Instructional Design in the Real World: A View from the Trenches

Anne-Marie Armstrong
US Government Printing Office, USA

Information Science Publishing
Hershey • London • Melbourne • Singapore

Chapter VIII

KABISA: Evaluation of an Open Learning Environment

Geraldine Clarebout
University of Leuven, Belgium

Jan Elen
University of Leuven, Belgium

Joost Lowyck
University of Leuven, Belgium

Jef Van den Ende
Institute of Tropical Medicine, Belgium

Erwin Van den Enden
Institute of Tropical Medicine, Belgium

ABSTRACT

This chapter focuses on the last phase of the systematic instructional design approach, ADDIE. This evaluation phase is illustrated through means of a case study, namely the evaluation of a computer-based training program, KABISA. The leading evaluation questions were whether students followed a criterion path and whether students used the embedded

help functions. Ninety-seven physicians following post-graduate training in tropical medicine participated in this evaluation. Log files were kept of the students and 21 students participated in thinking-aloud sessions. Results indicate that students do not follow the criterion path and that only poor use is made of help functions. This evaluation study shows that a systematic approach to instructional design remains highly valuable.

INTRODUCTION

Educational goals have generally shifted from knowing everything in a specific domain, to knowing how to deal with complex problems. Reasoning and information processing skills have become more important than the sheer amount of information memorized. In medical education the same evolution occurred. Diagnostic reasoning processes get more strongly emphasized. Whereas previously knowing all symptoms and diseases was stressed, reasoning skills have now become educationally more important. They must enable professionals to distinguish between differential diagnoses and to recognize patterns of illnesses (e.g., Myers & Dorsey, 1994).

Authentic or realistic tasks have been advocated to foster the acquisition of complex problem-solving processes (Jacobson & Spiro, 1995; Jonassen, 1997). In medical education this has led to the use in education of expert systems. Such systems were initially developed to assist practitioners in their practice (e.g., NEOMYCIN in Cormie, 1988; PATHMASTER in Frohlich, Miller, & Morrow, 1990; LIED in Console, Molino, Ripa di Meanan, & Torasso, 1992). These systems simulate a real situation and were expected to provoke or develop students' diagnostic reasoning processes. However, the implementation of such expert systems in regular educational settings has not been successful. Instead of developing reasoning processes, these systems assume them to be available. They focus on quickly getting to a solution rather than on reflecting on possible alternatives. Consequently, it was concluded that students need more guidance in the development of diagnostic reasoning skills (Console et al., 1992; Cromie, 1988; Friedman, France, & Drossman, 1991); instructional support was lacking.

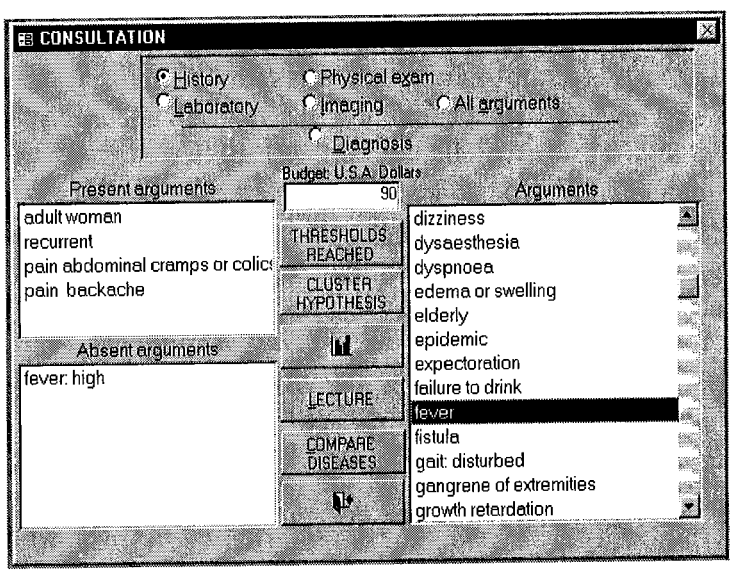
KABISA is one of the programs that was purposely designed to help students in the development of their diagnostic reasoning skills (Van den Ende, Blot, Kesten, Gompel, & Van den Enden, 1997). It is a dedicated computer-based training program for acquiring diagnostic reasoning skills in tropical medicine.

