \mathcal{K}$ for some problem q, then $L \cup \{q\} \in \mathcal{K}$.

In words: *If problem q is learnable from state K, then it is learnable from any state L including K.* Or: knowing more does not make a student less capable to learn something new.

[2]For elementary topics in mathematics or science, for example.

In studying these axioms, the exact interplay between the mathematics and the pedagogy must be understood clearly. By themselves, Axioms [K1] and [K2] do not guarantee that the collection of problems is learnable. For these axioms to be useful, they must express a pedagogical reality. A basic idea is that for a student to be able to master any problem q, there must some state K that can be reached by the student, such that q can be mastered from state K; thus, $K \cup \{q\}$ must also be a state. Axioms [K1] and [K2] generalize and systematize this idea.

Remark[3]**.** To understand what these axioms mean and do not mean, it is useful to realize that they are equivalent to two quite different looking conditions, which are much less obvious pedagogically. Cosyn and Uzun (2006) have shown that Axioms [K1] and [K2] are satisfied by a knowledge structure $(Q, \mathcal{K})$ if and only if the two axioms below also hold:

[K1*] The family $\mathcal{K}$ is *well-graded*, that is, if K and L are any two distinct states differing by exactly n problems, then there exists a sequence of states $K_0 = K, K_1, \ldots, K_n = L$ such that, for $0 \leq i < n$, the two states K_{i+1} and K_i differ by exactly one problem: either $K_{i+1} = K_i \cup \{q\}$ or $K_i = K_{i+1} \cup \{q\}$ for some problem q.

[K2*] The family $\mathcal{K}$ is *closed under union*, that is, if K and L are any two states, then $K \cup L$ is also a state.

Many results can be derived from these axioms, which are described in detail in Doignon and Falmagne (1999). One important consequence of [K1] and [K2] is recalled informally below (Fringe Theorem), which is essential from an educational standpoint. It relies on a crucial pair of concepts.

The two fringes of a knowledge state. The *outer fringe* of a state K is the set containing all the problems q not in K such that $K \cup \{q\}$ is also a state. Thus, the outer fringe of a student's state contains all the problems that the student is ready to learn. The *inner fringe* of a state K is the set of all the problems q such that $K \setminus \{q\}$ is also a knowledge state. In other words, the inner fringe contains all the problems representing the 'high points' of the student's competence.

Fringe Theorem. *The knowledge state of a student is defined by the inner fringe and the outer fringe of the state. Thus, if the results of an assessment are summarized in the form of the two fringes of a state, the state is completely specified* (Doignon and Falmagne, 1999, Theorem 2.8, (i) $\Rightarrow$ (v)).

[3]This remark can be skipped without loss of continuity.

Tables 1a and 1b contain an actual example of the two fringes of a knowledge state in beginning algebra, with each problem being represented by one instance. Taken together, the two fringes amount to 14 problems. This suffices to specify the 134 problems contained in that student's knowledge state. The economy is notable. Moreover, the summary is meaningful for an instructor.

Table 1: A KNOWLEDGE STATE IN BEGINNING ALGEBRA SPECIFIED BY ITS TWO FRINGES

Table 1a. Outer fringe (9 problems): WHAT THE STUDENT IS READY TO LEARN.
Word problem with linear inequalities: The sum of two numbers is less than or equal to 13. The second number is 5 less than the first. What are the possible values for the first of the two numbers?
Solving a rational equation that simplifies to a linear equation (Type 1): Solve for u: $-6 = -\frac{8}{u}$.
Word problem on mixed number proportions: A chocolate chip cookie recipe requires one and one quarter cups of flour to one cup of chocolate chips. If two and one half cups of flour is used, what quantity of chocolate chips will be needed?
Y-intercept of a line: Find the y-intercept of the line whose equation is $y = \frac{17}{15}x - \frac{5}{4}$.
Multiplying polynomials: Multiply and simplify: $(6z + 6w - 1)(5z + 3w - 3)$.
Word problem on inverse proportions: Suppose that 8 machines can complete a given task in 5 days. If there were 10 machines, how many days would it take for them to finish the same task?
Word problem on percentage (type 3): The price of a gallon of gas has risen to \$2.85 today. Yesterday's price was \$2.79. Find the percentage increase. Round your answer to the nearest tenth of a percent.
Area and perimeter of a rectangle: The length of a rectangle is twice its width. If the area of the rectangle is 162 ft^2, find its perimeter.
Union and intersection of sets: The sets F and A are defined by $F = \{x \mid x \text{ is an integer and } -4 < x \leq 0\}$, $A = \{x \mid x \text{ is an integer and } 1 < x \leq 3\}$. Find $F \cup A$ and $F \cap A$.

Table 1b. Inner fringe (5 problems) : WHAT THE STUDENT CAN DO (HIGH POINTS).
Power rule—Positive exponents: Write without parentheses: $(\frac{y^3}{-2x})^3$.
Squaring a binomial: Expand the square: $(6x-6)^2$
Properties of real numbers: For each equation below, indicate the property of real numbers that justifies the equation: $\frac{3}{4}+(m+b) = (\frac{3}{4}+m)+b$ $7\cdot\frac{1}{7} = 1$ $0 = 4+(-4)$ $m(\frac{3}{5}+7) = m\cdot\frac{3}{5}+m\cdot 7$
Solving a linear inequality (type 4): Solve for t: $-\frac{7}{2}t+9>-8t-3$.
Writing a negative number without a negative exponent: Rewrite without an exponent: $(-3)^{-1}$.

