2

Assessing Mathematical Knowledge in a Learning Space[1]

Eric Cosyn[2]
Christopher Doble[2]
Jean-Claude Falmagne[2,3]
Arnaud Lenoble[2]
Nicolas Thiéry[2]
Hasan Uzun[2]

2.1 Introduction

According to Knowledge Space Theory (KST) (cf. Doignon and Falmagne, 1999; Falmagne and Doignon, 2011), a student's competence in a mathematics or science subject, such as elementary school mathematics or first year college chemistry, can be described by the student's 'knowledge state,' which is the set of 'problem types' that the student is capable of solving. (In what follows, we abbreviate 'problem type' as 'problem' or 'item.') As the student masters new problems, she moves to larger and larger states. Some states are closer to the student's state than others, though, based on the material she must learn in order to master the problems in those states. Thus, there is a structure to the collection of states, and this structure gives rise to a 'learning space,' which is a special kind of knowledge space. These concepts have been discussed at length in Chapter 1 of this volume. We recall here that the collection of states forming a *learning space* always contains the 'empty state' (the student knows nothing at all in the scholarly subject considered) and the 'full state' (the student knows everything in the subject). The collection of states must also satisfy two pedagogically cogent principles, which we state below in non-mathematical language.

Consider two hypothetical students S and S', with S' knowing everything that S knows, and more. Then the following hold.

[L1*] Student S can catch up with Student S' by learning the missing concepts one at a time.

[1]We are grateful to Brian Junker, Don Laming and two anonymous referees for their useful comments on an early presentation of this material.

[2]ALEKS Corporation.
[3]University of California, Irvine.

J.-C. Falmagne et al. (eds.), *Knowledge Spaces: Applications in Education*,
DOI 10.1007/978-3-642-35329-1_2,

[L2*] Any new concept that Student S is ready to learn either was already mastered by Student S', or S' is also ready to learn it.

Set-theoretical formulations of [L1*] and [L2*] are given in Chapter 8, the introductory chapter to Part II entitled '*Learning Spaces: A Mathematical Compendium*' on page 131. While these two principles may seem rather trite at first blush, we have seen in Chapter 1 that they have strong, non-obvious implications. In particular, the *Fringe Theorem* results from these principles, which states that every state in a learning space is defined by its two fringes (cf. page 11). Thus, we can summarize the result of an assessment by the two fringes of the uncovered knowledge state without any loss of information. The significance of this result from an educational standpoint is that we can interpret the outer fringe as the set of problems that the student is ready to learn. In other words, the assessment gives a precise access to further learning.

Several assessment systems are founded on KST, the most prominent ones being `ALEKS` and `RATH` (see Hockemeyer, 1997a). The focus of this chapter is an examination of the `ALEKS` system, whose assessments are adaptive and taken on-line. We evaluate the validity of these assessments on the basis of student data. Our usage of the term 'validity' in this context requires some comment.

2.2 Measuring the Validity/Reliability of an Assessment

An assessment in a learning space contrasts with a psychometric test, whose aim is to obtain a numerical score[4] indicative of the competence of a student in a scholarly subject. In the latter case, the validity of such a measurement is paramount because, *a priori*, there is no immediate, obvious connection between a numerical score of competence in a subject, such as elementary algebra, and the ability to solve a particular problem, such as a quadratic equation by the method of the discriminants, or a word problem on proportions. Such a connection is especially questionable in view of the standard methods used in the construction of such a test, which are based on a criterion of homogeneity of the problems in the test: a problem whose response is poorly correlated with the overall result of the test may be eliminated, even though it may be an integral part of the relevant curriculum. In principle, the situation is quite different in the case of the learning space because the collection of all the items[5] potentially used in any assessment is, by design, a fully comprehensive coverage of a particular curriculum. Arguing that such an assessment, if it is reliable, is also automatically endowed with a corresponding amount of validity is plausible. In other words, granting that the database of problem types is a faithful representation of the curriculum, the measurement of reliability is confounded with that of validity.

[4] Or, in some cases, a numerical vector with a small number of terms.

[5] We recall that we use "item" or "problem" to mean "problem type." The actual question asked is an 'instance' of a problem type.

In any event, we use the following method to evaluate the reliability and validity of the results. In each assessment, an *extra* problem (item) $\boldsymbol{p}$ is randomly selected in a uniform distribution on the set of all problems. Then an instance of $\boldsymbol{p}$, also randomly selected, is given to the student, whose response is not taken into account in assessing the student's state. On the basis of the knowledge state uncovered by the assessment, a prediction can be made regarding the student's response to the extra problem $\boldsymbol{p}$, and the accuracy of the prediction can be evaluated. We shall also investigate the evolution of the accuracy of this prediction in the course of the assessment.

In the rest of this chapter, we describe a large scale study performed to test the validity and reliability of an assessment in elementary school mathematics covering grades 3 to 6. The analysis is based on the particular learning space for this subject used by the `ALEKS` system.

2.3 Statistical Methods

2.3.1 Outline of three methods. Three different types of data analysis were performed, based on a large number of assessments taken from June 2009 to December 2012.

1) The most obvious method of predicting the student's actual response (correct or false) to the extra problem $\boldsymbol{p}$ is to check whether or not $\boldsymbol{p}$ appears in the student's knowledge state selected by the assessment algorithm at the end of the test. The effectiveness of such predictions can be evaluated using standard measures of correlation between two dichotomous variables, such as the tetrachoric coefficient or the phi-correlation coefficient. This type of analysis is one of the cornerstones of this chapter and is contained in Paragraph 2.4.4. This analysis does not take possible careless errors into account. A variant of the above method, described in Paragraph 2.4.5, uses the same type of data, but corrects the predictions by a factor depending of the probability that the student commits a careless error in responding to a particular problem. The correlation coefficient used is the point biserial. However, limiting our analysis to the data of such 2×2 correlation matrices does not take full advantage of all the information available.

2) We have seen in Section 1.3 that the core mechanism of the assessment algorithm resides in updating, from one trial to the next and on the basis of the student's response to the question, the likelihood of the knowledge states. A correct response to some item $\boldsymbol{q}$ presented on a trial results in an increase of the probabilities of all the states containing $\boldsymbol{q}$, and an incorrect response in a decrease of all such probabilities. This means that, from the standpoint of the assessment engine, the probability of a correct response to the extra problem $\boldsymbol{p}$ on trial n of the assessment can be obtained by summing the probabilities of all the states containing $\boldsymbol{p}$ on that trial. We use the point biserial coefficient to evaluate the correlation between this continuous variable and the dichotomous 0/1 variable coding the student's response (false/correct) to problem $\boldsymbol{p}$. As this computation can in principle be performed on any trial, we can trace the

evolution of such a prediction in the course of the assessment. The results are the first ones reported in the next section (see Paragraph 2.4.2).

3) The third method for evaluating the validity of the assessment is based on a different idea. At the end of most assessments, the student picks an item in the outer fringe of the state assigned by the assessment engine and begins learning[6]. If the knowledge state assigned to the student is the true one or at least strongly resembles the true one, then the prediction of what the student is capable of learning at that time should be sound. Accordingly, we can gauge the validity of the assessments by the probability that the student successfully masters an item chosen in the outer fringe of the assessed state. We have computed such probabilities for a very large number of assessment/learning trials in both elementary school mathematics and first year college chemistry. Paragraph 2.4.6 contains the results.

