

Evaluation of ERST – An External Representation Selection Tutor

Beate Grawemeyer

Representation & Cognition Group

Department of Informatics, University of Sussex, Falmer, Brighton BN1 9QH, UK
b.grawemeyer@sussex.ac.uk

Abstract. This paper describes the evaluation of ERST, an adaptive system which is designed to improve its users' external representation (ER) selection accuracy on a range of database query tasks. The design of the system was informed by the results of experimental studies. Those studies examined the interactions between the participants' background knowledge-of-external representations, their preferences for selecting particular information display forms, and their performance across a range of tasks involving database queries. The paper describes how ERST's adaptation is based on predicting users' ER-to-task matching skills and performance at reasoning with ERs, via a Bayesian user model. The model drives ERST's adaptive interventions in two ways - by 1. hinting to the user that particular representations be used, and/or 2. by removing from the user the opportunity to select display forms which have been associated with prior poor performance for that user. The results show that ERST does improve an individual's ER reasoning performance. The system is able to successfully predict users' ER-to-task matching skills and their ER reasoning performance via its Bayesian user model.

1 Introduction

People vary in their knowledge of external representations (KER), in the range of representations that they can use effectively on reasoning tasks, and in their ability to match particular representations to tasks (i.e. in their knowledge of 'applicability conditions' ([19])).

Numerous factors are associated with ER-to-task matching skill. First, it is known that individuals differ widely in terms of their preferences for particular forms of external representation (ER) ([4],[6],[16]). Better reasoners organise their knowledge of ERs on a 'deeper' semantic basis than poorer reasoners, and are better at correctly naming various ER forms ([5],[7]).

Secondly, some types of tasks require a particular, specialised type of representation to solve whereas for other types of tasks, several different ER forms may be useful. The extent to which a problem is representationally-specific is determined by characteristics such as its degree of determinacy (extent to which it is possible to build a single, unique model of the information in the problem). ER selection skill requires, *inter alia*, knowledge of a range of ERs in terms of a) their semantic properties (e.g. *expressiveness*), b) their functional roles ([4],[23],[5],[2],[19])

together with information about the ‘applicability conditions’ under which a representation is suitable for use on a particular task ([19]).

This paper describes the evaluation of ERST - a prototype External Representation Selection Tutor. ERST is an adaptive system designed to help users to choose effective representations for use across a varied range of tasks that involve answering queries using graphically displayed information from a database.

ERST’s user model is being developed ([11],[12]) on the basis of empirical data gathered from a series of empirical studies. In these experiments a prototype automatic information visualization engine (AIVE) was used to present a series of questions about the information in a database. This approach is similar to that of [14], who used an empirical basis for the development of a user model for the READY system. That system models users’ performance capacity under various cognitive load conditions.

This paper focusses upon an evaluation of ERST in which two versions of the system were compared. One version (used by participants in the evaluation group) used the adaptive ERST system with its user-modelling system turned on. A comparison group used another version of ERST - one with the user modeling subsystem turned off.

2 ERST

The aim of ERST is to enhance users’ ER reasoning performance across a range of different types of database query tasks. The adaptive system is able to predict ER to task matching skills and ER reasoning performance, based on its user model. It drives ERST’s adaptive interventions (by hinting or advising) or by ‘hiding’ inappropriate display forms.

ERST’s user model and its user-adaptation mechanism have been developed on the basis of empirical data gathered from two experiments. The study [10] investigated the representation selection and reasoning behaviour of participants who were offered a choice of information-equivalent data representations (for example, tables, bar charts, etc.) for use on various database query tasks. Some tasks required the identification of unique entities, some required the detection of clusters of similar entities, and some involved the qualitative comparison of values.

A further study [11], investigated the degree to which some task types are more representation-specific¹ than others, with respect to reasoning performance and response latency.

The results showed that, display selection accuracy and database query answer performance were both significantly predicted by prior knowledge-of-external-representations (KER) pre-tests. Specifically, conceptual (classificatory) knowledge of ERs predicts success at appropriate information display selection on the AIVE tasks. In contrast, deeper, semantic (functional) knowledge of ERs was associated with success at *using* the selected ER i.e. reading-off information and using it to respond correctly to the database query.

¹ These are tasks for which only a few, specialised, representational forms are useful.