It was suggested by some readers of a previous draft that the outer fringe of a state could be seen as a formal implementation of the concept of 'zone of proximal development' (ZPD) in the sense of Vygotsky (1978) (for a recent reference, see Chaiklin, 2003). This remark is particularly apt because a strict reading of the literature on ZPD indicates that it explicitly includes the intervention of external agents, such as human teachers or other students (via cooperative problem solving), for the zone of proximal development to be conquered. In fact, some teachers use the outer fringe of students to select the subset of the class that is prepared for a pointed lecture on a particular topic.

C. Uncovering the Knowledge State: The Assessment Engine

The task of the assessment engine 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 GRE and other standardized tests (see, for example, Wainer et al., 2000)—except that the outcome of the assessment is a knowledge state, rather than a numerical estimate of a student's competence in the topic. The procedure follows a scheme outlined in Figure 1.

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 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

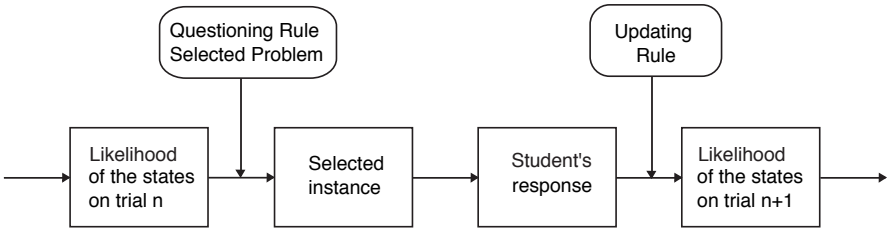

Figure 1. Diagram of the transitions in the assessment procedure. Three operations are involved: the selection of a maximally informative problem, the choice of a particular instance, and the updating of the likelihood distribution.

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[4]. If several problems 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 likelihoods 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 those of the remaining states decreased. The exact formula of the operator modifying the likelihood distribution will not be recalled here; see Definition 10.10 in Doignon and Falmagne (1999). It is proved there that the operator is commutative, in the sense that its cumulative effect in the course of a full assessment does not depend upon the order in which the problems have been proposed to the student. This commutativity property is consistent with the fact that, as shown by Mathieu Koppen (see Doignon and Falmagne, 1999, Remark 10.11), this operator is Bayesian. 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[5] 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

[4] A different interpretation of 'maximally informative' was also investigated, based on the minimization of the expected entropy of the likelihood distribution. This method did not result in an improvement, and was computationally more demanding.

[5] As compared to the case of a false response, which could be attributed to a careless error.

$\mathbf{p_1}$, 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 solved 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. As mentioned earlier, because all the problems have either open-ended responses or multiple choice responses 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 the learning space at some moment in the assessment of a student. For practical reasons, we chose an example from Falmagne et al. (2006) involving a part of arithmetic involving 108 problems, rather that the full beginning algebra domain whose large number of states would render the graphic representation computationally more difficult. In principle, each colored pixel in the oval shape of Figure 2 represents one of the 57,147 states of the learning space for that part of arithmetic. (Because of graphics limitations, some grouping of similar states into a single pixel was necessary.)

Knowledge states are sorted according to the number of problems 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 that part of 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 aesthetic 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 fewer than 100 states containing either more than 100 problems or fewer than 10 problems. The arrangement of the points on any vertical line is largely arbitrary. The color of a pixel represents the likelihood of the corresponding state. A color coded logarithmic scale, pictured on the right of Figure 2, is

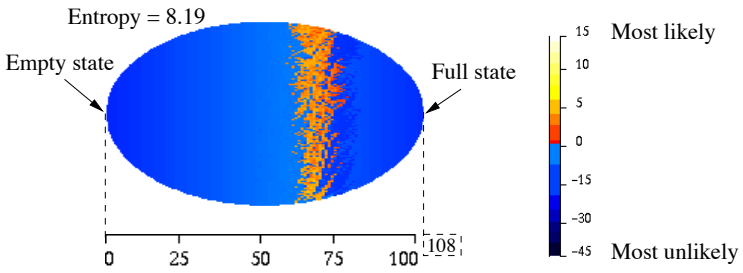

Figure 2. Likelihood map of the learning space representing the exemplary part of arithmetic under discussion.

used to represent the likelihood values. Red, orange, and yellow-white[6] indicate states with a likelihood exceeding the mean of the distribution, with yellow-white marking the most likely states. Conversely, dark blue, blue, and light blue represent states that are less likely than the mean, with dark blue marking the least likely states.

Figure 3 displays a sequence of likelihood maps describing the evolution of an assessment 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 complete assessment took 24 questions, which is close to the average for this part of 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 this curriculum.

Next to each map in Figure 3, 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 (indicated by the red arrow) 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, and so would not be very informative. In this particular case only 24 problems have

[6]Or shades of increasingly light grey in a black and white copy, and similarly darker and darker ones for the lower part of the scale.

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. With the full arithmetic curriculum from the 4th grade up, the assessment takes around 30-35 questions.