2.3.2 Concepts and notation. We write $\mathcal{K}$ for the collection of all the knowledge states in the particular subject considered. As recalled above, the likelihoods of the knowledge states are systematically modified in the course of an assessment as a result of a student's response (false or correct). We denote by ξ_n the likelihood distribution on $\mathcal{K}$ on the nth trial of the assessment, and by $\xi_n(K)$ the corresponding likelihood of state K on that trial. Thus, the assessment begins with an initial, or *a priori*, distribution ξ_1 on $\mathcal{K}$. An item $\boldsymbol{q}_1$ is chosen and given to the student; the student's response $\boldsymbol{r}_1$ is recorded, leading to the transformation of ξ_1 into ξ_2 by a Bayesian operator, etc. In symbols, we thus have the sequence

$$(\xi_1, \boldsymbol{q}_1, \boldsymbol{r}_1) \longmapsto \ldots \longmapsto (\xi_n, \boldsymbol{q}_n, \boldsymbol{r}_n) \longmapsto \ldots \longmapsto (\xi_{L-1}, \boldsymbol{q}_{L-1}, \boldsymbol{r}_{L-1}) \longmapsto \xi_L,$$

with L denoting the number of the last trial of the assessment. So, the likelihood distribution ξ_L on $\mathcal{K}$ gives the final result of the assessment. The likelihood distribution ξ_n may be regarded as a *probabilistic knowledge state* on trial n. If there were no careless errors or lucky guesses, ξ_n would provide, for every item $\boldsymbol{q}$ in the domain, the probability $P_n(\boldsymbol{q})$ of a correct response if an instance of $\boldsymbol{q}$ were presented on trial n. Indeed, in our theoretical framework, that probability is the sum of all the probabilities of the states containing $\boldsymbol{q}$. Writing $\mathcal{K}_{\boldsymbol{q}}$ for the subcollection of $\mathcal{K}$ of all the knowledge states containing $\boldsymbol{q}$, we could thus compute $P_n(\boldsymbol{q})$ by the equation

$$P_n(\boldsymbol{q}) = \sum_{K \in \mathcal{K}_{\boldsymbol{q}}} \xi_n(K). \tag{2.1}$$

However, the assumption that there are no careless errors is not warranted and we will shortly enlarge our notation to take care of this aspect of the data. We do, however, suppose that there are no lucky guesses. This assumption is justified because all the problems either have open responses or offer a multiple

[6] Exceptions are cases in which the assessment is a placement test or serves as the final exam of a course.

choice with a very large number of possibilities. We assume that no learning on the part of the student is taking place during the assessment. So, the index n in $\xi_n(K)$ and $P_n(\boldsymbol{q})$ only marks the progress of the assessment.

Note that the last trial number L and the extra problem $\boldsymbol{p}$ depend upon the student. More exactly, they depend on the particular assessment, and so do ξ_n and $P_n(\boldsymbol{p})$ for all trial numbers $n \leq L$. Making this dependence explicit, we refine our notation and define

$$L_{\mathbf{a}} \quad \text{as the number of the last trial in assessment } \mathbf{a}, \tag{2.2}$$

$$\boldsymbol{p}_{\mathbf{a}} \quad \text{as the extra problem in assessment } \mathbf{a}, \tag{2.3}$$

$$\xi_{\mathbf{a},n} \quad \text{as the probabilistic knowledge state on trial } n \leq L_{\mathbf{a}}, \tag{2.4}$$

$$\mathcal{A} \quad \text{as the set of all the assessments.} \tag{2.5}$$

We also define the collection of random variables

$$\mathbf{R}_{\mathbf{a}} = \begin{cases} 0 & \text{if the student's response to problem } \boldsymbol{p}_{\mathbf{a}} \\ & \text{of assessment } \mathbf{a} \text{ is incorrect} \\ 1 & \text{otherwise.} \end{cases}$$

There is no ambiguity in using the abbreviation $\mathbf{R}_{\mathbf{a}} = \mathbf{R}_{\boldsymbol{p}_{\mathbf{a}}}$ since any assessment $\mathbf{a}$ defines a unique extra problem $\boldsymbol{p}_{\mathbf{a}}$.

The possibility of careless errors must be taken into account. We suppose that a careless error probability is attached to each problem and define

$$\epsilon_{\boldsymbol{q}} \quad \text{as the probability of committing a careless error on problem } \boldsymbol{q}.$$

In the framework of the theory, the interpretation of this parameter is straightforward: $\epsilon_{\boldsymbol{q}}$ is the conditional probability that a student whose knowledge state contains $\boldsymbol{q}$ commits a careless error in attempting to solve that problem. We suppose that the parameter $\epsilon_{\boldsymbol{q}}$ only depends on the problem $\boldsymbol{q}$ and does not vary in the course of an assessment. In accordance with the abbreviation convention used above for $\mathbf{R}_{\mathbf{a}}$, from now on we write

$$\epsilon_{\mathbf{a}} = \epsilon_{\boldsymbol{p}_{\mathbf{a}}}, \qquad \text{and} \qquad \mathcal{K}_{\mathbf{a}} = \mathcal{K}_{\boldsymbol{p}_{\mathbf{a}}}$$

for the subcollection of states containing problem $\boldsymbol{p}_{\mathbf{a}}$. As mentioned, we assume that the probability of a lucky guess is zero for each item.

The probability $\mathbf{P}_{\mathbf{a},n}$ that a student correctly solves the extra problem $\boldsymbol{p}_{\mathbf{a}}$, based on the information accumulated by the assessment algorithm up to and including trial n of assessment $\mathbf{a}$, is thus defined by the following equation:

$$\mathbf{P}_{\mathbf{a},n} = (1 - \epsilon_{\mathbf{a}}) \sum_{K \in \mathcal{K}_{\mathbf{a}}} \xi_{\mathbf{a},n}(K) + 0 \times \sum_{K \in \mathcal{K} \setminus \mathcal{K}_{\mathbf{a}}} \xi_{\mathbf{a},n}(K) \tag{2.6}$$

$$= (1 - \epsilon_{\mathbf{a}}) \sum_{K \in \mathcal{K}_{\mathbf{a}}} \xi_{\mathbf{a},n}(K). \tag{2.7}$$

The 0 factor in (2.6) is included as a reminder of the hypothesis that the probability of a lucky guess is zero.

2.3.3 Estimating the careless error parameters. Our estimate of $\epsilon_{\boldsymbol{q}}$ is based on the assessments in which $\boldsymbol{q}$ has been presented as the extra problem and also, at least once, as part of the assessment[7]. Coding an error and a correct response as 0 and 1, respectively, our sample space for estimating the careless error probability $\epsilon_{\boldsymbol{q}}$ of item $\boldsymbol{q}$ is thus the set $\{(0,0),(0,1),(1,0),(1,1)\}$, in which, by convention, the first term of every pair codes the response to $\boldsymbol{q}$ as the extra problem, and the second term the response to the other relevant presentation of $\boldsymbol{q}$ in that assessment. Denoting by $p_{\boldsymbol{q}}(i,j)$ the probability of sampling the point (i,j) and by $\gamma_{\boldsymbol{q}}$ the probability that the knowledge state of the student belongs to $\mathcal{K}_{\boldsymbol{q}}$, we have the following (compare with Eq. (2.7)):

$$p_{\boldsymbol{q}}(0,0) = \epsilon_{\boldsymbol{q}}^2\gamma_{\boldsymbol{q}} + (1-\gamma_{\boldsymbol{q}}) \tag{2.8}$$

$$p_{\boldsymbol{q}}(0,1) = \epsilon_{\boldsymbol{q}}(1-\epsilon_{\boldsymbol{q}})\gamma_{\boldsymbol{q}} \tag{2.9}$$

$$p_{\boldsymbol{q}}(1,0) = (1-\epsilon_{\boldsymbol{q}})\epsilon_{\boldsymbol{q}}\gamma_{\boldsymbol{q}} \tag{2.10}$$

$$p_{\boldsymbol{q}}(1,1) = (1-\epsilon_{\boldsymbol{q}})^2\gamma_{\boldsymbol{q}}. \tag{2.11}$$

Writing $N_{\boldsymbol{q}}$ for the number of assessments having at least two presentations of item $\boldsymbol{q}$, with one of them as the extra problem, and $N_{\boldsymbol{q}}(i,j)$ for the number of times (i,j) is realized among the $N_{\boldsymbol{q}}$ assessments, we obtain the statistic

$$\texttt{Chi}_{\boldsymbol{q}}(\epsilon_{\boldsymbol{q}},\gamma_{\boldsymbol{q}}) = \sum_{i,j}\frac{(N_{\boldsymbol{q}}(i,j) - N_{\boldsymbol{q}}\cdot p_{\boldsymbol{q}}(i,j))^2}{N_{\boldsymbol{q}}\cdot p_{\boldsymbol{q}}(i,j)}, \tag{2.12}$$

in which the $p_{\boldsymbol{q}}(i,j)$'s are defined by (2.8)-(2.11). The parameters $\epsilon_{\boldsymbol{q}}$ and $\gamma_{\boldsymbol{q}}$ are estimated by minimizing the Chi-square statistic (2.12). The details are relegated to the Appendix 2.6 on page 49. The estimators for $\epsilon_{\boldsymbol{q}}$ and $\gamma_{\boldsymbol{q}}$ are

$$\hat{\epsilon}_{\boldsymbol{q}} = \frac{N_{\boldsymbol{q}}(0,1)+N_{\boldsymbol{q}}(1,0)}{N_{\boldsymbol{q}}(0,1)+N_{\boldsymbol{q}}(1,0)+2N_{\boldsymbol{q}}(1,1)}, \tag{2.13}$$

$$\hat{\gamma}_{\boldsymbol{q}} = \frac{(N_{\boldsymbol{q}}(0,1)+N_{\boldsymbol{q}}(1,0)+2N_{\boldsymbol{q}}(1,1))^2}{4N_{\boldsymbol{q}}(1,1)\cdot N_{\boldsymbol{q}}}. \tag{2.14}$$

We may regard $\texttt{Chi}_{\boldsymbol{q}}(\hat{\epsilon}_{\boldsymbol{q}},\hat{\gamma}_{\boldsymbol{q}})$ as a χ_1^2 random variable[8] with $3-2=1$ degree of freedom (three degrees of freedom in the 2×2 table of the $p_{\boldsymbol{q}}(i,j)$'s minus two estimated parameters). As such, $\texttt{Chi}_{\boldsymbol{q}}(\hat{\epsilon}_{\boldsymbol{q}},\hat{\gamma}_{\boldsymbol{q}})$ may serve as an additional indicator of the adequacy of a model based on an "all-or-none" assumption regarding the mastery of an item, and a zero probability of a lucky guess.