It was also found (unsurprisingly) that appropriate representation selection results in better query answering performance. Taken together the results suggested that for predicting query response accuracy, a participant's KER can be as powerful a predictor of question answering accuracy as display selection accuracy.

The selection latency results show that a speedy selection of a display type is associated with a good display-type choice. This could be interpreted to imply that users tend to do one of two things - either they recognise the 'right' representation and proceed with the task or they procrastinate and hesitate because of uncertainty about which display form to choose. Hence less time spent responding to the database query question is associated with a good display-type choice and correct query response. This suggested that the selection and database query latencies may be used in the system's user model as predictors of users' ER expertise.

However these effects differed somewhat across types of task. On highly representationally-specific tasks (e.g. correlate variables) high prior KER significantly predicted the user's ability to identify and select the optimal information display and hence to perform well on the database query task. In contrast, on less ER-specific tasks (eg 'locate'), on which several different types of information display are potentially equally effective, prior KER predicted performance less strongly.

The results of the experiments indicated that ERST (the adaptive version of AIVE), needs to take into account a) individual differences (like user's ER preferences), b) their level of prior ER knowledge and c) the domain task characteristics, in order to coach ER-to-task matching skills in an individualised way reflecting the individual's needs.

2.1 User Model

As described above, ERST's user model is derived from two experimental studies ([10],[11]) which examined the relationship between participants' background knowledge of external representations (KER) and their ability to select appropriate information displays in the course of responding to various types of database queries.

The experimental results show that particular types of data are crucial for modeling. A Bayesian network approach (e.g. [21],[17]) was chosen as a basis for ERST's user model. Bayesian networks have been applied successfully in ITS (e.g. [3]) and are suitable, *inter alia*, for recognizing and responding to individual users, and they can adapt to temporal changes.

The structure of a simple Bayesian network based on the experimental data can be seen in figure 1. The Bayesian network in ERST's user model has been 'seeded' with the empirical data from the experiments in order that it could, from the outset, usefully monitor and predict users' ER selection preference patterns within and across query types.

The aim was for ERST to be able to relate query response accuracy and latencies to particular display selections and contrive query/display option com-

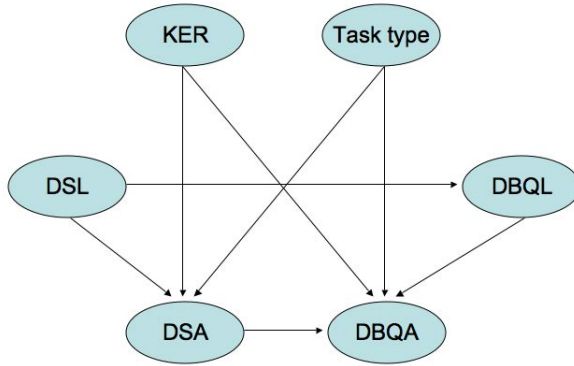


Fig. 1. Graph of a Bayesian network for ERST’s user model. KER = knowledge-of-external-representations; DSL = display selection latency; DBQL = database query answer latency; DSA = display selection accuracy; DBQA = database query answer performance.

binations to ‘probe’ an individual users’ prior knowledge of ERs. The empirical data was used to instantiate values in the relevant conditional probability tables (CPTs) at each node of the model. The bayesian network is sensitive to the representation specificity of different types of tasks. The network dynamically adjusts the CPT values and evolve individualised models for each of its participants in real time as they interact with the system. For example, for each ER selection and resulting database query performance score the corresponding CPT values will be updated and used from the system for an individual adaptation. The learned network is able to make the following inferences:

- **Predicting ER preferences and performance with uncertainty about background knowledge**

If there is uncertainty about users’ background knowledge of ERs, the system is able to make predictions about the dependent variables (e.g. background knowledge-of-external-representations (KER)), through a probability distribution of each these variables.

- **Learning about users’ ER preferences and performance**

Users’ ER preferences and performance can be learned incrementally, through users’ interaction with the system. The network can be updated with the individual characteristics and used to predict future actions and system decisions.

These inferences are used as a basis for ERST’s adaptive interventions based on background knowledge, task type and ER preferences.

2.2 The Adaptation Process

As mentioned above, ERST’s interventions consist of both overt hints or advice to users and also covert adaptations such as not offering less-appropriate display

forms² in order to prevent users from selecting them. The system is able to adapt to the individual user in the following ways:

- **Hiding ‘inappropriate’ display forms**

The system varies the range of ‘permitted’ displays as a function of each tasks’ ER-specificity and the users’ ER selection skill.