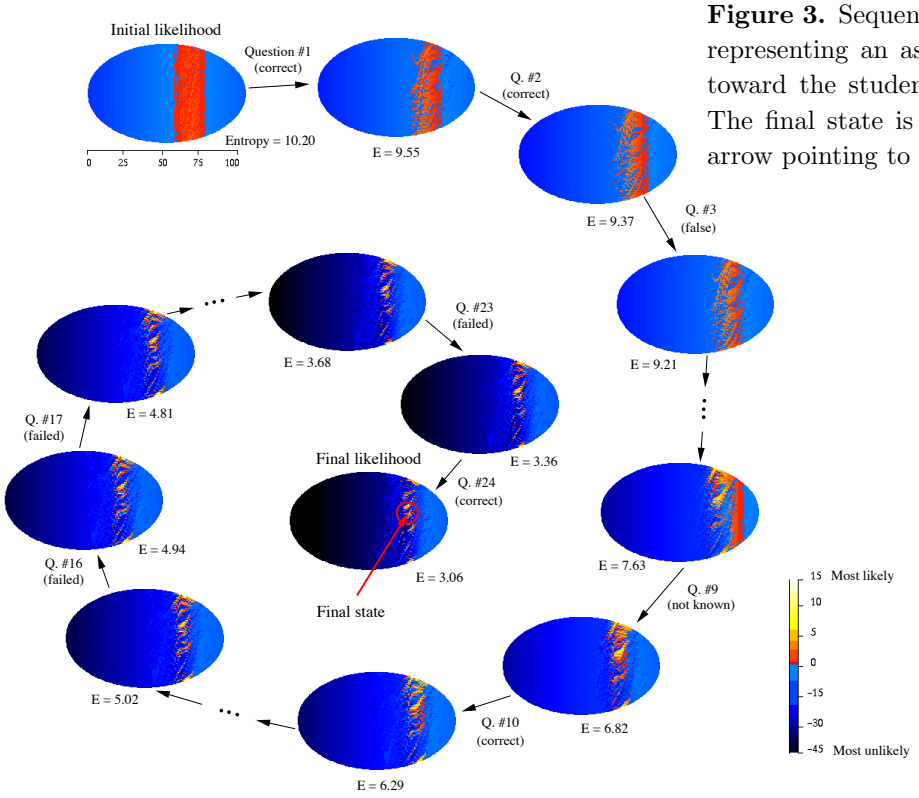

Figure 3. Sequence of likelihood maps representing an assessment converging toward the student's knowledge state. The final state is marked by the long arrow pointing to the circle.

D. Validity and/or Reliability of the Assessment

As indicated in our introductory section, an assessment by the method described above contrasts with a standardized test, whose aim is to obtain a numerical score indicative of a degree of competence in a topic. In the latter case, the issues of the reliability and, especially, the validity of the measurement, are paramount in view of the method used to construct the test. The situation is different in the learning space case because the collection of all the problems potentially used in any assessment represents a fully comprehensive coverage of a particular curriculum, such as beginning algebra, the leading example in this paper. Arguing that such an assessment, if it is reliable, is also automatically endowed with a corresponding amount of validity is plausible. In other words, assuming that the database of problem types is a faithful representation of the curriculum, the measurement of reliability is confounded with that of validity. (We shall go back to this issue, which is fundamental, in our discussion section.) In any event, a different approach is taken here to evaluate the reliability–validity

of the results. In practically all assessments, an *additional* test problem is randomly selected in a uniform distribution on the set of all problems not used in the assessment and given to the student, whose answer is not taken into account in assessing the state. However, at the end of the assessment, the student's response to this problem—correct or false—can be predicted on the basis of the assessed knowledge state. For all topics (arithmetic, beginning algebra, etc.) extensive data are available for each problems in the form of a 2×2 table, with one of the two dichotomies referring to the response predicted from the estimated state, and the other to the observed one (see matrix (1)). How well the assessed state is capable of predicting the response to the additional problem can be evaluated by standard statistical indices. In this paper, we report the results for two such indices: the tetrachoric coefficient and the log odds ratio, together with their correlation.

The subjects were college or high school students (85% and 15%, respectively) who took assessments in beginning algebra via the internet, either in school or at home. In most cases, such assessments were taken in the framework of a computerized course on this subject (see Remark (b) below). These data were provided by 210,102 assessments. For each of the problems in the database, we have a 2×2 table of the following form, in which 'in' stands for 'problem is **in** the assessed state' (and so the prediction is that the response should be correct, barring a careless error), and '**out**' for the alternative:

$$\begin{array}{cc} & \begin{array}{cc} \text{out} & \text{in} \end{array} \\ \begin{array}{c} \text{false} \\ \text{correct} \end{array} & \begin{pmatrix} a & b \\ c & d \end{pmatrix} . \end{array} \tag{1}$$

These data matrices are available at the URL `http://www.aleks.com/paper_psych`, together with a few exemplary instances of some of the problems. Thus, for a particular problem $\boldsymbol{p}$, the letters a and c represent the numbers of cases where $\boldsymbol{p}$ was '**out**' of the assessed state, and the student's responses were false, and correct, respectively. The interpretation of b and d are similar in the '**in**' case.

For each of these data matrices, we compute two statistics: the tetrachoric coefficient and the log odds ratio. The standard rationale for the tetrachoric coefficient is the assumption that the data originated from sampling an underlying 2-dimensional Gaussian distribution, with the results gathered in a 2×2 table; thus, splitting each of the dimensions into two half intervals. This assumption is debatable in our particular case because the all-or-none character of the prediction of success or failure based on the assessed state is a priori inconsistent with the Gaussian assumption. Our choice of this coefficient was dictated by the wish to compare our results with similar statistics in standardized tests, such as the item-test correlation. In any event, the log odds ratios were also

computed. We will see that the results provided by the two statistics are closely related (see Figure 7). The approximation used for the computation of the tetrachoric coefficient is the AS 116 Algorithm of Brown (1977). We used the routine of J. S. Uebersax — `http://ourworld.compuserve.com/homepages/jsuebersax/tetra.htm`— translated into VBA by Keizo Hori.)