[7] We only consider the first of these non-extra-problem presentations of $\boldsymbol{q}$.

[8] However, see Remark 2.3.4 (b).

2.3.4 Remarks.

(a) Another possibility for estimating the parameter $\epsilon_{\boldsymbol{q}}$ would rely on those assessments in which the final state chosen by the assessment algorithm contains item $\boldsymbol{q}$, with this item being presented at least once during the assessment (either as the extra problem or otherwise) and answered incorrectly by the student (in the first presentation).

The objection to this method is that the choice of a knowledge state for a student at the end of some assessment $\mathbf{a}$ is based on choosing the most likely state in the final probabilistic state $\xi_{\mathbf{a},L_{\mathbf{a}}}$ (cf. (2.2), (2.4) for this notation). At that time, there may still be several states with a maximally high likelihood. The algorithm then chooses randomly among them. This choice is not very critical from an assessment or even learning standpoint because these states very much resemble each other.[9] Nevertheless, the remaining uncertainty regarding the exact state of the student makes this method for estimating careless errors questionable. Note that the method actually used, i.e., the one described in Paragraph 2.3.3, does not suffer from this shortcoming. In particular, the method used does not rely on the final assessed state.

(b) The phrase "*probability that the knowledge state of the student contains item* $\boldsymbol{q}$" is ambiguous. Its interpretation as the parameter $\gamma_{\boldsymbol{q}}$ in the chi-square statistic of Eq. (2.12) is a device allowing us to estimate $\epsilon_{\boldsymbol{q}}$, which is our primary concern. However, while $\epsilon_{\boldsymbol{q}}$ has a legitimate place in our theory, the parameter $\gamma_{\boldsymbol{q}}$ that we estimate by minimizing $\mathtt{Chi}_{\boldsymbol{q}}(\epsilon_{\boldsymbol{q}}, \gamma_{\boldsymbol{q}})$ lies outside. We could have estimated $\gamma_{\boldsymbol{q}}$ differently, for example by averaging the final $\xi_{\mathbf{a},L_{\mathbf{a}}}$ values appropriately. However, this method suffers from the same objection spelled out in (a) above.

2.3.5 Temporal course of the assessments. The Vincent curves.
For every assessment $\mathbf{a}$ and every trial number n of that assessment, we have a pair $(\mathbf{P}_{\mathbf{a},n}, \mathbf{R}_{\mathbf{a}})$ of numbers[10]. The first one $\mathbf{P}_{\mathbf{a},n}$ is computed from Eq. (2.7) and is the prediction of the algorithm for the probability of a correct response to the extra problem $\boldsymbol{p}_{\mathbf{a}}$ on trial n of assessment $\mathbf{a}$, taking into account possible careless errors. The second number $\mathbf{R}_{\mathbf{a}}$ is a dichotomous (0/1) variable coding the student's (false/correct) response to $\boldsymbol{p}_{\mathbf{a}}$. This number is constant for a given assessment. If we align all the assessments on their last trial, we can compute such a correlation between the $\mathbf{P}_{\mathbf{a},L_{\mathbf{a}}}$ values, which vary continuously between 0 and 1, and the dichotomous values $\mathbf{R}_{\mathbf{a}}$. This correlation, computed

[9] Continuing the assessment and asking more questions might help in discriminating among these highly likely states, but the benefit would be slight and the cost to the student heavy.

[10] To avoid complicating our notation, we do not distinguish between the random variables $\mathbf{P}$ and $\mathbf{R}$ on the one hand, and their realizations on the other hand.

for all pairs $(\mathbf{P}_{\mathbf{a},L_\mathbf{a}}, \mathbf{R}_\mathbf{a})$ for $\mathbf{a} \in \mathcal{A}$, is a measure of the success of the assessment algorithm in uncovering the student's knowledge state. We have computed such correlations as part of a more extensive analysis of the evolution of the correlation in the course of the assessment. This analysis requires aligning all the assessments, which are typically of different lengths, from the first to the last trial. The method of choice to perform such an alignment is the *Vincent curve*, which is a standard tool for the analysis of learning data[11]. In our case, this consists of splitting all the assessments into the same number of parts—we have chosen 10—and gathering data pertaining to the same parts in all the assessments. Specifically, for each of the ten deciles of each assessment, we combine the data of the last trial for that decile for all the assessments.

An example involving three hypothetical assessments **a**, **b**, and **c** is displayed on Table 2.1. The first line in the body of the table concerns the assessment **a** which takes 40 trials. Each of the 10 deciles for that assessment consists of four trials, the last one of which is retained for the construction of the Vincent curve. We also include the initial trial in this analysis. In other words, only the probabilistic knowledge states $\mathbf{P}_{\mathbf{a},1}$, $\mathbf{P}_{\mathbf{a},4}$, $\mathbf{P}_{\mathbf{a},8}$, ..., $\mathbf{P}_{\mathbf{a},36}$, $\mathbf{P}_{\mathbf{a},40}$ are taken into account for the computation of the correlation coefficients. The second and third line concern the assessments **b** and **c**, which take 30 and 20 trials, respectively. The last trial in each of their deciles is aligned with the corresponding trial of assessment **a**. We also include the very first trial in our analysis.

The value of the correlation coefficient for the first trial measures how much the learning space $\mathcal{K}$, equipped with the *a priori* probability distribution on $\mathcal{K}$, already knows about the student population before the beginning of the assessment. When the length $L_\mathbf{d}$ of some assessment is not a multiple of 10, the rule for specifying the decile is a generalization of the above rule, namely, the trial number retained for the computation of the correlation coefficient in decile i is the smallest integer greater than or equal to $i \times L_\mathbf{d}/10$. For example, for $N_\mathbf{d} = 17$ the trial numbers retained (including the initial trial) are 1, 2, 4, 6, 7, 9, 11, 12, 14, 16, and 17.

The columns of the table that are relevant to the analysis in terms of Vincent curves are printed in red, and so is the initial column, corresponding to the probabilities at the outset of the assessment. For each of these 11 columns, we have computed the correlations between the **P** values measuring the probability that the student responds correctly to the extra question, and the actual response **R** coded as 0 or 1 for false or correct, respectively.

[11] From Stella B. Vincent, who is the first researcher on record to have used this type of analysis (Vincent, 1912).

Table 2.1. The initial trial and the ten deciles in the Vincent curve analysis. The first column lists the assessments. The data of the 11 columns in red are the only ones retained for the correlation analysis.