The evaluation of students' performance while using KABISA involved comparing students' paths to a pre-specified 'criterion path' for working with KABISA and analyzing the use of embedded help functions. The evaluation concludes with a discussion of the results, a reflection on the evaluation process itself, and the implications for the evaluation phase within the ADDIE instructional design-model.

DESCRIPTION OF THE KABISA PROGRAM

KABISA confronts the user with cases or 'virtual patients.' The virtual patient is initially presented by three 'arguments'[1] randomly selected by the computer. After the presentation of the patient (three arguments), students can ask additional arguments gathered through anamnesis, physical examination, laboratory, and imaging (see Figure 1). If students click on a particular argument, such as physical examination or test, they receive feedback. Students are informed about the presence of a certain disease characteristic, or whether a test is positive or negative. If students ask a 'non-considered' argument, i.e., an argument that is not relevant or useful in relation to the virtual patient, they are informed about this and asked whether they want to reveal the diagnosis they were thinking about. After selecting a diagnosis, students receive an overview of the arguments that are explained by their selection and which

Figure 1: User Interface of KABISA



ones are not, as well as the place of the selected diagnosis on a list that ranks diagnoses according to their probability given the arguments at hand. If students do not want to show the diagnosis they were thinking about, they can just continue asking additional arguments. A session is ended with students giving a final diagnosis. KABISA informs them about its correctness. If their diagnosis is correct, students are congratulated.

If the diagnosis is not correct, students may be either informed that it is a very plausible diagnosis but that the threshold is not yet reached, or they may get a ranking of the diagnosis and an overview of the disease characteristics that can and cannot be explained by their answer (this feedback is similar to the one students receive when they show the diagnosis they were thinking about after asking a non-considered argument).

In addition to the information on non-considered arguments, students may use other support systems available in KABISA. KABISA offers the following support systems:

1. *Threshold reached:* this tool gives a ranking of the possible diseases at that moment and provides information on whether a threshold is reached. Reaching a threshold means that, given the disease characteristics present and absent at a certain moment, a diagnosis can be given, although one is not absolutely certain, but sufficiently certain to start a treatment.
2. *Cluster hypothesis:* presents an overview of all possible diagnoses, given the disease characteristics known at a certain moment.
3. *Graphic:* presents the different diagnoses with the disease characteristics known by the student. The graphic indicates the contribution for each disease characteristic and how it contributes to thresholds for different possible diseases.
4. *Lecture:* in this section, students can ask for information about a certain disease. They get all disease characteristics that occur if a patient has a particular disease, or by clicking on the characteristics they get information on their diagnostic power.
5. *Compare diseases:* gives the opportunity to compare disease characteristics of two diagnoses. The comparison reveals the unique and shared characteristics of the alternatives considered.

There are two different versions of KABISA, a junior and senior version. These versions do not differ in structure or content, but with respect to difficulty level. In the junior version virtual patients always have all the typical arguments for a disease. If a student asks an argument that should be present given a 'textbook' description of the diagnosis, the program will confirm its presence. In the

senior consultation, some of the arguments that should typically be present for a specific diagnosis might be missing.

Parallel to these two versions, there are also exam-versions of KABISA. These versions are similar to the junior and senior version. However, students can no longer use the support systems.

EVALUATION METHODOLOGY

In order to gain insight in how students use KABISA, an evaluation was performed. To do this, first a criterion path was drawn that represents the 'ideal' way of working with KABISA. This criterion path served as the benchmark for the evaluation. This criterion path was elaborated in close collaboration between the evaluators and the domain experts.

Evaluation Questions

Two evaluation questions were focused upon, namely:

1. Do students follow the criterion path when working on KABISA? And if not, how many, and how serious do students deviate from this path?
2. Do students use the different embedded help functions?

To answer these questions, relationships between program and student characteristics were explored.

Participants

The students involved in this evaluation are general physicians following a post-graduate training in tropical medicine at the Institute of Tropical Medicine in Antwerp. Thirty-seven Dutch-speaking students and 60 French-speaking students participated. For the complete group log files were kept, and 21 volunteers (10 from the Dutch-speaking group and 11 from the French-speaking group) participated in think-aloud sessions.