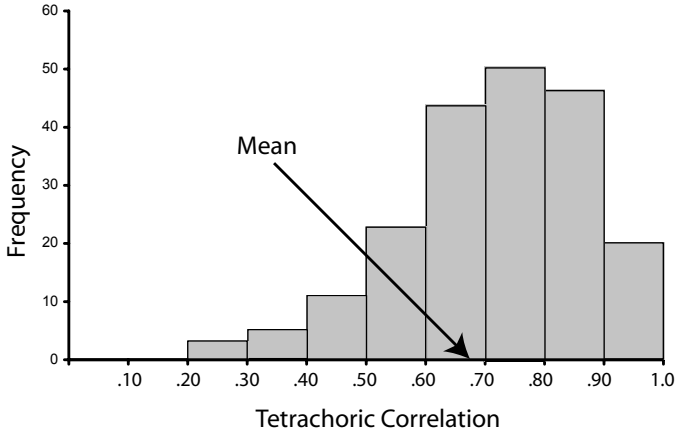

Figure 4. Distribution of the values of the tetrachoric coefficient for 204 problems in beginning algebra, computed from 210,102 assessments. The average number of data points per problem is $210,102/204 \approx 1,030$.

Figure 4 displays the distribution of 204 tetrachoric coefficient values, each corresponding to one of 204 problems from a database of 250 problems in beginning algebra. Forty-six problems were removed because the relevant data were too meager or the approximation formula for the computation of the tetrachoric coefficient was not defined (one of the margins contained a zero). The mean and the median of the distribution are .67 and .68, respectively. In the distribution pictured in Figure 4, the values of the tetrachoric coefficient are indicated by the abscissa, and the number of problems by the ordinate. Note that the problems yielding a relatively low value (below .5) for the tetrachoric coefficient are not regarded as candidates for elimination from the assessment. Rather, the low value is regarded as an indication that the problem may be misplaced in the structure of the learning space, or possibly formulated ambiguously. For reference, we also display in Figure 5 a diagram indicating the sample size of the data for each of the 204 tetrachoric coefficient values, that is, the sum $a + b + c + d$ in Matrix (1). The average sample size is 210,102/204 $\approx$ 1,030.

A similar analysis was performed in terms of the log odds ratio (Breslow, 1981; Agresti, 1995), based on 185 out the 204 matrices. (Nineteen problems were discarded because the 2×2 matrix contained a zero cell.) The distribution of the values of this index is displayed

in Figure 6. The mean and the median of the distribution were approximately 2.75 and 2.58. The diagram of the correlation between the tetrachoric coefficient and the log odds ratio, based on the 185 data points for which both indices were available, is given in Figure 7.

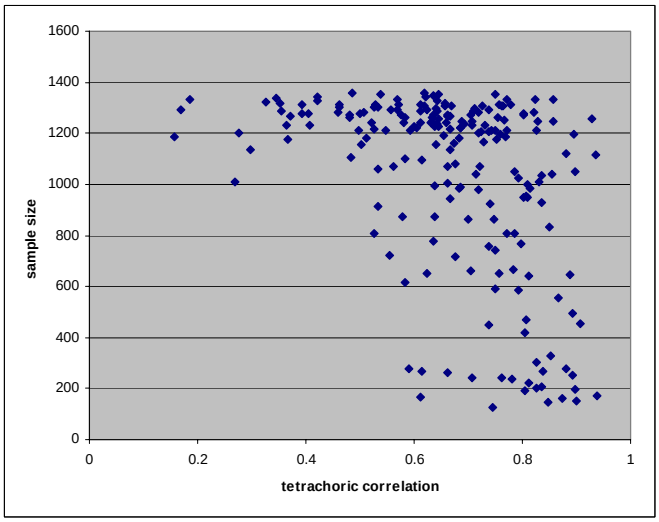

Figure 5. Correlation diagram of the values of the tetrachoric coefficient vs the sample size: $a+b+c+d$ in Matrix (1). Each point represents one of the 204 matrices.

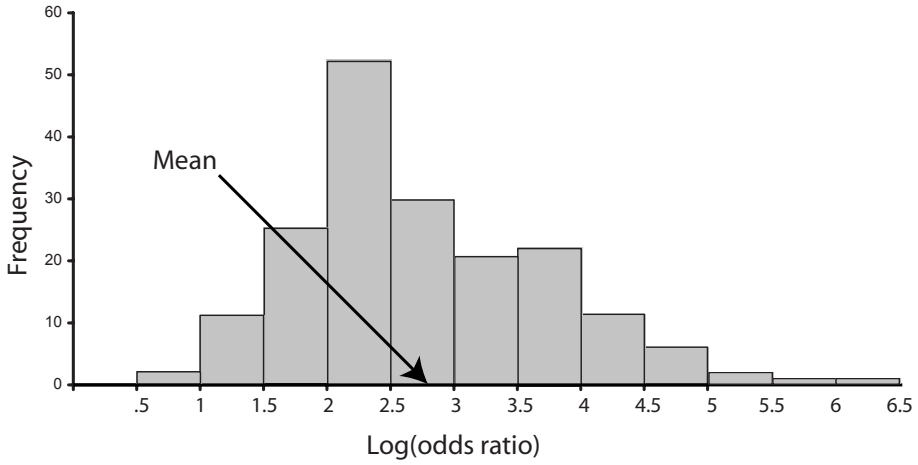

Figure 6. Distribution of the values of the log odds ratio statstic for 185 problems in beginning algebra (out of 204; 19 problems with a zero cell were discarded).

Remarks. (a) We must point out that, from the standpoint of predicting the mastery of beginning algebra on the basis of the assessment, the tetrachoric coefficient is a somewhat underestimating index[7]. Indeed, while there are practically no lucky guesses, there are careless errors, which are bound to lower the values of all the tetrachoric coefficients, albeit unequally[8].