- **Recommending ERs**

The system will interrupt and highlight the most appropriate ER (based on its user model) if too much time is spent on selecting a representation, after learning an individuals’ selection display selection latency patterns.

Based on users’ interactions the system is able to adapt the range of displays and/or recommend ERs. For example, if a user manifests a particularly high error rate for particular task/ER combinations, then the system will limit the ER selection choice if it believes that this task could be answered with an appropriate different ER. Additionally, after the system detected users average display selection latencies, ERST will recommend the most appropriate ER to the user, if the system believes that the user is unclear what kind of ER to choose and spends too much time in selecting a representation for a particular database query.

3 Evaluation of ERST

In order to evaluate ERST two version of the system will be compared. One version with the adaptive system turned on and the other version with the user modeling subsystem turned off, similar to the evaluation approach employed by [1] or [22]. It was hypothesised that ERST would improve an individuals ER reasoning performance across a range of different types of database query tasks.

3.1 Participants

Thirty two participants, 20 in the comparison and 12 in the evaluation group, were recruited for this evaluation. The twenty participants in the comparison group³, which used non adaptive ERST, includes 5 software engineers, 1 graphic designer, 1 html programmer, 2 IT business managers, 7 postgraduate students, and 4 research officers/fellows (6 female/14 male). The evaluation group, which used the adaptive version of ERST consists of 2 software engineers, 1 IT business manager, 6 postgraduate students, and 3 research fellows (4 female/8 male).

3.2 Procedure

Participants in both groups were administered 4 pre-tasks, designed to assess their knowledge of external representations (KER) [8], before completing ERST’s database query problem solving task.

² ERST having observed the user attempt to use such ERs unsuccessfully over several previous trials.

³ The data from the second AIVE study is used in the control group.

3.3 Knowledge of External Representations (KER) Tasks

Four knowledge-of-external-representation tasks were employed. These consisted of a series of cognitive tasks designed to assess ER knowledge representation at the perceptual, semantic and output levels of the cognitive system [8]. A large corpus of external representations (ERs) was used as stimuli. The corpus contained a varied mix of 112 ER examples including many kinds of chart, graph, diagram, tables, notations, text examples, etc.

The first task was a decision task requiring decisions, for each ER in the corpus, about whether it was ‘real’ or ‘fake’⁴. This was followed by a categorisation task designed to assess semantic knowledge. Participants categorised each representation as ‘graph or chart’, ‘icon/logo’, or ‘map’, etc. In the third (functional knowledge) task, participants were asked ‘*What is this ER’s function?*’ An example of one of the (12) multiple-choice response options for these items is ‘*Shows patterns and/or relationships of data at a point in time*’. In the final task, participants chose, for each ER in the corpus, a specific name from a list. Examples include ‘venn diagram’, ‘timetable’, ‘scatterplot’, ‘Gantt chart’, ‘entity relation (ER) diagram’.

The 4 tasks were designed to assess ER knowledge representation using an approach informed by picture and object recognition and naming research [13]. The cognitive levels ranged from the perceptual level (real/fake decision task) to through production (ER naming) to deeper semantic knowledge (ER functional knowledge task).

3.4 ERST’s Database Query Tasks

Following the KER tasks, participants’ performed the ERST database query tasks. Participants were asked to make judgments and comparisons between cars and car features based on database information. The database contained information about 10 cars: manufacturer, model, purchase price, insurance group, CO2 emission, engine size, horsepower, etc.

Each subject responded to 30 database questions, which were of 6 types: identify; correlate ; quantifier-set; locate; cluster (similarity); compare negative. For example, a typical correlate task was: ‘Which of the following statements is true? A: Insurance group and engine size increase together. B: Insurance group increases and engine size decreases. C: Neither A nor B?’; or a typical locate task: ‘Where would you place a Fiat Panda with an engine size of 1200 cc inside the display?’.

Participants were informed that to help them answer the questions, the system (ERST) would supply the appropriate data from the database. Further details of the tasks can be found in [10].

ERST also offered participants a choice of representations of the data. They could choose between various types of ERs, e.g. set diagram, scatter plot, bar chart, sector graph, pie chart and table. Only in the comparison group were all representations offered to participants, as in Figure 2.

⁴ Some items in the corpus are invented or chimeric ERs.