Evaluation Instruments

For the first question, two evaluation approaches were used: the analysis of think-aloud protocols and the analysis of log files. For the second question, only log files were analyzed. The think-aloud procedure involved students performing two consultations on KABISA (one with the junior version and one with the senior version), while externalizing their reasoning processes. Students were instructed to think aloud and work with KABISA the way they would

normally do, i.e., without an evaluator sitting next to them. It was explicitly mentioned that they could use all functionalities of the program. If students did not say anything while working on KABISA, they were prompted by the evaluator to think aloud with a question like: "What are you thinking of?" Everything students said was audiotaped. Notes were taken of the different steps students took. The think-aloud method allows detailed insight in the reasoning process of students, and the path they follow during a session (Karat, 1997; Ericsson & Simon, 1993).[2]

Audiotapes and notes were transcribed and used for protocol analysis.

For analyzing the log files, a dedicated program was developed. The program registered different actions of students while working on KABISA. For data gathering through log files, students were asked to do three junior and three senior consultations, and three exams (junior version) on KABISA. The advantage of log files is their unobtrusive character. The registration of students' actions has no effect on the behavior of students (Jackson, 1990).

Procedure

Question 1: Do students follow the criterion path? And if not, how many, and how serious deviations do students make?

As previously mentioned, a criterion path was constructed. This path was compared to students' paths when working with KABISA. Actual paths followed by the students were reconstructed based on the think-aloud protocols.

For the second part of this evaluation question, all possible deviations from the criterion path were identified and scored on their 'seriousness' by two experts of the Tropical Institute. Both experts gave a score for the possible errors on a six-point scale (from 0, not serious, to 5, very serious mistake). The sum of these scores was used in further analyses.

This approach corresponds to what Elstein and Rabinowitz (1993) call a 'normative approach.' Uncertainties and risks involved in a clinical situation are translated in probabilities that allow the construction of an ideal model.

A comparison was made between sessions with the senior and the junior version as well as between consultations by the French-speaking and the Dutch-speaking groups.

Using the think-aloud protocols, students' deviations from the optimal path were identified, summed, and given a score. An average number of deviations and an average score for the seriousness of the deviations were calculated for every student think-aloud. Different groups and different versions were compared.

Question 2: Do students use the embedded help functions?

For this question, the frequency students' use of help function was analyzed. A mean was calculated per help function for each session. These means were used as dependent variables. Three-way ANOVAs or MANOVAs were performed (depending on the correlation between the different help functions) with 'group,' 'version,' and 'finding the correct diagnosis' as independent variables.

RESULTS

Question 1a: Following the Optimal Path

Of 21 students participating in the think-aloud procedure, only one French student followed the criterion path during a session with the senior version. All other students did not follow the criterion path. Log file analysis reveals only eight consultations in which the criterion path was followed. Five out of 44 students followed this path (see Table 1).

Question 1b: Number of Deviations

For the number of mistakes, the analysis of the think-aloud protocols reveals that with the junior version, on average four mistakes are made per session. Almost five mistakes per senior version are made. The French group makes fewer deviations from the criterion path than the Dutch group. Concerning the seriousness of the mistakes, it seems that the Dutch group makes more serious deviations than the French group, both for sessions with the junior and the senior versions (see Table 2). It should be noted, however, that the data presented here relate only to a limited number of students (N = 21) and

Table 1: Number of Students Following the Optimal Path

Group (max. N sessions)	**Version**	**N students**	**N sessions**
French (n = 218)	Junior	1	2
	Senior	1	1
Dutch (n = 142)	Junior	2	4
	Senior	1	1

Table 2: Means for the Number of Mistakes and Seriousness of Deviations

		Number of deviations ($\bar{x}$)		Seriousness of deviations ($\bar{x}$)	
Junior	Dutch	4.70 (SD = 3.16)	3.95 (SD = 2.57)	14.40 (SD = 10.37)	11.38 (SD = 8.77)
	French	3.20 (SD = 1.93)		8.64 (SD = 6.30)	
Senior	Dutch	7.00 (SD = 5.70)	4.90 (SD = 4.72)	20.00 (SD = 17.40)	13.43 (SD = 14.09)
	French	3.10 (SD = 2.69)		7.45 (SD = 6.50)	
Total		*4.40 (SD = 3.65)*		*12.41 (SD = 11.40)*	

consultations (N = 42). Hence, no statistics were performed to test the significance of the observed differences.

However, the log file analysis reveals that even if students deviate from the criterion path, the correct diagnosis is found in more than half of the sessions (Table 3). This result will be discussed later as an implication for other evaluation studies.

