

A Comparison of Non-Deterministic Procedures for the Adaptive Assessment of Knowledge¹

Cord Hockemeyer²
Karl-Franzens-Universität Graz

Zusammenfassung

Die Wissensraumtheorie wurde von Doignon und Falmagne als Grundlage für adaptive Wissensdiagnose entwickelt. Seit 1986 wurden verschiedene Verfahren für die deterministische und nicht-deterministische Diagnose entwickelt. Bislang war die Forschung im Bereich der nicht-deterministischen Diagnosealgorithmen auf Wissensräumen theoretischer Natur. In diesem Beitrag werden erste Ergebnisse einer Simulationsstudie zum Vergleich verschiedener nicht-deterministischer Diagnoseverfahren beschrieben. Vergleichskriterien sind hierbei die Zahl der gestellten Fragen und die Genauigkeit und Robustheit der Diagnoseverfahren.

Schlüsselwörter: Adaptive Diagnose - Wissensraumtheorie - nicht-deterministische Verfahren

Abstract

The theory of knowledge spaces has been developed by Doignon and Falmagne as a basis for the adaptive assessment of knowledge. Since 1986, several procedures for deterministic and non-deterministic assessment have been developed. However, most research on non-deterministic assessment algorithms has been of theoretical kind. This paper presents first results of a simulation study comparing the performance of various non-deterministic procedures for adaptive assessment on empirical knowledge spaces. Criteria for the comparisons are the number of questions to be asked and the accuracy and robustness of the assessment procedure.

Keywords: Adaptive assessment - Knowledge Space Theory - non-deterministic procedures

¹ Part of the work described in this paper was done at the Institut für Psychologie, Technische Universität Braunschweig and financially supported through a grant of the Ministère de l'Éducation, Luxembourg to Cornelia E. Dowling. This paper is the revised and extended version of a manuscript of a talk given at the 30th European Mathematical Psychology Group (EMPG) Meeting 1999 in Mannheim.

² Cord Hockemeyer, Institut für Psychologie, Universitätsplatz 2/III, A-8010 Graz, URL: <http://css.uni-graz.at/staff/hockemeyer/>, Email: Cord.Hockemeyer@uni-graz.at.

Introduction: The Theory of Knowledge Spaces

A central idea behind adaptive assessment of knowledge is that of a teacher assessing an individual student's knowledge. Such a teacher would start with a question of intermediate difficulty and would select subsequent questions depending on the student's prior answers. Thus, the teacher could assess the student's knowledge by asking only a relatively small number of questions. Having in mind such an adaptive assessment, Doignon and Falmagne (1985, 1999) have developed the theory of knowledge spaces. Their basic idea is to structure a domain of knowledge by specifying *prerequisite relationships* between different items of knowledge. In a first approach, such prerequisite relationships may be defined through *surmise relations*: two items a and b are in a surmise relation (aSb) if, from a student mastering item b , we can surmise that this student also masters item a . Thus, the surmise relation is equivalent to a set of rules of the form "if a student masters b this student will also master a ". In a second, more elaborated approach, this concept has been extended to *surmise mappings* specifying alternative sets of prerequisites for an item. These surmise mappings are equivalent to rules of the form "if a student masters a this student will also master b_1, \dots, b_n or c_1, \dots, c_m or ...".

If the *knowledge state* of a student is defined as the subset of items which this student is able to master then the set of possible knowledge states is limited by the aforementioned rules. If we have, for example, aSb for two items a and b , i.e. a is a prerequisite for b then no knowledge state is possible that contains b but not a . The set of all possible knowledge states within a given domain of knowledge and its prerequisite relationships is called *knowledge space*. Knowledge spaces and surmise mappings are equivalent representations for the structure of a domain of knowledge. Knowledge spaces have been investigated in detail by Doignon and Falmagne (1985, 1999). Various methods for a knowledge space for a given domain obtaining (see, e.g., Dowling, 1993; Albert & Lukas, 1999) as well as for applying such a knowledge space in adaptive testing, training, and teaching (see, e.g., Dowling, Hockemeyer, & Ludwig, 1996; Hockemeyer, Albert, & Held, 1998) have been suggested. In the sequel, I will focus on the application of knowledge spaces for the adaptive assessment of knowledge.

Procedures for the Adaptive Assessment of Knowledge

Looking at procedures for the adaptive assessment of knowledge based on knowledge spaces we distinguish between *deterministic* and *non-deterministic* procedures. The former assume that the answer behaviour of the student is completely determined by the student's knowledge while the latter take into consideration also other influences or noise, for example careless errors and lucky guesses.