[7]For that matter, so are any other correlation indices.

[8]Problems may differ widely in their propensity to elicit a careless error.

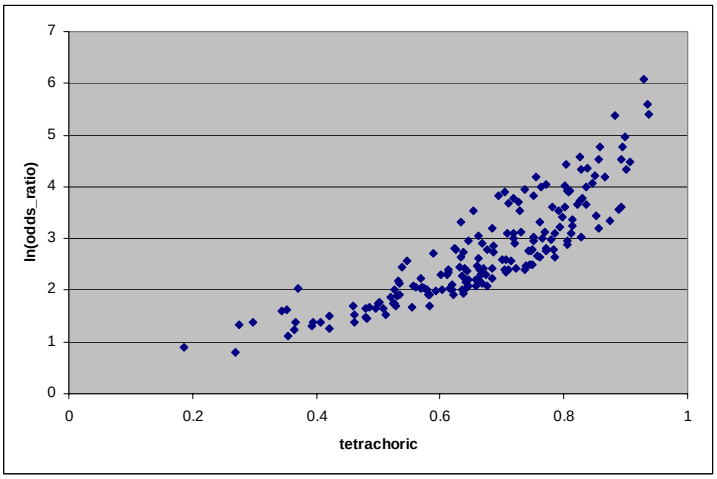

Figure 7. Diagram of the correlation between the tetrachoric coefficient and the log odds ratio, based on the 185 problems for which both indices were available.

The presence of careless errors may be appraised by comparing the average estimates of two conditional probabilities. Denoting by a^*, b^*, c^* and d^* the sum of each of the numbers a, b, c and d in Matrix (1) over all 204 problems, we obtain the average estimates

$$\mathbb{P}(\text{correct response} \,|\, \text{problem in student's state}) \approx \frac{d^*}{b^* + d^*} = .70 \tag{2}$$

$$\mathbb{P}(\text{incorrect response} \,|\, \text{problem not in student's state}) \approx \frac{a^*}{a^* + c^*} = .87. \tag{3}$$

(If there is no noise of any sort and the model is perfect, without careless errors, both of these probabilities are equal to 1.) We could certainly refine the theory to explain the difference between these estimates, which is most probably due to careless errors, but we shall not do so here.

(b) It may perhaps be argued that our experimental conditions are far from ideal because we had no control over the situations in which these assessments were taken. Because the assessments were part of a course, they were in some instances supervised. In other cases, it is possible that a student taking an assessment at home may have received substantial help from someone. This objection is not as damaging as it may seem. Suppose indeed that a student has not worked alone. The assessed knowledge state would obviously not be that of the student. This state would be either the state of whoever has been helping the student, or a combination—a union, actually—of the states of the student and the helper. This would happen if the help was consistent, namely, the assessment was taken jointly. Remember, however, that the collection of states is closed under union. (This is a consequence of the fact that [K1] and [K2] imply [K2*].) Thus, the union of the student's state and the helper's state is a genuine knowledge state, and we can regard the corresponding data as legitimate.

Another reason for accepting our results, that some readers may find more convincing, resides of course in the very large number of assessments on which our data are based.

(c) Our tetrachoric coefficients are notably high by comparison with the typical values obtained for the item-test correlation in standardized testing. However, the lower correlations obtained in the latter case may conceivably be due to the multiple-choice procedure used in most cases of standardized testing, which unavoidably increases the noise in the data. To evaluate the potential effect of such a procedure on our tetrachoric coefficients, we recalculated all of them under the assumption that the student's responses resulted from a multiple choice procedure with five possible responses. Specifically, we replaced each of our 2×2 matrices (1) by the matrix below

$$\begin{array}{c} \\ \text{false} \\ \text{correct} \end{array} \begin{array}{c} \begin{array}{cc} \text{out} & \text{in} \end{array} \\ \begin{pmatrix} .8a & .8b \\ .2a + c & .2b + d \end{pmatrix} \end{array} . \tag{4}$$

This manipulation is bound to reduce the precision of our predictions. For instance, the .87 value reported in Eq. (3) for the estimate of the average conditional probability of an error, given that a problem is not in the student's state, is now down to .74. Not surprisingly, we observe a concomitant shift to the left in the distribution of Figure 4. This shift is a minor one, however.

(d) Note that the tetrachoric coefficient and the log odds ratio could be corrected to account for careless errors. Indeed, because the additional problem $\boldsymbol{p}$ is selected randomly in a uniform distribution on the set of all problems, some students may be given two instances of the same problem. The data from such cases, which were not used for the computation of the two correlation indices, has in fact been used in our study to estimate the probability of making a careless error in solving problem $\boldsymbol{p}$. In turn, these estimates, obtained for all the problems, can then be applied to manufacture a tetrachoric coefficient and a log odds ratio *corrected for careless errors*. The importance of such considerations is that these corrected indices could be even better tools to gauge the accuracy of the learning space as a cognitive model of the mental organization of the scholarly material. Exploring this issue in depth would go beyond the scope of this paper.

The outer fringe and the learning efficiency. We also mention an indirect, but nevertheless meaningful way of gauging the validity of an assessment. As the outer fringe of a state contains exactly those items that the student is supposedly ready to learn, the learning success for those problem types should be revealing of the accuracy of the assessment. This may be evaluated by the conditional probability that a student is capable of mastering a

problem type, given that it is located in the outer fringe of the student's knowledge state (and so accessible for learning). These probabilities can be estimated from our learning data. For beginning algebra, the median of the distribution of the estimated (conditional) success probabilities is .92. To understand the import of this number, some details about the learning process must be given.