Question 2: Use of Help Functions

For this question a difference is made between the response to discuss a non-considered argument and the remaining help functions. For most help functions students themselves have to take the initiative to use it. In case of a non-considered argument, however, KABISA explicitly invites students to provide a response.

Non-Considered Arguments

To gain insight in the way students deal with non-considered, or irrelevant arguments, the average number of considered and non-considered arguments was calculated. Within the last group, the average of non-considered discussed arguments was calculated as well. From Table 4, it can be seen that students ask more for considered arguments than for non-considered ones. On average, three non-considered arguments are asked for almost 13 considered ones.

Table 3: Cross Table for Finding the Diagnosis

		Version		
		Junior	**Senior**	**Total**
Diagnosis found	No	23	47	70
	Yes	155	134	289
Total		*178*	*181*	*359*

Table 4: Number of Considered, Non-Considered, and Discussed Non-Considered Arguments

	Version	N sessions	$\bar{x}$	SD
Number of considered arguments	Junior	178	13.11	68.32
	Senior	181	12.10	48.32
Number of non-considered arguments	Junior	178	3.28	17.32
	Senior	181	2.59	12.34
Number of discussed non-considered arguments	Junior	178	.90	6.06
	Senior	181	.85	4.47
Total amount of arguments	*Junior*	*178*	*16.39*	*85.45*
	Senior	*181*	*14.69*	*60.41*

However, less than one out of three non-considered arguments are discussed. In other words, students who ask a non-considered argument ignore the opportunity to discuss the diagnosis they were thinking about.

The large standard deviation (SD) reveals large differences between the sessions with respect to the number of arguments asked.

To study the influence of group, version, and finding the diagnosis, proportions were calculated for the number of considered and non-considered arguments, by taking the sum of the considered and non-considered arguments as 100%. For the number of discussed non-considered arguments, the proportion was calculated by taking the number of non-considered arguments as 100%.

To decide whether a MANOVA or ANOVA should be done, a correlation (Pearson) was calculated between the proportion of considered arguments and the proportion of non-considered discussed arguments. This resulted in a low but significant positive correlation ($r = .16$, $p = .02$). The larger the

Figure 2: Main Effect of Version on the Proportion of Considered, Non-Considered, and Discussed Arguments

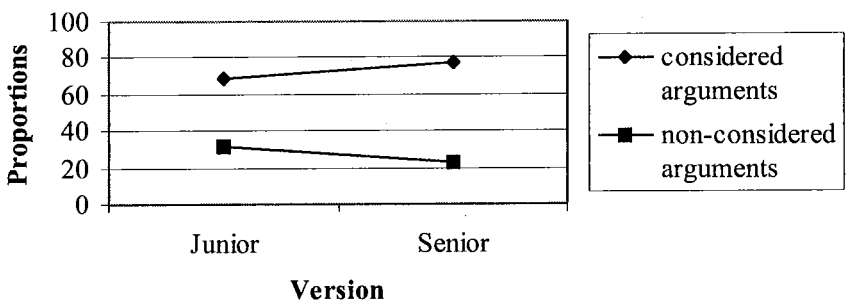
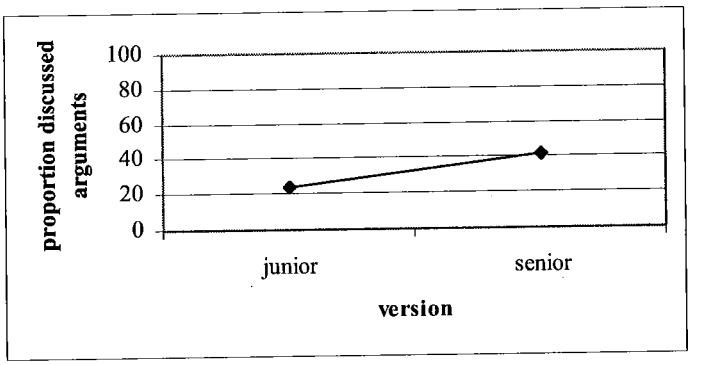


Figure 3: Proportion of Non-Considered Discussed Arguments



proportion of considered arguments asked during a session, the greater the probability that students will discuss non-considered arguments.

Given this correlation a multivariate three-way analysis of variance (MANOVA) was performed with group, version, and finding the diagnosis as independent variables and the proportion of considered and non-considered discussed arguments as dependent variables. A significant main effect of version was found ($\lambda = .95$, $F(2,221) = 5.43$, $p = .01$) (see Figure 2[3]). In the senior version the proportion of considered arguments is significant higher than in the junior version.