Assessment	Deciles										
	1			2			...	10			
a	$\mathbf{P}_{\mathbf{a},1}$	$\mathbf{P}_{\mathbf{a},2}$	$\mathbf{P}_{\mathbf{a},3}$	$\mathbf{P}_{\mathbf{a},4}$	$\mathbf{P}_{\mathbf{a},5}$	$\mathbf{P}_{\mathbf{a},6}$	$\mathbf{P}_{\mathbf{a},7}$	$\mathbf{P}_{\mathbf{a},8}$	...	$\mathbf{P}_{\mathbf{a},37}$	$\mathbf{P}_{\mathbf{a},38}$	$\mathbf{P}_{\mathbf{a},39}$	$\mathbf{P}_{\mathbf{a},40}$
b	$\mathbf{P}_{\mathbf{b},1}$		$\mathbf{P}_{\mathbf{b},2}$	$\mathbf{P}_{\mathbf{b},3}$	$\mathbf{P}_{\mathbf{b},4}$		$\mathbf{P}_{\mathbf{b},5}$	$\mathbf{P}_{\mathbf{b},6}$	...	$\mathbf{P}_{\mathbf{b},28}$		$\mathbf{P}_{\mathbf{b},29}$	$\mathbf{P}_{\mathbf{b},30}$
c	$\mathbf{P}_{\mathbf{c},1}$			$\mathbf{P}_{\mathbf{c},2}$	$\mathbf{P}_{\mathbf{c},3}$			$\mathbf{P}_{\mathbf{c},4}$	...	$\mathbf{P}_{\mathbf{c},19}$			$\mathbf{P}_{\mathbf{c},20}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

2.3.6 The correlation coefficients. We computed the correlations between the variables **P** and **R** by the point biserial coefficient

$$r_{pbis} = \frac{M_1 - M_0}{s_n}\sqrt{\frac{n_1 n_0}{n^2}}$$

where

n	is the number of pairs $(\mathbf{P}, \mathbf{R})$,
s_n	is the standard deviation of the continuous variable **P**,
n_1, n_0	are the numbers of $\mathbf{R} = 1$, $\mathbf{R} = 0$ cases, respectively,
M_1, M_0	are the conditional means of **P** given $\mathbf{R} = 1$ and $\mathbf{R} = 0$

(see e.g. Tate, 1954). This correlation coefficient provides an estimate of a Pearson correlation coefficient ρ under some hypotheses regarding the joint distribution of the two random variables involved. Applied to our situation, these hypotheses are as follows:

(1) the variable **R** is obtained by dichotomizing some underlying continuous random variable **X**;

(2) the joint distribution of **P** and **X** is Gaussian;

(3) the marginal distributions have equal variances;

(4) the conditional variances of **P** given $\mathbf{R} = 0$ and $\mathbf{R} = 1$ are equal.

These hypotheses are not satisfied in our case. For one thing, **P** is bounded, taking its values in the interval $[0, 1]$. More critically, experimental plots of the distribution of **P** indicate that the underlying random variable is bimodal, which is at odds with Hypothesis (2). Our data also point to a contradiction of Hypothesis (3), showing that the conditional variance of **P** given $\mathbf{R} = 1$ is substantially larger than that given $\mathbf{R} = 0$. Thus, the values obtained for r_{pbis} should not generally be regarded as estimates of a Pearson correlation

coefficient[12]. Nevertheless, we adopted the point biserial coefficient in view of its frequent use in psychometrics to compute the item-test correlation. Even though our variable $\mathbf{P_a}$ is different from the overall result of a standardized test, it quantifies the final result of an assessment, making a comparison worthwhile.

Two other correlation coefficients have also been used for cases in which the data take the form of double dichotomies, namely the tetrachoric coefficient and the phi coefficient. The tetrachoric coefficient tends to give higher values than the phi coefficient, and that is what we shall observe in Paragraph 2.4.4 where we show that these coefficients give substantially different numbers for the same data. Their values should be regarded as complementary.

2.4 Data Analysis

We recall that the data pertains to elementary school mathematics.

2.4.1 The participants. The participants were elementary school students in grades 3 to 6. The students took assessments via the internet, either in school or at home. In most cases, such assessments were carried out in the framework of a computerized course on the subject. The data only concern the initial assessments taken by the students. This ensures that performance on the assessment is independent from any knowledge acquired on the ALEKS computerized courses[13]. The students used an interface that made it clear what form the answer to a given problem should take (a number, an algebraic expression, a graph, etc.). Students were given a short, 15-minute tutorial on using the computer system's answer input tools and were offered (pre-packaged) online help with the tools during the assessment. Otherwise, no help or feedback was given.

2.4.2 The Vincent curve analysis. The data used here are based on the assessment algorithm outlined in Section 8.8 of Chapter 8 of this volume. This algorithm allowed us to conduct assessments on the very large knowledge structure associated to the domain. The algorithm partitions the domain into several parts on which the structure is 'projected' (see Section 1.4)[14].

[12] There are indications in the literature (cf. Kraemer, 2006; Karabinus, 1975) that r_{pbis} is a robust statistic in some situations, such as for testing the $\rho = 0$ hypothesis. This robustness of r_{pbis} does not seem to extend to cases where ρ is far from 0.

[13] As discussed in Section 2.2, this precaution is somehow superfluous since the domain of knowledge is confounded with the curriculum by design.

[14] Cf. the Glossary on page 25. We discussed in Section 1.4 how and why a large learning space may be split into several parts for the purpose of performing assessments. The concept of a 'projection' was developed for this purpose. In particular, for each item **a**, Eq. (2.15) is defined on the projection of the knowledge structure on the sub-domain that contains **a**.

As mentioned earlier, we use the point biserial coefficient to compute the correlation between the probability of a correct response to the extra problem on trial n, specified by the equation

$$\mathbf{P}_{\mathbf{a},n} = (1 - \epsilon_{\mathbf{a}}) \sum_{K \in \mathcal{K}_{\mathbf{a}}} \xi_{\mathbf{a},n}(K), \tag{2.15}$$

and the actual response to that problem coded as 0 or 1. Note that estimates of the careless error rate $\epsilon_{\mathbf{a}}$ were not available for all items. However, since this rate is defined as constant for a given item, it does not affect the point biserial correlation for that item.

Figure 2.1 traces the evolution of the medians of the distributions of the point biserial coefficient in the course of the assessment. These data concern 300 items and are based on 125,786 assessments. (Seventy items were discarded because the relevant data were too sparse.) All the assessments available were used except those in which the extra problem had also been presented as part of the assessment, which occasionally happened since the extra problem was selected randomly among the available problems.

The initial values (at the zero abscissa) of the correlation are obtained from the *a priori* distribution on the set of knowledge states, before the assessment begins. The fact that the median value, about .19, is substantially above zero shows that something is already known by the assessment engine about the population of students: they are more or less ready to take an assessment in elementary school mathematics, or are already engaged in a course on that subject. Then, the curve reaches what appears to be an asymptote at .46. The upper quartile reaches the value .55 and the lower quartile the value .35.

2.4.3 Vincent curves for different categories of problems. We first investigated how sensitive the Vincent curves were to the careless error rate of the items. The analysis was restricted to the items with sufficient data to estimate their careless error rates as defined by Eq. 2.13. As explained earlier in Section 2.3.3, these estimates rest on assessments where an item as been presented twice: once as the extra question and once as part of the assessment. To reduce the selection bias arising from one presentation being part of the adaptive assessment, we only kept those assessments that presented the question within the first five questions. We obtained estimates for 143 items, with a median careless error rate value of 19%. We divided the items in two groups, the first group comprising the 67 items with a careless error rate less than 19% and the second group comprising the remaining 76 items. The Vincent curves for the median value of each group are displayed in Figure 2.2.

Items with a lower careless error rate tend to fare better than items with a higher careless error rate. The median r_{pbis} reaches the value of .51 for the former and .48 for the latter. The difference between the two Vincent curves however is quite small in view of the difference between quartiles in Fig.2.1.

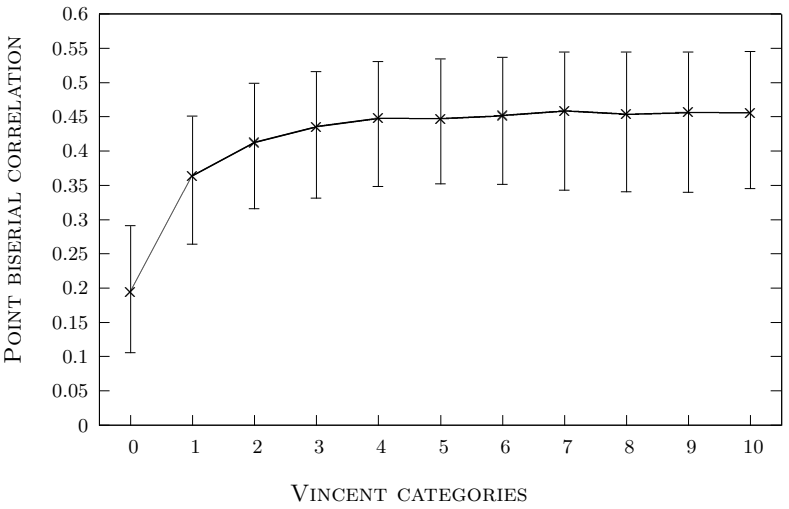

Figure 2.1. Evolution, during the initial part of the assessment, of the point biserial correlation between the probability of a correct response to the extra problem predicted by the assessment engine via Eq. (2.15) and the actual response to that problem. The midpoint of each vertical segment represents the median value of the correlation, the bottom end the 25th percentile value and the top end the 75th percentile value. The abscissae 1, ..., 10 correspond to the ten Vincent categories (cf. Table 2.1). The zero abscissa indicates the initial correlation, before the first trial of the assessment.