When a problem type is located in the outer fringe of a student's state, the student may select that problem type as the next one to learn. This choice initiates a random walk keeping track of the learning stages for that problem type. The random walk takes place on an interval of the integers and has two absorbing barriers. At each step of that random walk, an instance of the problem type is proposed and the student is asked to solve it. In case of failure, an explanation of the solution is offered which is centered on this instance. The problem type enters the random walk at the point 0 and moves left or right depending on the student's response to successive instances of the problem. The general principle is that a success in the solution of an instance provokes a move to the right, and an error a move to the left. The item is considered to be learned when the random walk hits the right barrier. Hitting the left barrier means that the student is not capable of learning the problem type at that time. The learning of this problem is then postponed and the student's knowledge state may be readjusted.

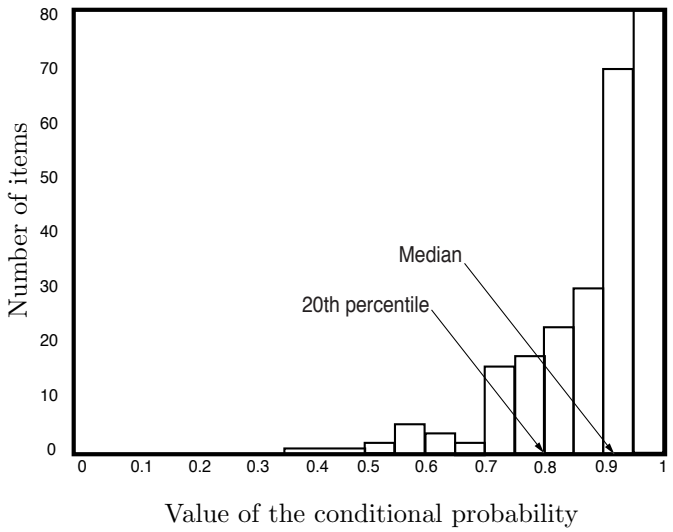

Figure 8. For the 256 items in beginning algebra, a graph of the estimated values of the conditional probabilities that a student entering the random walk reaches its right bound. The problem is then regarded as having been mastered. The median probability is .92, with the 20th percentile at .8. These data are based on 1,564,296 random walks.

Figure 8 displays the distribution of the conditional probabilities that a problem entering the random walk ends up at the right bound, and so is regarded as mastered, at least for the present. (A later assessment would verify the fact.) An examination of the graph shows that many items are satisfactorily handled: 80% of them have a probability of success of at least .8, with the median of the distribution at .92. Nevertheless, the left tail of the distribution indicates that some problems are not learned easily, and that adjustments have to be made: some intermediate problems may be missing and should then be added; also, the problem may be misplaced in the structure; or the explanation may be defective and should be rewritten. Great care is taken in the design of the various instances of a problem so that a correct response should never be attributed to a trivial device, unrelated to the understanding of the problem.

The construction of a learning space. An important aspect of our work has not been discussed so far in this paper and must be commented upon. Building a learning space in a practical situation is still today a very demanding task, taking months, and relying in part on the judgment of experts responding to probing questions about the curriculum. These questions can be generated systematically by the QUERY routine developed by Koppen (1993), Dowling (1993a), Dowling (1993b) and further elaborated by Cosyn and Thiéry (2000) (see also Villano, 1991; Müller, 1989; Kambouri et al., 1994; Dowling, 1994; Doignon and Falmagne, 1999). This routine stores each response of the expert using QUERY, and makes sophisticated inferences in choosing each successive question, so as to maximize the information and shorten the questioning. Even so, each expert spends many hours responding to the queries. Moreover, the resulting structure qualifies only as a preliminary step, potentially plagued by inconsistencies in the experts' responses (see Kambouri, 1991), and in need of refinements and corrections. The refinement of this initial learning space relies then on a statistical analysis of students' data, involving the tetrachoric coefficient discussed in this paper, or related indices. Note that, by contrast with the practice common in psychometrics, where an item is often rejected when its item-test correlation is regarded as unacceptably low[9], such a rejection of a problem is exceedingly rare in an empirical test of a learning space. Rather, either one finds in such a case that the statement of the problem is defective or ambiguous, and a revision takes place, or the local structure of the learning space affected by the problem is reexamined and suitably altered. Such a treatment is required for those problems having values of their tetrachoric coefficient below .4 or even .5 in the distribution of Figure 4. One might be rightfully concerned by the painstaking manipulations involved in the refinement of a learning space. We go back to this point in the discussion section.

[9] A minimum value exceeding .15 is sometimes required (e.g. Kehoe, 1995).

E. Summary and Discussion

The aim of this research was to evaluate the extent to which an assessment in a learning space is predictive of the mastery of the topic. In other words, we wanted to appraise the validity of such an assessment. We took beginning algebra as an exemplary scholarly topic. The method used for such an appraisal was systematically to ask the student an additional problem, randomly selected, and predict the student's response to that problem on the basis of the student's knowledge state diagnosed by the assessment engine. This prediction is possible because, by definition, a knowledge state is a set containing all the problems mastered by the student. Thus, for each problem, the available data takes the form of a 2×2 table of the kind displayed in (1) with the numbers of observed responses (false or correct) in the rows, and the number of predicted ones ('out of' or 'in' the state) in the columns. A tetrachoric coefficient of correlation was computed for 204 out of 250 problems in the database for beginning algebra (46 problems were discarded because there were not enough data or the tetrachoric coefficient was not defined for these problems). These data were obtained from 210,102 assessments. This means that, on the average, $210{,}102/204 \approx 1{,}030$ assessments per problem were used to compute the tetrachoric coefficients. The distribution of the values of this coefficient was displayed in Figure 4. The median of that distribution is around .68. A similar analysis was performed with the log odds ratio (see Figure 6). The correlation diagram presenting the joint results for the two indices was given in Figure 7. We also presented an analysis of the learning efficiency achieved by students because it is directly related to the validity of the assessment. Indeed, the problem types that are proposed to the student for learning are those located in the outer fringe of that student's state, as revealed by the assessment. The argument is that a valid assessment leads to a correct gauging of the outer fringe, which should entail efficient learning. The distribution of the conditional probabilities of learning successes are displayed in Figure 8. The median of that distribution is .92, which may be regarded as an indirect evidence of the validity of the assessment.