Looking only at non-considered arguments, results also show that significantly more non-considered arguments are discussed when working with the senior version than with the junior version (Figure 3). The more difficult the consultations are, the higher the proportion of considered arguments (Figure 2) and the more students discuss non-considered arguments (Figure 3). No significant effects were found for group and finding the diagnosis.

Other Help Functions

In Table 5, an overview is presented of the frequency of consulting a particular help function; the number of sessions, and the number of sessions in which a help function was consulted; and a correct diagnosis was found. From this table it can be derived that 'clusters' and 'thresholds' are consulted most frequently. Results also indicate that these help functions are consulted repeatedly in a session. The number of consulting thresholds is almost twice the number of sessions.

Table 5: Consultation of the Different Help Functions (HFs)

Help function	N consulted	N sessions in which a HF was consulted/ total N sessions	N correct diagnoses/ N sessions in which a HF was consulted	N students consulting a HF
Clusters	150	108/350 (= 30.86%)	80/108 (= 74.07%)	34/51 (= 66.67%)
Thresholds	297	174/350 (= 49.71%)	137/174 (= 78.44%)	40/51 (= 78.43%)
Graphic	45	33/350 (= 9.43%)	24/33 (= 72.73%)	18/51 (= 35.29%)
Lecture	69	47/350 (= 13.43%)	32/47 (= 68.09%)	17/51 (= 33.33%)
Compare diseases	29	25/350 (= 7.14%)	12/25 (= 48%)	25/51 (= 49.02%)

In order to detect the influence of group and version on the consultation of these help functions, first correlations (Pearson) were calculated for consulting the help functions (Table 6). A positive correlation indicates a tendency to use several help functions during one session.

Table 6: Correlation Between the Use of the Different Help Functions

	Thresholds	Graphic	Lecture	Compare diseases
Clusters	.35**	.22**	.31**	.35**
Thresholds		.18**	.38**	.26**
Graphic			.28**	.35**
Lecture				.26**

A MANOVA with group and version as independent variables and consultation of help functions as dependent variables resulted in two main effects, one for group ($1 = .91, F(5,350) = 6.81, p = .01$) and one for version ($1 = .96, F(5,350) = 3.20, p = .01$). For all help functions, (except for clusters where there was no significant difference between the versions), students more frequently consult the help functions in the senior consultation than in the junior version. With respect to differences between the two groups, the French-speaking group more frequently consults the help functions than the Dutch-speaking group does. Graphic is an exception; the Dutch group (see Table 7 for an overview) more frequently consults it.

Table 7: Two Main Effects for Group and Version on the Use of the Different Help Functions

	Version			
Help function	$\bar{x}$ *junior (SD)*	$\bar{x}$ *senior (SD)*	*(F1,357)*	*p*
Clusters	.32 (.61)	.52 (.82)	.96	.33
Thresholds	.63 (.87)	1.02 (1.68)	13.26	.00
Graphic	.01 (.26)	.19 (.64)	4.04	.05
Compare diseases	.00 (.15)	.14 (.41)	4.26	.04
Lecture	.01 (.35)	.29 (.77)	11.46	.00
	Group			
Help function	$\bar{x}$ *Dutch (SD)*	$\bar{x}$ *French (SD)*	*F(1,357)*	*p*
Clusters	.37 (.71)	.45 (.85)	4.08	.04
Thresholds	.51 (.82)	1.04 (1.57)	3.92	.05
Graphic	.18 (.67)	.01 (.33)	7.60	.01
Compare diseases	.00 (.22)	.17 (.35)	10.20	.00
Lecture	.01 (.26)	.28 (.73)	6.25	.01

CONCLUSION AND SUGGESTIONS FOR OPTIMIZATION

The evaluation of KABISA reveals that students do not follow a criterion path when working with KABISA. As evidenced by log file analysis, a criterion path was followed only in eight sessions. The think-aloud procedure reveals only one such session. These findings might be explained by the perception of students of consulting help functions. In the think-aloud protocol analysis, indications were found that students conceive consulting a help function as cheating or as failing:

"I'm going to cheat now and have a look at the thresholds."
"I really don't know what it is (...) I'm going to look at the thresholds."
"I give up, can I look at the thresholds?"