We also notice that the median value for both groups of items is greater than the overall median value of .46 from Fig.2.1. The 143 items used for the two Vincent curves in Fig. 2.2 were the most popular items in the early part of the assessment because of the way their careless error rates were computed. Such items tend to be of a middle level of difficulty. The next paragraph provides a closer look at this aspect.

Second, we examined how sensitive the Vincent curves were to the overall difficulty of the items. A simple difficulty index was defined as the proportion of incorrect answers when the item was presented as the extra question. Such a ratio can be used as a rough and straightforward measure of the item difficulty. In particular, it does not attempt to correct for careless errors. We divided the 300 items with sufficient data to compute their point biserial correlations in three groups. The first group has 22 items with a difficulty index lower than 33%, the second group has 84 items with a difficulty index between 33% and 66%, and the third group has 194 items with a difficulty index greater than

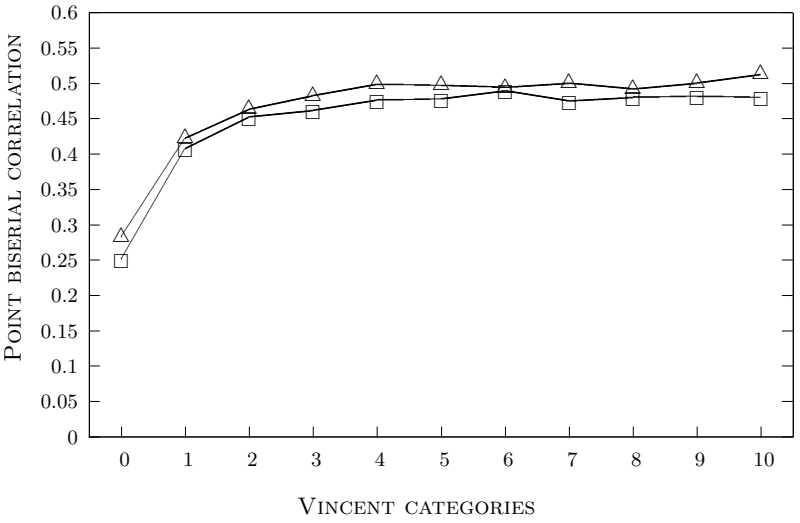

Figure 2.2. Vincent curves similar to the one of Figure 2.1. The curve connecting triangular dots concerns problems with a careless error rate less than 19%. The curve connecting square dots concerns problems with a careless error rate greater than or equal to 19%.

66%. Figure 2.3 traces the Vincent curves for the median values of the point biserial correlation of each group.

The figure makes it clear that items of medium difficulty yield better correlations than items that are either easier or more difficult, with ending correlation values of .51, .44, and .44, respectively. Let us recall that the assessments under examination are initial assessments that took place before any learning. In other words, the population who took them had a limited knowledge of the curriculum as evidenced by most items being classified in the higher difficulty group.

2.4.4 The final state predictions. The Vincent curve analysis that we just discussed does not tell us how successful the assessment ultimately is. We now consider the final result of the assessment, that is, the knowledge state chosen by the assessment engine, and investigate how predictive of the competence of a student that final state is. We still rely on the extra problem methodology. The relevant data for a particular extra problem **a** is a 2×2 matrix of the form of Table 2.2:

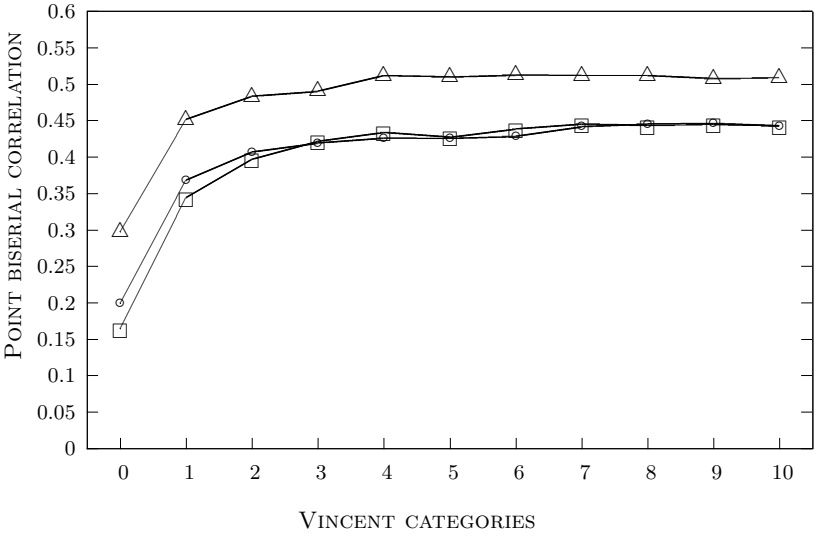

Figure 2.3. Vincent curves similar to the one of Figure 2.1. The curves connecting circular, triangular, and square dots concern problems of lower, medium, and higher difficulty, respectively.

Table 2.2. Basic data matrix for the computation of the correlation between the cases 'in or out of the final state' and the student's response coded as 0/1 for false/correct by the variable $\mathbf{R_a}$.

		$\mathbf{R_a}$	
		0	1
State	**a** in	x	y
	a out	z	w

The letter x in this matrix represents the number of cases in which the extra problem **a** was in the final state assigned to the student and the response was incorrect; y, z, and w have similar interpretations. From the standpoint of the assessment engine, x can thus be regarded as the number of careless errors committed in the $x + y + z + w$ cases in which **a** was presented.

As no correlation coefficient is available that would be completely adequate for our situation and without bias of some sort, we have used three of them. The first two, which are the tetrachoric and the phi coefficients, act directly on matrices of the type displayed in Table 2.2 and do not take the careless errors into account. The third coefficient is the point biserial. It is applied

to the same data matrices but involves a correction for careless errors. The results are described in Paragraph 2.4.5.

The data analyzed in this section are based on the same 125,786 assessments used for the Vincent curve analysis. For the purpose of this analysis, the assessments were replayed with a simple extracting rule that ascribes to the student's knowledge state any item $\boldsymbol{q}$ such that

$$\sum_{K \in \mathcal{K}_{\boldsymbol{q}}} \xi_n(K) > 0.5,$$

where ξ_n is the final probability distribution on the knowledge structure over the sub-domain containing $\boldsymbol{q}$. (For practical reasons, assessments sometimes stopped while there were still items for which the left hand side of the above inequality was around .5, making them potential questions to ask. Such items would typically not be ascribed to the student's actual knowledge state.)

Figure 2.4 displays the distribution of the tetrachoric coefficient values pertaining to 324 problems out of the 370 problems forming the elementary school mathematics domain in ALEKS. Forty-six problems were discarded because the relevant data were too sparse for reliably estimating the coefficient. The tetrachoric coefficient is based on the hypothesis of an underlying 2-dimensional Gaussian random variable. Thus, the double dichotomy matrix arises from splitting each of the two Gaussian marginals into two categories[15]. The median of the distribution is around .68, which is quite high in such a context. The grouped data, obtained from gathering the 324 individual 2×2 matrices into one, yields a still higher correlation of about .80. For high values, the tetrachoric coefficient is sometimes regarded as biased upward (however, see Greer et al., 2003).

Figure 2.5 contains a similar analysis using the phi coefficient. All the values are noticeably lower, yielding a median of .43 (in contrast to the .68 value obtained for the tetrachoric coefficient) and a grouped data value of .58 (instead of .80). Of particular interest are the very low correlation values obtained for some problems. In our view, any problem with a correlation below .2 for this coefficient deserves some examination. A low correlation value could be due, for example, to a high careless error rate or a misplacement in the structure. We go back to this issue in Paragraphs 2.4.5 and 2.4.6.

[15] Admittedly, this hypothesis does not fit the situation very well. See Drasgow (1988) or Nunnally and Bernstein (1994) for a discussion of the hypotheses underlying the use of the tetrachoric and the phi coefficients.