Fig. 2. Representation selection interface

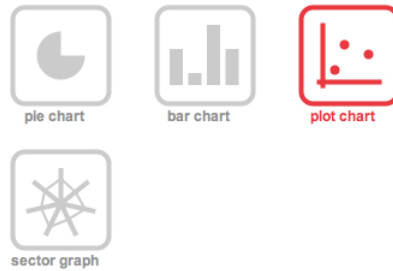


Fig. 3. Representation selection interface

For evaluation group participants (who used adaptive ERST), the list of representation choices might have been reduced or a single representation recommended (as discussed in 2.2) if the system believed that this intervention would enhance the individuals' ER reasoning performance on the basis of the current user's interaction and performance history. An example is provided in Figure 3.

Participants in both groups were told that they were free to choose any ER, but that they should select a form of display they thought was most likely to be helpful for answering the question. Participants then proceeded to the first question, read it and selected a representation.

The spatial layout of the representation selection buttons was randomized across the 30 query tasks in order to prevent participants from developing a set pattern of selection.

Based on the literature (e.g. [9]) a single 'optimal' ER for each task was identified (display selection accuracy scores were based on this - see results). However, each query type could *potentially* be answered with any of the representations offered by the system (except for set diagrams, which were only usable in quantifier-set tasks).

After the participant made his/her representation choice, ERST recorded the selection time (display selection latency) and then generated and displayed the representation instantiated with the data required for answering the question e.g. Figure 4).

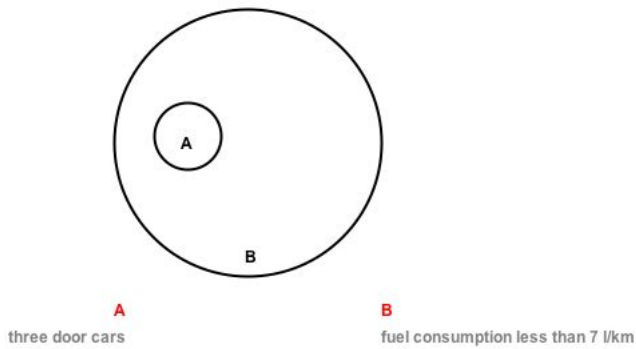
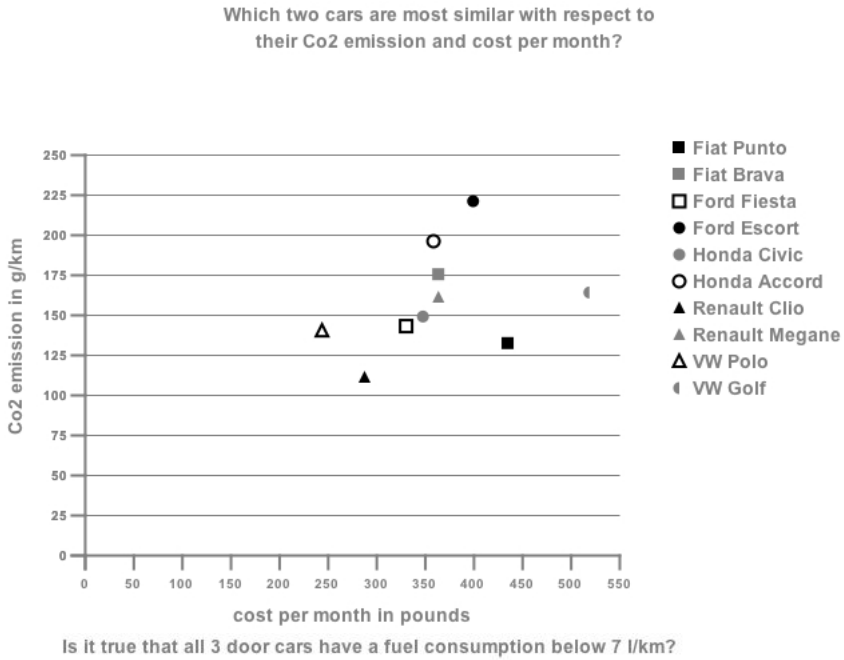


Fig. 4. Examples of ERST Euler's circles and plot representations

Participants then answered the question using the chosen visualization. ERST logged the time to answer the question using the chosen ER. Participants were not permitted to select a different representation following their initial selection. This constraint was imposed in order to encourage participants to carefully consider which representation was best matched to the task. Following a completed response, participants were presented with the next task and the sequence was repeated.