As far as the tetrachoric coefficients and the log odds ratio statistics are concerned, our data analysis is similar to an item-test correlation study in a psychometric test, inviting a comparison. The data presented here suggest that an assessment in a learning space is capable of more reliable/valid predictions. An offhand conclusion in that direction would be hasty, however, and perhaps misleading in view of the fundamental differences between the two types of instruments. Let us review them here.

1. The objectives and the philosophy. A learning space is precisely tailored to assess the knowledge states of students in a well defined area. Not only is its database of problems curriculum driven, but it is intended to be comprehensive for that curriculum. It seems possible to extend our type of assessment to very large learning spaces, such as all of K-12 mathematics. However, while such a far reaching learning space has been constructed, no solid results concerning such a large scale application are yet available. By contrast, the ambition of a psychometric instrument is to arrive at a numerical evaluation of one or a few aptitudes or competencies. Its philosophy owes much to nineteenth century physics, with Galton, Pearson and Kelvin, whose credo held that precision in science was tantamount to numerical measurement.

2. The theories. The learning spaces are defined by two axioms, [K1] and [K2], solely motivated by pedagogical considerations. One axiom formalizes the possibility of learning the material gradually, one problem at the time. The other expresses the principle that 'knowing more does not make one less able to learn something new.' The mathematical tools come from combinatorics and stochastic processes. The theoretical framework of standardized tests is psychometric theory with all its variants. This theory is formulated in the framework of calculus and statistics.

3. The results of the assessment or the test. The outcome of an assessment is a knowledge state, which is exactly represented, from a theoretical viewpoint (cf. the Fringe Theorem), by the two relatively short lists of problems in its outer and inner fringes, such as those contained in Tables 1a and 1b. Together, these two lists pinpoint one among possibly 10^7 feasible knowledge states in the learning space. The result of a psychometric test is a number or a numerical vector with a small number of components. By design, the number of possible results of a test is several orders of magnitude smaller than the typical number of feasible knowledge states in a learning space.

4. The principles underlying the construction of the test. They are dictated by the respective objectives. In the case of a learning space, there should be a consensus among educators that the database of problems is a comprehensive compendium for testing the mastery of a scholarly subject. This phase is relatively straightforward. On the other hand, the construction of the learning space, that is, the delineation of the collection of feasible knowledge states, is extremely painstaking, and can be regarded as satisfactory only after a prolonged period of successive revisions of the learning space, based on data of the kind reported here in Figures 4 and 6, and yielding an acceptable degree of predictive power. What was not clear at first is that the size of

the collection of feasible knowledge states would be manageable, considering that it is a subfamily of a family containing 2^n sets (where n is the number of problems; we have 2^{250} possible subsets in the case of beginning algebra discussed here). Actually, in all the cases investigated so far, which include only elementary mathematics and other quantitative or highly structured topics, the size of the collection of states has been on the order of 10^7, rendering learning space theory applicable. At this point, it is not yet clear that very different subjects, such as language or history, will be as amenable to such a treatment. The construction of a psychometric test follows very different, much stricter rules concerning the selection of the items, in that one is not free to choose items simply because they are desirable to have. The database of items forms a highly particular ensemble. It is obtained by successive modifications—consisting in adding or removing items, for example—leading to a model involving a representation of the students and sometimes the items in a Euclidean space. Validity and reliability are primary concerns, but they are not the only concerns because the model is quite constraining: by definition, it has to be a measurement instrument.

The last is of course the key difference between the two approaches. From a theoretical viewpoint, a learning space is much less demanding than a psychometric model, and its validity is entirely grounded on its reliability, that is, its predictive power regarding problems not tested, or so we will argue here. This is so if there is a broad consensus among educators that its database of problems represents a comprehensive coverage of the curriculum, and that a student capable of solving randomly chosen instances of all the problem types has completely fulfilled the educational goals. Some may feel uneasy about such a point of view, and demand, for example, that the results of an assessment in a learning space be correlated with those of standardized achievement tests such as the Stanford 9. While such a demand would seem to be a reasonable requirement, it is predicated on the belief that such tests as the Stanford 9 are themselves fully valid, which is debatable because of the chance factors entering in the choice of the questions from one year to the next[10]: no single Stanford 9 test can pretend to cover the complete curriculum. To be sure, providing results showing that an assessment in the relevant learning space yields knowledge states that are highly predictive of the test results for each *individual* item of a Stanford 9 would carry much weight. This particular study has not been performed yet. Such a project occupies a top position in our agenda.

[10]Note to mention the noisy character the standardized test data in those frequent cases where a multiple choice format is used.