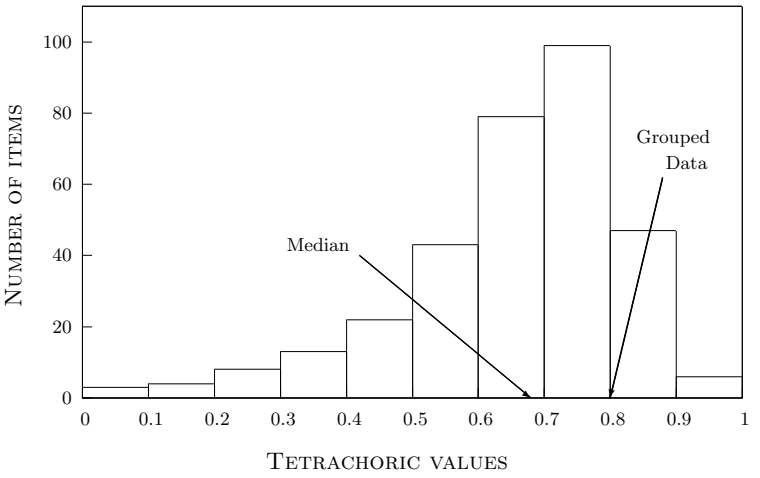

Figure 2.4. Correlation between the in/out cases for the final knowledge state and the actual false/correct response of the student. The figure displays the distribution of the tetrachoric coefficient values for 324 of the 370 problems forming the domain of the learning space for elementary school mathematics.

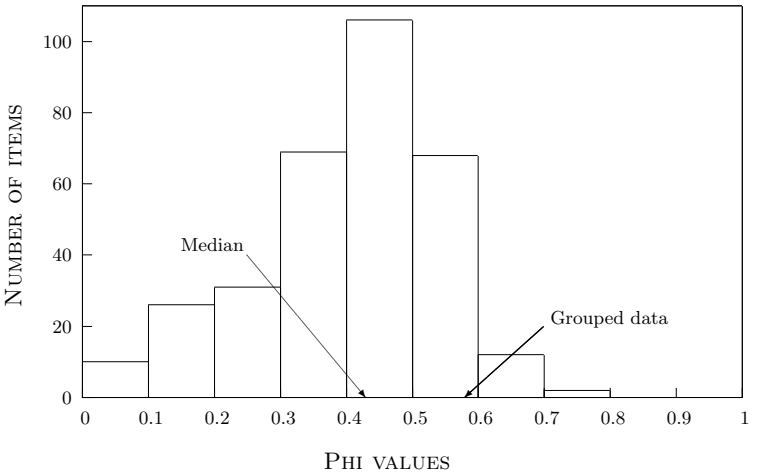

Figure 2.5. The similar distribution, based on the same data, for the phi coefficient.

2.4.5 Correcting for careless errors. The analyses in terms of the tetrachoric and phi coefficients that we just described did not take the careless errors into consideration. However, the basic data matrix (2.2) can be amended to include the effect of careless errors. To this end, we define the new variable

$$\mathbf{S_a} = \begin{cases} (1 - \epsilon_{\mathbf{a}}) & \text{if the final state contains the extra question } \mathbf{a} \\ 0 & \text{otherwise.} \end{cases}$$

Thus, $\mathbf{S_a}$ is the probability of not committing a careless error on the extra problem **a** computed based on the final state at the end of the assessment. The variable $\mathbf{S_a}$ is neither exactly continuous nor exactly discrete[16]. Nevertheless, again for the purpose of comparison with similar analyses performed in psychometric situations, we have used the point biserial coefficient r_{pbis} to compute the correlation between the variables $\mathbf{S_a}$ and $\mathbf{R_a}$ for the grouped data over the same 143 items used in the first paragraph of Section 2.4.3. The value obtained for r_{pbis} was .67, sensitively higher than the .57 obtained for the phi coefficient for the same grouped data.

2.4.6 Validity and learning readiness. Finally, we discuss an indirect but nevertheless revealing way of gauging the validity of an assessment. We recall from Chapter 1 that the outer fringe of a knowledge state is the set of problems that the student in that state is ready to learn[17]. Consider a situation in which an assessment is a prelude to learning and the system routinely prompts the student to start learning how to solve the problems in the outer fringe of her state. The capability of the student to learn ("master") such problems should be revealing of the validity of the assessment. This may be evaluated by the conditional probability that a student is capable of mastering a problem, given that it is located in the outer fringe of her knowledge state (and so accessible for learning). These probabilities can be estimated from our learning data. For elementary school mathematics, the median of the distribution of the estimated (conditional) success probabilities is .93. To understand the import of this number, some details about the learning process in the system examined here must be given.

When a problem is located in the outer fringe of a student's state, the student may select that problem as the next one to learn. This choice initiates a random walk keeping track of the learning stages for that problem. The random walk takes place on an interval of the integers and has two absorbing barriers (see Figure 2.6). At each step of that random walk, an instance of the problem is proposed and the student is asked to solve it. In case of failure, an explanation of the solution that is centered on that instance is offered. The problem enters the random walk at the point 0 and moves left or right depending on the student's response to the instance presented. 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 problem is considered to be learned when the random walk hits the right barrier. Hitting the left barrier means

[16] For example, the distribution of $\mathbf{S_a}$ vanishes in a positive neighborhood of 0, but is positive at the point 0 itself.

[17] See Paragraph 1.2.1 on page 11 for an introduction to this concept, and Definition 8.3.1 on page 136 for a formal definition.

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.

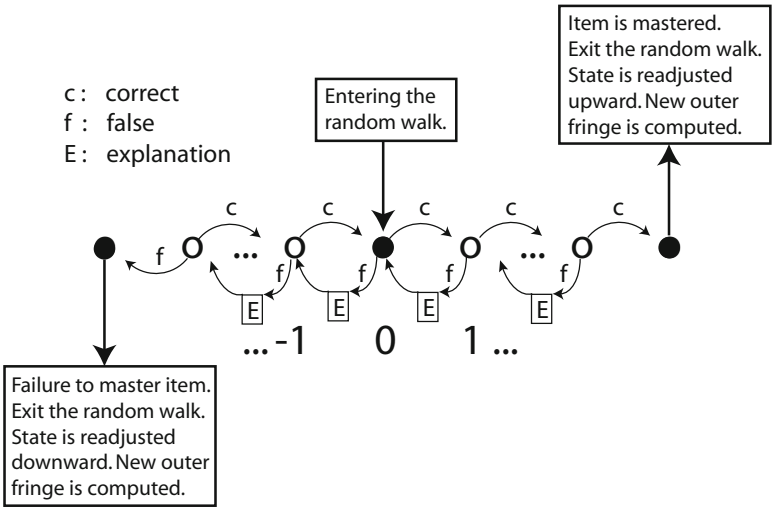

Figure 2.6. Illustration of the random walk on the integers with two absorbing barriers. The left barrier corresponds to the failure to master the problem at this time, and the right barrier to the success. The random walk keeps track of the intermediate learning stages for a problem. The problem enters the random walk at the point 0 and moves left or right depending on the student's response to the instance presented. In case of a false response, an explanation is given to the student, and the random walk moves one unit to the left. A correct response initiates a move of one unit to the right. The reader should keep in mind that the successive instances proposed to the student may be quite different.

Figure 2.7 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: 90% of them have a probability of success of at least .83, with the median of the distribution at .93. Nevertheless, the left tail of the distribution indicates that a few problems are not learned easily and that adjustments deserve to be made. For example, some intermediate problems may be missing and should then be added to the domain. Also, a given problem may be misplaced in the structure, or its explanation may be defective and should be rewritten. Care is taken in the design of the various instances of a problem so that a correct response would not be due to a trivial device,

unrelated to the understanding of the problem. The data analyzed are based on 1,940,473 such random walks.

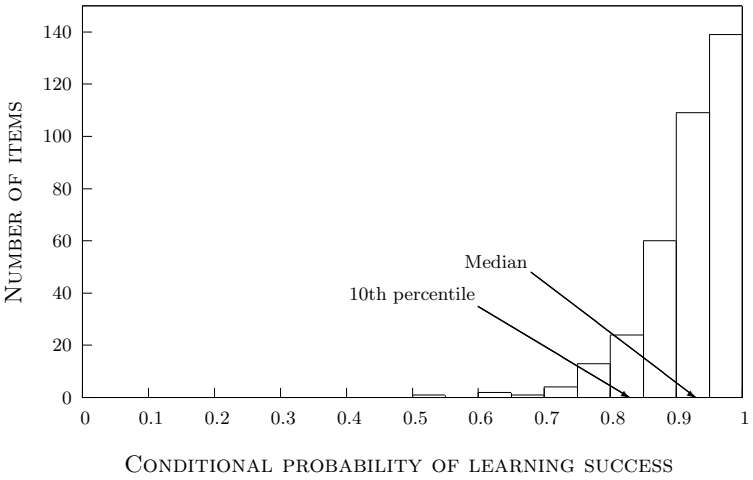

Figure 2.7. For the 353 items in elementary school mathematics (out of 370), the distribution of the estimated values of the conditional probabilities that a student entering the random walk reaches its right barrier. The problem is then regarded as having been mastered. The median probability is .93, with the 10th percentile at .83. These data are based on 1,940,473 random walks.