The students anticipate the feedback provided by KABISA when a non-considered argument is asked for. They rephrase it to 'stupid argument':

"If I would now ask fever, he will tell me that this is a stupid question."
"Stool, but he will say that it is a stupid question."
"I will ask something but the computer will probably not find it very interesting."

Log file analyses reveal also that students seldom consult help functions. Given the limited use of help functions, their impact on the learning process cannot be but limited.

Concerning version and group, differential effects were found for the use of the help functions. The Dutch-speaking group less frequently discusses non-considered arguments. In general, students in the French group more frequently consult help functions. For version, the proportion of considered arguments is larger in the senior version than in the junior version. Similarly, non-considered arguments are more often discussed in the senior version than in the junior version. Help functions are more consulted in a senior consultation than in a junior consultation.

However, in spite of the limited use of the help functions and in spite of the observation that in only a small number of consultations the optimal path was followed, students do find the diagnosis in 80% of the consultations (Table 3).

It might be concluded that KABISA provides easy tasks for students. Or, the program may allow too easily for guessing and may not sufficiently link the criterion path to finding the correct diagnosis. Students can easily follow another path and still make the correct diagnosis. Overall, students approach the program as being product directed rather than learning directed. Finding the correct diagnosis seems to be more important than the reasoning process to arrive at a diagnosis. Differences between the French-speaking and the Dutch-speaking group further suggest that the way in which KABISA is introduced to the students influences their use of KABISA.

The results of this evaluation suggest that KABISA is currently not used by students to foster their diagnostic reasoning skills. Rather, it enables them to train readily available skills.

IMPLICATIONS

Looking at the ADDIE-model, it can be said that the evaluation phase remains important. Through the use of evaluation methods, feedback is given with respect to other phases of the design process. In the case of KABISA for example, the evaluation gave some indications that more attention should have been given to the analysis phase. A more thorough analysis of student characteristics could have provided a means to adapt the difficulty level to the level of the students or to identify what guidance students actually need. Apparently, the feedback given to students does not encourage them to adapt their problem-solving process. Being product rather than process oriented, feedback may not be adapted to students' actual needs. Or, students' perception of the program (a game versus an educational application) may

influence the use of the program. These perceptions should be taken into account throughout the design process of the program.

The difference between the French-speaking group and the Dutch-speaking group indicates a need for a different type of introduction for the program. In the introduction, the aims of the program, the different functionalities and the relationship with the different courses are clearly defined (see Kennedy, Petrovi, & Keppell, 1998 for the importance of introductory lessons). This relates to the implementation phase.

In order to perform a thorough evaluation, the use of different evaluation instruments provides more information than using only one instrument. With respect to the evaluation of KABISA, the think-aloud method resulted in both quantitative and qualitative results and, hence, a more detailed insight in the reasoning process of student. This think-aloud method allowed, for example, to find out why students make only limited use of these help functions. However, given the time investment that is needed for collecting the information and analyzing the protocols, the method may not always be applicable. Log files on the other hand are automatically generated and allow one to easily gather data from a large group. However, they do not provide insight in the reasoning processes of students (Drury, 1990; Kirwan & Ainsworth, 1992). This underscores the need for multiple evaluation tools and methods to obtain optimum results.

Open and realistic learning environments make it difficult to anticipate and to take into account during the design phase potential problems or difficulties students might encounter. A recommendation in this respect would be to break the linearity of the ADDIE-model and to introduce a formative evaluation after each phase. This would enable the redirection of the program while developing it, rather than after the implementation of the program. Rather than only evaluating a final product, the development process should be taken into consideration as well. Rapid prototyping for testing the program at different phases of the development might be indicated.

The presented case study of KABISA illustrates the importance of the evaluation in the ADDIE process. It revealed students to be able to state a correct diagnosis without using the diagnostic skills the program purports to be training.

For various reasons (limited time, limited budget, etc.), this phase often receives limited attention or is quickly dealt with through a questionnaire measuring students' attitudes towards the program. Restricted evaluations on the other hand may be both cheap and non-productive. Kirkpatrick (1994) has

already revealed that such restricted evaluations have only limited value. However, a more thorough evaluation can point out weaknesses and flows otherwise remaining undiscovered, thus making an investment in evaluation (formative and summative) worthwhile.

ENDNOTES

1 The term 'argument' refers to either a symptom or disease characteristic, as well as a request for results of a physical examination, laboratory test, or imaging.