Deterministic Assessment of Knowledge

Two different approaches for the deterministic assessment of knowledge have been suggested. Dowling and Hockemeyer (2001) have introduced a procedure based on the rule-oriented representation of surmise mappings. A student's answers are propagated through the network of rules specified by the surmise mapping, and new questions are selected from those items for which it cannot yet be decided whether or not they are mastered by the student. Subsequent questions are selected through a *medium difficulty* rule where the difficulty of an item is computed through the number of its prerequisites for which it is not yet known whether the student masters them or not. Items for which the mastery by the current student is already known are not considered at all.

Another procedure suggested earlier by Degreeef, Doignon, Ducamp and Falmagne (1986) is based on the knowledge space, i.e. the set of knowledge states consistent with the prerequisite relationships specified through a surmise mapping. They reduce the set of possible knowledge states stepwise by eliminating those states which are not in accordance with the received answers. In this procedure, new questions are selected from those items which are in some but not all of the knowledge states still under consideration. An optimal rule for selecting new questions is to take those items which are contained in approximately half of the states still under consideration (*halfsplit rule*).

Dowling and Hockemeyer (2001) have shown that both propagating mechanisms are equivalent. Furthermore, they are both irredundant, i.e. new information is gained with each question as long as new questions are selected as described above. This is also reflected by the fact that the number of questions needed for a complete assessment is close to the optimal value which is given by the height of a balanced decision tree, i.e. the binary logarithm of the number of knowledge states.

A comparison of the deterministic procedures through simulation studies (Dowling and Hockemeyer, 2001) has shown that due to the different query rule their approach needs slightly more questions for a complete assessment. On the other side, it is much faster in selecting new questions, especially for larger knowledge spaces. In some cases it may even be the only possibility for capacity reasons as knowledge spaces can grow exponentially with the number of items.

Non-Deterministic Assessment of Knowledge

Falmagne and Doignon (1988a,b; Doignon, 1994) have suggested two different approaches to non-deterministic assessment of knowledge based on knowledge space theory. Both of these approaches are based on the deterministic procedure by Degreeef et al. (1986).

A first, *discrete* procedure (Falmagne & Doignon, 1988b; Doignon, 1994) starts with a deterministic assessment. Afterwards, it is assumed that despite any noise that might occur (e.g. careless errors or lucky guesses) the result of this deterministic assessment is quite close to the correct result. This means that errors

should be somewhere in the *fringe* of the preliminarily resulting knowledge state, i.e. they should be in some items distinguishing that state from its neighbouring states in the knowledge space. Therefore, such items are tested updating the preliminary knowledge state whenever necessary. Falmagne and Doignon (1988a) have shown that, for knowledge spaces with certain properties, this procedure will converge to the correct knowledge state.

A second, *continuous* approach (Falmagne & Doignon, 1988a) is based on a probability distribution estimate over the complete knowledge space, i.e. for each knowledge state the probability is estimated that this knowledge state describes the current student. During the assessment process, the probability distribution estimate is updated by applying, for example, the Bayesian update rule assuming item-specific fixed probabilities for careless errors and lucky guesses. From the estimated probability distribution over the knowledge space, estimates for the probability of each item to be mastered by the student can easily be computed. The halfsplit rule for selecting new questions can be applied here by selecting those items which have a probability of about .5 to be mastered by the current student. The deterministic procedure of Degreef et al. (1986) as described above is a special case of the continuous procedure where all probabilities for careless errors and lucky guesses are 0.

While Falmagne and Doignon (1988a,b) have investigated the theoretical properties of these approaches quite in detail, their practical usability in terms of, for example, robustness, accuracy, or efficiency has been neglected. The simulation study described below shall be a first step into this direction.

A Simulation Study comparing Non-Deterministic Assessment Procedures

The two non-deterministic assessment procedures were compared in a simulation study using empirical knowledge spaces. For this purpose, several implementation issues had to be decided that had been left open by Falmagne and Doignon in their theoretical papers (1988a,b). An issue open for both procedures was the criterion when an assessment should be considered as finished. For the discrete procedure, a certain level of stability in the second phase of the assessment process was selected, i.e. the assessment was interrupted when the preliminarily assessed knowledge state kept unchanged for a certain number of questions. For the continuous procedure, the concentration of probability mass onto a single state was selected as criterion, i.e. the assessment was interrupted when one knowledge state reached a certain estimated probability of being the knowledge state of the simulated student. For the latter procedure, also update and query function had to be decided upon as Falmagne and Doignon had proposed a class of different procedures. As an update function, the Bayesian update rule was selected; for the query function, the halfsplit rule was taken which ensures higher comparability of the two procedures. These implementation decisions lead to two programs *discrete-assess* and *halfsplit-assess*

sharing most of their source parts and also leaving the possibility to implement different versions especially for the continuous procedure. Simulations were run using three knowledge spaces on the domain "usage of AutoCAD³". These knowledge spaces were obtained by Dowling (1991) through querying different experts on prerequisite relationships. For each of these knowledge spaces, student were simulated with probabilities for lucky guesses and careless errors of 0, 5, and 10% maximum (*noise*). For each item, the noise probabilities were selected randomly on an equal distribution between 0 and the respective maximum. For each of these conditions, 1000 students were simulated.