2.5 Summary and Discussion

The aim of this work was to evaluate the extent to which an assessment in a learning space, performed by the `ALEKS` system, is predictive of a student's mastery of a scholarly subject. In other words, we wanted to appraise the validity of such an assessment. We took elementary school mathematics as the exemplary subject. The method used for the appraisal was systematically to ask the student an extra problem $\boldsymbol{p}$, randomly selected, and predict the student's response to that problem on the basis of the rest of the information provided by the assessment. The accuracy of the prediction was measured by correlation coefficients (three of them were used in different parts of the study). Two aspects of the data were taken into account for this correlation analysis.

First, we computed Vincent curves tracing the evolution of the median point biserial correlations during the assessment. The graph of Figure 2.1

shows that, beginning with a median correlation of about .19, the median grows steadily to an asymptote that appears to be around .46. We also computed Vincent curves comparing problems with respect to their careless error rate (Figure 2.2) and with respect to their overall difficulty (Figure 2.3). The overall difficulty turned out to account for a greater variability of the correlation than the careless error rate. Specifically, items of medium difficulty (defined as items which were answered correctly between 33% and 66% of the time when asked as the extra problem) reached a median point biserial correlation of .51, significantly better than the median value of .44 for the other items.

Second, we relied again on the extra problem methodology, but with a different predictor, namely, the final knowledge state chosen by the assessment engine. Our basic data, for each of 324 problems (out of 370)[18], take the form of 2×2 matrices, or double dichotomies. Whether or not the selected state contains the extra problem provides the first dichotomy. The student's response to that problem—false or correct—yields the second one. The distributions of the correlations for the 324 problems were computed using the tetrachoric coefficient (Figure 2.4) and the phi coefficient (Figure 2.5). The median correlation and the grouped correlation (gathering all the 324 contingency matrices into one) were computed for both coefficients. The values obtained are recalled below:

	tetrachoric	phi
median	.68	.43
grouped data	.80	.58

These correlations do not take into account the careless errors. So, we also computed the correlation between the response to the extra problem and a different prediction variable, which also depends on the final state of the student. This variable has value 0 when the final state of assessment $\mathbf{a}$ does not contain the extra problem $\boldsymbol{p}_{\mathbf{a}}$, and value $1 - \epsilon_{\mathbf{a}}$ (the probability that no careless error occurs) when the final state contains $\boldsymbol{p}_{\mathbf{a}}$. We used the point biserial coefficient to evaluate this correlation and obtained $r_{pbis} = .67$, which is a high number in a psychometric context.

Our last analysis was of a different type, and its bearing on the validity of the assessment, while indirect, is important in practice. If the final state selected by the assessment is a valid representation of a student's competence, then the outer fringe of that state should contain problems that the student is ready to learn. Thus, if a student chooses to work on one of these problems, the probability of success should be high. We computed these probabilities for 1,940,473 learning experiences. The median probability of learning success was .93. The details are given by Figure 2.7.

[18] As indicated, 46 problems were discarded because the correlation coefficients could not be reliably computed.

While overall the validity results presented in this chapter should be regarded as satisfactory, an examination of Figures 2.1, 2.5, and 2.7 reveals the weaknesses of some items: the correlations measuring the validity are too low, and the probability of learning success is also too low. As argued earlier in this chapter, eliminating low performing items is not part of a solution: all items are included in the domain because they are considered an integral part of the curriculum. In particular, Figure 2.3 tells us that items of medium difficulty with respect to the population yielded better point biserial correlations. We only considered in this analysis initial assessments and a large majority of items were of high difficulty with respect to the current knowledge of the students. We may expect that, tested against the same students at a later stage after they have learned and increased their knowledge, the point biserial correlation of some of these items would show improvement. For general improvements, we discuss possible remedies below.

On improving the items and the structure. Let us first consider the careless errors. One might posit that a high value of $\epsilon_{\mathbf{a}}$ might be due to a mistake in the placement of item **a** in the structure of the learning space. Actually, this is unlikely in view of the procedure used to estimate the careless error probabilities, which does not involve the learning space. Indeed, the four equations (2.8)-(2.11) of Paragraph 2.3.3 rely on a dummy parameter $\gamma_{\boldsymbol{q}}$ measuring the probability that the knowledge state contains $\boldsymbol{q}$. The estimated value of this parameter resulting from minimizing the chi-square expression (2.12) has no formal relation with the structure. The most plausible hypothesis is that the high careless error of an item is intrinsic. Assuming this, there are two major reasons why an item could have a high careless error rate.

1. The instances of an item are not homogeneous. This could arise, for example, if the instances are not of equal difficulty, or if some of them are ambiguously phrased. So, an item may be identified largely by its "good" instances, but a "bad" instance will occasionally be asked as the extra problem. The cure is straightforward, if tedious: all the instances of items having a high careless error rates must be carefully examined, and adjustments made if need be.
2. The nature of an item induces possible careless errors. This could arise, for example, when an item involves several numerical operations, each of which can elicit a careless error. This is unavoidable but there are ways to mitigate the high careless error rates of such items. `ALEKS` already uses specialized feedbacks, in some cases, to induce the student to check for careless errors. Such feedbacks give no hint of what the correct response is. A more thorough use of such feedbacks could probably palliate some of these high careless error rates.

There are also a couple of reasons for an item to have a low correlation value (whether in the form of the point biserial, tetrachoric, or phi) and/or a poor

learning success rate (say, below the 10th percentile in Figure 2.7). First, a lack of instance homogeneity could contribute to either situation. Second, and in contrast to high careless error rates, low correlation or learning success values could certainly be due to a misplacement of an item in the structure. This means that the item must be removed from some states, and possibly added to some other states. While such a restructuring is a substantial enterprise, the tools are available to achieve it. We only sketch the main steps here. Suppose that some item q is distinguished as misplaced in a learning space $\mathcal{K}$. Removing this item from the domain of $\mathcal{K}$ creates a structure $\mathcal{K}^{-q}$ that is still a learning space; that is, Axioms [L1*] and [L2*] are satisfied. We can now regard $\mathcal{K}^{-q}$ as if it were the result of a partial construction of $\mathcal{K}$ and use the standard tools to complete the construction; that is, add the item q and give it a better placement in the structure. One of these tools is the QUERY algorithm[19], which can be used either with human experts, for whom it was originally conceived, or with statistics of conditional probabilities of solving or failing to solve problems, estimated from the data. The original implementation of the QUERY algorithm resulted in a structure that was a knowledge space but not necessarily a learning space. However, the current version of the algorithm corrects this shortcoming and substantially eases the completion process. Alternatively, how to complete a knowledge space into a learning space by a minimal addition of states formed from items in the same domain is the topic of Eppstein et al. (2009).

Finally, we must consider the learning success data reported in Figure 2.7, which shows that not all of the items are successfully learned when selected in the outer fringe of the student. Any item with a success rate below .8 may be regarded as flawed in some fashion. There are about 6% of such items. It is possible that these items are also misplaced in the structure and that the defect would be corrected by restructuring the learning space. A second possibility is that the explanation of these items is not clear enough and should be rewritten. A third possibility is that the domain lacks 'granularity' with respect to the skills tested by theses items and that the addition of 'step items' would benefit their learning.

From the analysis presented in this chapter, the `ALEKS` system appears to provide a valid assessment of what a student knows, does not know, and is ready to learn. Especially noteworthy are the high probabilities of learning success revealed by Figure 2.7 for all but a few items in elementary school mathematics. The system is not perfect: we have remarked that some items deserve improvements, whether in their placement in the structure, their instance formulation or their explanation. This discussion indicates that the means are available to achieve the necessary developments.

[19] Cf. page 25 in the Glossary.