For a parting comment, we return to what is certainly the most critical phase in the application of learning space theory, which is the actual construction of the collection of knowledge states. We have noted such a construction was delicate and labor intensive. However, the need for such painstaking manipulations may very well be temporary because an automatization of this process is conceivable. Indeed, it turns out that the collection of **all** learning spaces of the type used here for beginning algebra—which forms an important class of learning spaces—can be represented as a connected graph, each vertex of which stands for a particular learning space. A random walk can be defined on the set of all vertices of such a graph, with transitions dictated probabilistically by statistical indices based on student data. This random walk would then evolve toward vertices representing learning spaces increasingly well adapted to the population of students. Thus, the learning space would be self adapting. We are not alluding here at some developments envisaged for some very distant future. The first steps in that direction have already been taken by Thiéry in his doctoral dissertation (Thiéry, 2001). A comprehensive application of these concepts is currently under way (see Eppstein et al., 2007, for some early theoretical results along these lines).

References

A. Agresti. *An Introduction to Categorical Data Analysis*. John Wiley & Sons, New York, 1995.

D. Albert and J. Lukas, editors. *Knowledge Spaces: Theories, Empirical Research, Applications*. Lawrence Erlbaum Associates, Mahwah, NJ, 1999.

A. Björner, M. Las Vergnas, B. Sturmfels, N. White, and G.M. Ziegler. *Oriented Matroids*. Cambridge University Press, Cambridge, London, and New Haven, second edition, 1999.

N. Breslow. Odds ratio estimator when the data are sparsed. *Biometrika*, 68:313–324, 1981.

M.B. Brown. Algorithm as 116: The tetrachoric correlation and its asymptotic standard error. *Applied Statistics*, 26:343–351, 1977.

S. Chaiklin. The zone of proximal development in vygotsky's analysis of learning and instruction. In A. Kozulin, B. Gindis, V. Ageyev, and S. Miller, editors, *Vygotsky's Educational Theory and Practice in Cultural Context*. Cambridge University Press, Cambridge, MA, 2003.

E. Cosyn and N. Thiéry. A Practical Procedure to Build a Knowledge Structure. *Journal of Mathematical Psychology*, 44:383–407, 2000.

E. Cosyn and H.B. Uzun. Axioms for learning spaces. Accepted for publication in the *Journal of Mathematical Psychology*, 2006.

J.-P. Doignon and J.-Cl. Falmagne. *Knowledge Spaces*. Springer-Verlag, Berlin, Heidelberg, and New York, 1999.

J.-P. Doignon and J.-Cl. Falmagne. Spaces for the assessment of knowledge. *International Journal of Man-Machine Studies*, 23:175–196, 1985.

C.E. Dowling. Applying the basis of a knowledge space for controlling the questioning of an expert. *Journal of Mathematical Psychology*, 37:21–48, 1993a.

C.E. Dowling. On the irredundant construction of knowledge spaces. *Journal of Mathematical Psychology*, 37:49–62, 1993b.

C.E. Dowling. Integrating different knowledge spaces. In G.H. Fischer and D. Laming, editors, *Contributions to Mathematical Psychology, Psychometrics, and Methodology*, pages 149–158. Springer-Verlag, Berlin, Heidelberg, and New York, 1994.

P.H. Edelman and R. Jamison. The theory of convex geometries. *Geometrica Dedicata*, 19: 247–271, 1985.

D. Eppstein, J.-Cl. Falmagne, and H.B. Uzun. On verifying and engineering the wellgradedness of a $\cup$-closed family. 2007. To be submitted.

J.-Cl. Falmagne and J-P. Doignon. A class of stochastic procedures for the assessment of knowledge. *British Journal of Mathematical and Statistical Psychology*, 41:1–23, 1988.

J.-Cl. Falmagne, E. Cosyn, J.-P. Doignon, and N. Thiéry. The assessment of knowledge, in theory and in practice. In B. Ganter and L. Kwuida, editors, *Formal Concept Analysis, 4th International Conference, ICFCA 2006, Dresden, Germany, February 13–17, 2006*, Lecture Notes in Artificial Intelligence, pages 61–79. Springer-Verlag, Berlin, Heidelberg, and New York, 2006.

M. Kambouri. *Knowledge assessment: A comparison between human experts and computerized procedure*. PhD thesis, New York University, New York, 1991.

M. Kambouri, M. Koppen, M. Villano, and J.-Cl. Falmagne. Knowledge assessment: Tapping human expertise by the QUERY routine. *International Journal of Human-Computer Studies*, 40:119–151, 1994.

J. Kehoe. Basic item analysis for multiple-choice tests. *Practical Assessment, Research & Evaluation*, 4(10):19–36, 1995.

M. Koppen. Extracting human expertise for constructing knowledge spaces: An algorithm. *Journal of Mathematical Psychology*, 37:1–20, 1993.

C.E. Müller. A procedure for facilitating an expert's judgments on a set of rules. In E.E. Roskam, editor, *Mathematical Psychology in Progress*, Recent Research in Psychology, pages 157–170. Springer-Verlag, Berlin, Heidelberg, and New York, 1989.

J. Nunnally and I. Bernstein. *Psychometric Theory*. MacGraw-Hill, New York, 1994.

M. Villano. *Computerized knowledge assessment: Building the knowledge structure and calibrating the assessment routine*. PhD thesis, New York University, New York, 1991. In *Dissertation Abstracts International*, vol. 552, p. 12B.

L.S. Vygotsky. *Mind and society: The development of higher mental processes*. Harvard University Press, Cambridge, MA, 1978.

D.J.A. Welsh. Matroids: Fundamental concepts. In R.L. Graham, M. Grötschel, and L. Lovász, editors, *Handbook of Combinatorics*, volume 1. The M.I.T. Press, Cambridge, MA, 1995.