In the assessment, both procedures were tested with three different levels of interrupt criteria used. The discrete procedure was interrupted after 1, 3, or 5 loops without a change in the preliminarily assessed state, the continuous procedure was interrupted after one state had reached an estimate of at least 51%, 70%, or 90% probability to be the state of the assessed student (*strictness*). For the continuous procedure, the same probabilities for noise were used as they had been applied in the simulation. All simulated students were assessed with all procedures and assessment parameters.

For each assessment, two values were obtained: the assessment errors (number of wrongly assessed items for a student) as a measure for the accuracy of the assessment result and the number of questions posed to reach this result. It was expected that the accuracy should increase with higher strictness and decrease with higher noise. The number of questions should increase with higher strictness and should be rather independent of the noise. Furthermore, it can be expected to obtain generally better results for the continuous procedure because the discrete procedure is based on a much coarser estimate of probabilities: this implicit estimate assigns to each state currently under consideration the probability of $1/n$ where n is the number of knowledge states currently under consideration, and the probability of 0 to all other knowledge states.

Results of the Simulation Study

The evaluation of results from simulation studies faces a methodological problem: the easiness with which an arbitrary high number of simulated students can be simulated changes any test for statistical significance into a mere function of the aforementioned number of simulations. Therefore, only some typical examples will be presented.

Accuracy of the Assessment

The accuracy of the results obtained from the assessment procedure is the most important criterion for the quality of the procedure. Table 1 presents average assessment error rates for simulations with an AutoCAD knowledge space with 28 items and 2261 knowledge states, and with minimal strictness (1 loop or 51%) for

³ AutoCAD is a trademark of AutoDesk.

different maximal noise probabilities. As it can be easily seen, the error rate increases with increasing noise as it was expected. Furthermore, the table shows a tendency towards higher accuracy for the continuous procedure, especially for high noise rates.

Table 1
Average assessment errors for different noise rates

Maximal noise rates [%]	Assessment errors (Percentage of correct assessments; max err.)	
	Discrete procedure	Continuous procedure
0	0 (100; 0)	0 (100; 0)
5	0.44 (81.1; 9)	0.45 (79.8; 8)
10	1.00 (59.3; 9)	0.90 (61.5; 11)

Remark: The simulations were run for a knowledge space with 2261 knowledge states applying a minimal strictness criterion for finishing the assessment process.

The distributions of assessment errors are quite skew. As a consequence, standard deviations cannot be used to describe these distributions in more detail. For all simulations shown here, the median of the assessment error is 0; to give an idea of the distributions, the percentage of correct answers, and the maximum number of assessment errors for the respective simulation conditions are specified in Table 1 and 2.

The second variable important for the accuracy of the assessment result is the strictness of the interrupt criterion for the assessment. Table 2 shows for the same knowledge space and for a medium noise level of maximal 5% that the average assessment error decreases as expected with higher strictness. Furthermore, the overall performance of the discrete procedure is - comparably to Table 1 - weaker than that of the continuous procedure, notably the effect of applying a stricter interrupt criterion is stronger.

Table 2
Average assessment errors for different strictness criteria

Strictness	Assessment errors (Percentage of correct assessments; max err.)	
	Discrete procedure	Continuous procedure
Low	0.44 (81.1; 9)	0.45 (49.8; 8)
Medium	0.40 (84.0; 8)	0.39 (80.1; 7)
High	0.40 (81.7; 9)	0.17 (89.8; 6)

Remark: The simulations were run for a knowledge space with 2261 knowledge states applying a maximal noise rate of 5% for careless errors and lucky guesses. Efficiency of the Assessment

The efficiency of the assessment which is measured by the number of questions to be asked for a complete assessment also has psychological implications. The more questions are to be asked, the stronger the learner will be exhausted and, therefore, will be less able to concentrate. As Dowling and Hockemeyer (2001) have already shown, the number of questions to be asked depends strongly on the number of knowledge states. This can also be seen in Table 3 where simulations with medium noise and strictness have been computed. The number of questions increases steadily with growing knowledge spaces.

Finally, the number of questions has also been regarded with respect to the strictness of the criterion for finishing the assessment process. As Table 4 shows, the number of questions needed for a complete assessment increases with the strictness as it was expected.

Discussion and Further Research

Regarding the accuracy of the assessment, the results presented in the previous section show a quite robust reaction of both procedures to noise in the simulated learners' response behaviour. This holds especially for the continuous procedure with higher thresholds.