The following data was recorded: (1) the users' representation choices (display selection accuracy - DSA); (2) time to read question and select representation (display selection latency - DSL); (3) time to answer the question using chosen representation (database query latency - DBQL); (4) participants' responses to questions (database query answer - DBQA); and (5) the randomized position of each representation icon from trial to trial;

In addition, the ERST system used by evaluation group participants recorded details of any adaptations to the user that it made, ie: (6) representations eliminated from the selection interface, for a particular user, for a particular database question (display selection reduction - DSR); (7) any representations that were actively recommended to the user for a particular query task (display selection highlight -DSH); and (8) timing of such adaptive responses.

4 Findings

Over all tasks the evaluation group, which used the adaptive version of ERST scored slightly higher on database query answer (DBQA) performance (88% compared to 83% response accuracy) than the comparison group.

ERST dynamically updates its user model according to its user interactions. Not all adaptations were able to take place in early trials, because of sparse user data. After a few trials, enough data for ERST's user model was gathered and the system started to adapt. If the user spent too much time selecting an appropriate representation, the system recommended an ER - one which it believed would be optimal for answering that particular task for that user. The user could then either follow ERST's recommendation or decide to choose a different representation.

The second type of ERST's adaptation consists of hiding ERs in cases where it believes the user is not be able to successfully answer the particular database query with that representation, based on the user's history.

Database query performance (DBQA) for ERST's adaptations combined (highlighting and reducing ERs) for early to late trials (first and last 6 trials) shows an increase of 78% to 94% within the evaluation group. In contrast the equivalent comparison group data (where the non adaptive version of ERST was used) shows a similar increase from an average DBQA performance of 80% on early trials to 90% on late trials.

An evaluation method recommended by e.g. [25] is to break the adaptation down to its constituents. Figure 5 shows database query performance (DBQA) over ERST's different types of adaptation decisions for early and late trials.

Starting from early DBQA performance (78%) the highest increase can be seen in cases where users followed ERST's advice and selected the recommended representation (display selection highlight selection - DSHS). Here, every database query was answered correctly on late trials (DBQA = 100%). In contrast, *not following* ERST's advice resulted in poorer DBQA late-trial performance 75% (even lower than that on early trials).

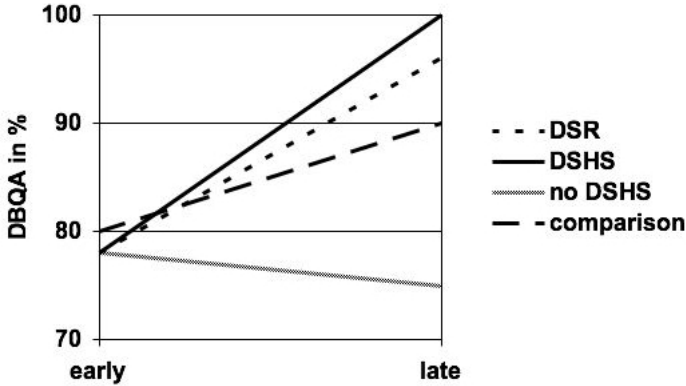


Fig. 5. Database query performance (DBQA) over ERST's adaptations in % for early and late trials. ERST's adaption for hiding a representations - DSR (display selection reduction); The user selected the recommended representation - DSHS (display selection highlight selection); User did not select ERST's highlighted representation - no DSHS; Comparison group performance (which used non adaptive ERST) - comparison.

ERST's other strategy (hiding particular ERs) resulted in an increase of DBQA performance to 96% on late trials. This is similar to trials where ERST did not adapt at all (within the evaluation group), DBQA on late trials increased to 96% as well. Hence 'recommending' seems to be a better intervention strategy than 'hiding'.

4.1 ERST's Adaptation in Relation to Task-Types and the Representational-Specificity of Different Tasks

The adaptations ERST conducted for each task type can be seen in table 1. The system was more active on its adaptation for some types of representations than others (like plot or bar charts), because these representations were used more often. For example, tables were selected 116 times, and scatterplots 114 times, whereas sector graphs were selected only 9 times (all of out of 360).

Some representations were highlighted by ERST on some tasks, but hidden for other tasks, such as the bar chart which was highlighted in the negative comparison task 22 times, whereas it was reduced 6 times in the cluster task, and 5 times in the identify task 1.