2.6 Appendix

We recall that $N_{\boldsymbol{q}}$ stands for the number of assessments having at least two presentations of item $\boldsymbol{q}$, with one of them as the extra problem, and $N_{\boldsymbol{q}}(i,j)$, $i,j \in \{0,1\}$, stands for the number of times (i,j) is realized among the $N_{\boldsymbol{q}}$ assessments[20]. We estimate for each item $\boldsymbol{q}$ the careless error probability $\epsilon_{\boldsymbol{q}}$ and the probability $\gamma_{\boldsymbol{q}}$ that the students has mastered item $\boldsymbol{q}$ by minimizing the Chi-square statistic

$$\begin{aligned}\texttt{Chi}_{\boldsymbol{q}}(\epsilon_{\boldsymbol{q}},\gamma_{\boldsymbol{q}}) &= \sum_{i,j}\frac{\left(N_{\boldsymbol{q}}(i,j)-N_{\boldsymbol{q}}p_{\boldsymbol{q}}(i,j)\right)^2}{N_{\boldsymbol{q}}p_{\boldsymbol{q}}(i,j)} \qquad (2.16)\\ &= \frac{\left(N_{\boldsymbol{q}}(0,0)-N_{\boldsymbol{q}}(\epsilon_{\boldsymbol{q}}^2\gamma_{\boldsymbol{q}}+1-\gamma_{\boldsymbol{q}})\right)^2}{N_{\boldsymbol{q}}(\epsilon_{\boldsymbol{q}}^2\gamma_{\boldsymbol{q}}+1-\gamma_{\boldsymbol{q}})}+\frac{\left(N_{\boldsymbol{q}}(0,1)-N_{\boldsymbol{q}}\epsilon_{\boldsymbol{q}}(1-\epsilon_{\boldsymbol{q}})\gamma_{\boldsymbol{q}}\right)^2}{N_{\boldsymbol{q}}\epsilon_{\boldsymbol{q}}(1-\epsilon_{\boldsymbol{q}})\gamma_{\boldsymbol{q}}}\\ &+\frac{\left(N_{\boldsymbol{q}}(1,0)-N_{\boldsymbol{q}}(1-\epsilon_{\boldsymbol{q}})\epsilon_{\boldsymbol{q}}\gamma_{\boldsymbol{q}}\right)^2}{N_{\boldsymbol{q}}(1-\epsilon_{\boldsymbol{q}})\epsilon_{\boldsymbol{q}}\gamma_{\boldsymbol{q}}}+\frac{\left(N_{\boldsymbol{q}}(1,1)-N_{\boldsymbol{q}}(1-\epsilon_{\boldsymbol{q}})^2\gamma_{\boldsymbol{q}}\right)^2}{N_{\boldsymbol{q}}(1-\epsilon_{\boldsymbol{q}})^2\gamma_{\boldsymbol{q}}}.\end{aligned}$$

Thus, the probabilities $p_{\boldsymbol{q}}(i,j)$ in (2.16) are defined by (2.8)-(2.11).

We obtain the minimum of $\texttt{Chi}_{\boldsymbol{q}}(\epsilon_{\boldsymbol{q}},\gamma_{\boldsymbol{q}})$ by the Lagrange multipliers method. We simplify the notation and set

$$x = N_{\boldsymbol{q}}(\epsilon_{\boldsymbol{q}}^2\gamma_{\boldsymbol{q}}+1-\gamma_{\boldsymbol{q}}) \qquad (2.17)$$

$$y = N_{\boldsymbol{q}}\epsilon_{\boldsymbol{q}}(1-\epsilon_{\boldsymbol{q}})\gamma_{\boldsymbol{q}} \qquad (2.18)$$

$$z = N_{\boldsymbol{q}}(1-\epsilon_{\boldsymbol{q}})^2\gamma_{\boldsymbol{q}}. \qquad (2.19)$$

Notice that the ratio of the last two equations only depends upon $\epsilon_{\boldsymbol{q}}$ since

$$\frac{z}{y}=\frac{1-\epsilon_{\boldsymbol{q}}}{\epsilon_{\boldsymbol{q}}},$$

and so

$$\hat{\epsilon}_{\boldsymbol{q}} = \frac{y}{z+y}. \qquad (2.20)$$

Replacing $\epsilon_{\boldsymbol{q}}$ in (2.18) by its expression in (2.20) yields

$$y = N_{\boldsymbol{q}}\frac{yz}{(z+y)^2}\gamma_{\boldsymbol{q}}.$$

Canceling the y's and rearranging, we obtain

$$\hat{\gamma}_{\boldsymbol{q}} = \frac{(z+y)^2}{zN_{\boldsymbol{q}}}. \qquad (2.21)$$

[20] Thus, for example, $(1,0)$ stands for the event that the student responds correctly to the first presentation of $\boldsymbol{q}$ in the assessment, and incorrectly to the second one.

Thus, $\epsilon_{\boldsymbol{q}}$ and $\gamma_{\boldsymbol{q}}$ are defined by y and z. To obtain the values of y and z, we minimize the function

$$(x, y, z) \mapsto \frac{(N_{\boldsymbol{q}}(0,0) - x)^2}{x} + \frac{(N_{\boldsymbol{q}}(0,1) - y)^2}{y} + \frac{(N_{\boldsymbol{q}}(1,0) - y)^2}{y} + \frac{(N_{\boldsymbol{q}}(1,1) - z)^2}{z}$$

with respect to x, y, and z, subject to the constraint

$$x + 2y + z = N_{\boldsymbol{q}}.$$

To this effect, we define

$$\begin{aligned} \Lambda(x, y, z, \lambda) = {} & \frac{(N_{\boldsymbol{q}}(0,0) - x)^2}{x} + \frac{(N_{\boldsymbol{q}}(0,1) - y)^2 + (N_{\boldsymbol{q}}(1,0) - y)^2}{y} \\ & + \frac{(N_{\boldsymbol{q}}(1,1) - z)^2}{z} + \lambda(x + 2y + z - N_{\boldsymbol{q}}). \end{aligned} \tag{2.22}$$

We then compute the derivatives of Λ with respect to its four variables and solve, with respect to x, y, z, and λ, the system of the four equations

$$\frac{d\Lambda}{x} = \frac{d\Lambda}{y} = \frac{d\Lambda}{z} = \frac{d\Lambda}{\lambda} = 0.$$

This gives

$$\frac{N_{\boldsymbol{q}}(0,0)^2}{x^2} = \frac{N_{\boldsymbol{q}}(0,1)^2 + N_{\boldsymbol{q}}(1,0)^2}{2y^2} = \frac{N_{\boldsymbol{q}}(1,1)^2}{z^2} = \lambda + 1, \tag{2.23}$$

leading to

$$\frac{z}{y} = \sqrt{\frac{2N_{\boldsymbol{q}}(1,1)^2}{N_{\boldsymbol{q}}(0,1)^2 + N_{\boldsymbol{q}}(1,0)^2}}. \tag{2.24}$$

Replacing $\frac{z}{y}$ in (2.20) by its expression in (2.24) gives

$$\hat{\epsilon}_{\boldsymbol{q}} = \frac{\sqrt{N(0,1)^2 + N(1,0)^2}}{\sqrt{2N(1,1)^2} + \sqrt{N(0,1)^2 + N(1,0)^2}}. \tag{2.25}$$

The derivations giving an estimation formula for $\gamma_{\boldsymbol{q}}$ are similar. From (2.17), (2.19) and (2.23), we get, after canceling the $N_{\boldsymbol{q}}$'s,

$$\frac{x}{z} = \frac{N_{\boldsymbol{q}}(0,0)}{N_{\boldsymbol{q}}(1,1)} = \frac{\epsilon_{\boldsymbol{q}}^2 \gamma_{\boldsymbol{q}} + 1 - \gamma_{\boldsymbol{q}}}{(1 - \epsilon_{\boldsymbol{q}})^2 \gamma_{\boldsymbol{q}}}.$$

The last equation is linear in $\gamma_{\boldsymbol{q}}$. Solving this equation for $\gamma_{\boldsymbol{q}}$ and replacing $\epsilon_{\boldsymbol{q}}$ by its expression in (2.25) gives finally, after rearranging,

$$\hat{\gamma}_{\boldsymbol{q}} = \frac{\left(\sqrt{2N_{\boldsymbol{q}}(1,1)^2} + \sqrt{N_{\boldsymbol{q}}(0,1)^2 + N_{\boldsymbol{q}}(1,0)^2}\right)^2}{2N_{\boldsymbol{q}}(1,1)\left(N_{\boldsymbol{q}}(0,0) + N_{\boldsymbol{q}}(1,1) + \sqrt{2(N_{\boldsymbol{q}}(0,1)^2 + N_{\boldsymbol{q}}(1,0)^2)}\right)}. \tag{2.26}$$