Table 3

Average number of questions to be asked for knowledge spaces of different size

Number of knowledge states	Number of questions (Standard deviation)	
	Discrete procedure	Continuous procedure
2261	14.7 (1.3625)	11.6 (1.0467)
14569	17.2 (1.2225)	14.5 (1.1132)
41395	18.8 (1.1718)	16.2 (1.2297)

Remark: Simulations were run with a maximal noise rate of 5% and with a medium strictness criterion.

Table 4

Average number of questions asked for different strictness values

Strictness	Number of questions (Standard deviation)	
	Discrete procedure	Continuous procedure
Low	12.5 (0.8814)	11.4 (0.8363)
Medium	14.7 (1.3625)	11.6 (1.0467)
High	17.1 (2.1924)	14.9 (2.0127)

Remark: Simulations were run for a knowledge space with 2261 knowledge states and with a maximal noise rate of 5%.

Looking at the efficiency of the assessment, i.e. the number of questions asked, we can still save about 50% of the items in knowledge domains with such a rather weakly structured area (cf. Dowling, 1991) as the usage of AutoCAD. For the discrete procedure, the increase of the number of questions had to be expected to be at least as much as the increase of strictness. For the continuous procedure, there is a lower increase than for the discrete one, especially for a medium level of strictness.

Summarising these results it can be said that the continuous procedure appears to be slightly superior to the discrete one in all investigated dimensions and should, therefore, also be in the focus of further research. In particular, there are two open issues. First, the practical comparison should be made in more detail including a deeper evaluation of the results as well as the variety of continuous procedures introduced by Falmagne and Doignon (1988a). A second open issue is the investigation and development of continuous non-deterministic assessment procedures using surmise mappings instead of knowledge spaces as representation for the structure of the knowledge domain.

References

- Albert, D. & Lukas, J. (ed.). (1999). *Knowledge Spaces: Theories, Empirical Research, Applications*. Mahwah, NJ : Lawrence Erlbaum.
- Degreef, E., Doignon, J.-P., Ducamp, A., & Falmagne, J.-Cl. (1986). Languages for the assessment of knowledge. *Journal of Mathematical Psychology*, *30*, 243-256.
- Doignon, J.-P. (1994). Probabilistic assessment of knowledge. In Dietrich Albert, (ed.), *Knowledge Structures* (pp. 1-56). New York: Springer.
- Doignon, J.-P. & Falmagne, J.-Cl. (1985). Spaces for the assessment of knowledge. *International Journal of Man-Machine Studies*, *23*, 175-196.
- Doignon, J.-P. & Falmagne, J.-Cl. (1999). *Knowledge Spaces*. Berlin: Springer.
- Dowling, C.E. (1991). *Constructing Knowledge Structures from the Judgements of Experts*. Braunschweig, Germany: Technische Universität Carolo-Wilhelmina.
- Dowling, C.E. (1993). Applying the basis of a knowledge space for controlling the questioning of an expert. *Journal of Mathematical Psychology*, *37*, 21-48.
- Dowling, C.E. & Hockemeyer, C. (2001). Automata for the assessment of knowledge. *IEEE Transactions on Knowledge and Data Engineering*, *13*, 451-461.
- Dowling, C.E., Hockemeyer, C., & Ludwig, A.H. (1996). Adaptive assessment and training using the neighbourhood of knowledge states. In Claude Frasson, Gilles Gauthier, & Alan Lesgold (ed.), *Intelligent Tutoring Systems*, volume 1086 of *Lecture Notes in Computer Science* (pp. 578-586). Berlin: Springer.
- Falmagne, J.-Cl. & Doignon, P. (1988a). A class of stochastic procedures for the assessment of knowledge. *British Journal of Mathematical and Statistical Psychology*, *41*, 1-23.
- Falmagne, J.-Cl. & Doignon, P. (1988b). A Markovian procedure for assessing the state of a system. *Journal of Mathematical Psychology*, *32*, 232-258.
- Hockemeyer, C. (2002). A Comparison of Non-Deterministic Procedures for the Adaptive Assessment of Knowledge. *Psychologische Beiträge*, *44*, 495-503.

Hockemeyer, C., Held, T., & Albert, D. (1998). RATH - a relational adaptive tutoring hypertext WWW--environment based on knowledge space theory. In Alvegård, C. (ed.), *CALISCE`98: Proceedings of the Fourth International Conference on Computer Aided Learning in Science and Engineering* (pp. 417-423). Göteborg, Sweden: Chalmers University of Technology.

Hockemeyer, C. (2002). A Comparison of Non-Deterministic Procedures for the Adaptive Assessment of Knowledge. *Psychologische Beiträge*, 44, 495-503.