2 There is an ongoing discussion whether thinking aloud interferes in the reasoning process of students and whether the results are reliable. Criticisms have been made and have been rejected (e.g., Karat, 1997; Veenman, Elshout, & Groen, 1993). But given the structured environment of KABISA, it is assumed that thinking aloud does not interfere to the extent that students change their behavior.

3 In Figure 2, the proportion of non-considered arguments is also presented, although this was not entered as a variable in the analysis since it is the inversed proportion of the considered arguments.

REFERENCES

Console, L., Molino, G., Ripa di Meana, V., & Torasso, P. (1992). LIED-liver: Information, education and diagnosis. *Methods of Information in Medicine,* 31, 284-297.

Cromie, W.J. (1988). Expert systems and medical education. *Educational Researcher,* 17(3), 10-12.

Drury, C.G. (1990). Computerized data collection in ergonomics. In Wilson, J.R. & Carlett, N.I. (Eds.), *Evaluation of Human Work* (pp. 229-243). London: Taylor and Francis.

Elstein, A.S. & Rabinowitz, M. (1993). Medical cognition: Research and evaluation. In Rabinowitz, M. (Ed.), *Cognitive Science Foundation of Instruction* (pp. 189-201). Hillsdale, NJ: Lawrence Erlbaum Associates.

Ericsson, K.A. & Simon, H.A. (1993). *Protocol Analysis: Verbal Reports as Data* (revised ed.). Cambridge, MA: MIT Press.

Friedman, C.P., France, C.L., & Drossman, D.D. (1991). A randomized

comparison of alternative formats for clinical simulations. *Medical Decision Making,* 11(4), 265-271.

Frohlich, M.W., Miller, P.L., & Morrow, J.S. (1990). PATHMASTER: Modeling differential diagnosis as "Dynamic Competition" between systematic analysis and disease-directed deduction. *Computers and Biomedical Research,* 23, 499-513.

Hannafin, M.J., Hall, C., Land, S., & Hill, J. (1994). Learning in open-ended learning environments: Assumptions, methods, and implications. *Educational Technology,* 34(10), 48-55.

Jackson, G.A. (1990). Evaluating learning technology. *Journal of Higher Education,* 61(3), 294-311.

Jacobson, M.J. & Spiro, R.J. (1995). Hypertext learning environments, cognitive flexibility and the transfer of complex knowledge. *Journal of Educational Computing Research,* 12(4), 301-333.

Jonassen, D.H. (1997). Instructional design models for well-structured and ill-structured problem-solving learning outcomes. *Educational Technology Research and Development,* 45(1), 65-91.

Karat, J. (1997). User-centered software evaluation methodologies. In Helander, M., Landauer, T.K., & Brabhu, P. (Eds.), *Handbook of Human-Computer Interaction (2nd ed.)* (pp. 689-704). Amsterdam: Elsevier Science.

Kennedy, G., Petrovic, T., & Keppell, M. (1998). The development of multimedia evaluation criteria and a program of evaluation for computer aided learning. In Corderoy, R.M. (Ed.), *Proceedings of the Fifteenth Annual Conference of the Australian Society for Computers in Tertiary Education (ASCILITE)* (pp. 407-415). Wollongong, Australia: University of Wollongong.

Kirkpatrick, D.L. (1994). *Evaluating Training Programs. The Four Levels.* San Francisco, CA: Berrett-Koehler Publishers.

Kirwan, B. & Ainsworth, L.K. (1992). Observational techniques. In Kirwan, B. & Ainsworth, L.K. (Eds.), *A Guide to Task Analysis* (pp. 53-58). London: Taylor & Francis.

Myers, J.H. & Dorsey, J.K. (1994). Using diagnostic reasoning (DxR) to teach and evaluate clinical reasoning skills. *Academic Medicine,* 69, 429.

Shaw, E., Johnson, W.L., & Ganeshan, R. (1999). *Pedagogical Agents on the Web.* Available online at: http://www.isi.edu/isd/ADE/papers/agents99/agents99.htm.

Van den Ende, J., Blot, K., Kestens, L., Van Gompel, A., & Van den Ende, E. (1997). KABISA: An interactive computer-assisted training program for tropical diseases. *Medical Education,* 31, 202-209.

Veenman, M.V., Elshout, J.J., & Groen, M.G. (1993). Thinking aloud: Does it affect regulatory processes in learning? *Tijdschrift voor Onderwijsresearch,* 18(6), 322-330.