For highly representation-specific tasks (e.g. cluster or correlate), ERST was more robust in its adaptations. For example, on cluster tasks, the system recommended the plot chart 19 times but reduced the bar chart's availability 6 times (equivalent figures for sector graphs and set diagrams were 2 and 1 respectively).

Whereas for example, on the low representationally-specific quantifier-set task, ERST recommended set diagrams 17 times and tables 5 times. It reduced access to the table ER once and plot chart twice. It can be seen that in quantifier-set tasks,

Table 1. ERST’s adaptations during evaluation according to task and representation type. Values for highlighting a particular representation in **bold** and reducing a representation from the selection menu in *italic*.

<i>Task type</i>	<i>Bar chart</i>	<i>Pie chart</i>	<i>Plot chart</i>	<i>Sector graph</i>	<i>Set diagram</i>	<i>Table</i>
Identify	5		1			8
Correlate	3		17		2	
Quantifier-set			2		17	5, 1
Locate			1		1	13
Cluster	6		19	2	1	
Compare-neg	22	1			1	1

ERST recommending the set diagram to some users, and tables to other users, and sometimes reduced access to tables on some trials and encouraged their use on other occasions.

4.2 ERST’s Adaptation Behaviour Related to Users’ Prior Knowledge of ERs (KER)

Lower performance on the KER pre-tests was associated with more intervention by ERST during the database query trials. Low functional knowledge of maps, node and arrow/arc network and set diagrams was associated with high levels of ERST adaptation. Functional knowledge of maps is significantly negative correlated to ERST adaptation ($r=-.70$, $p<.05$), functional knowledge of node and arrow/arc networks is also significantly negative correlated ($r=-.71$, $p<.01$) as well as functional knowledge of set diagrams ($r=-.64$, $p<.05$).

It is interesting to note that these three KER sub-scores are derived among the most spatial types of ERs. Set diagrams use a spatial (containment) diagrammatic metaphor to represent set membership. In contrast to e.g. maps which are isomorphic and scaled representations of real-world space.

Exploratory regression analyses suggests that pre-test knowledge of highly spatial ER forms (set diagrams, maps etc.) are particularly predictive of ER selection accuracy and hence ERST’s degree of adaptation. In future systems, this knowledge could be used to systematically test the level of individuals ER-to-task matching skills and knowledge of ERs - further research on this issue is planned.

5 Discussion

In the current study it was interesting to note that the ‘best’ representation was sometimes associated with poorer performance than other (suboptimal) representations, presumably due to individual differences in prior knowledge. This was observed in participants 1 and 9. Here, because of low prior knowledge scores, ERST recommended the tabular representation instead of the set diagram for

quantifier-set tasks resulting in 100% success rate in database query answer performance. In this case adaptive ERST provides better reasoning support to such an individual than a system that recommends ‘optimal’ representations to all its users.

It was noted that subjects sometimes failed to follow the adaptive systems advice, which resulted in worse performance. The more representation-specific the task, the poorer the performance, if advice was not followed. For example, performance decreased dramatically on the highly representation-specific cluster task if ERST’s advice was not followed (database query performance of 94% if participants followed the recommendation, in contrast to a performance of 0% if ERST recommendation was not followed). In contrast performance stayed the same whether following or ignoring the recommendation on the low-representation specific tasks (e.g. the locate task with database query performance of 100% for following/not following ERST’s advice). A future system could indicate to its user the ‘cost’ associated with following/not following its advice.

Looking at individual recommendations, here some users refused to follow ERST advice on particular representations, which then were used successfully on other tasks. Participant 5, who refused to select the recommended ER, bar chart in negative comparison tasks, did select the bar chart on cluster and quantifier-set tasks with a surprisingly high (100%) database query response success. This can be contrasted with the recommendation of set diagrams. Here some participants refused to follow the recommendation and also never selected it on any other tasks. For example, participant 10 never selected the set diagram and did not follow ERST’s advice to choose it on quantifier-set tasks. The findings suggest that participant 10 did not know how to use the recommended ER. A future ERST-tutor version of the system could offer tutorial clarification about the functionality and cognitive and semantic properties of particular ER forms.

6 Conclusion

The current version of ERST is able to increase users’ ER reasoning performance through recommending the most appropriate display. ERST’s other adaptation strategy (varying the range of ‘permitted’ displays available to the user for selection) was not as successful in terms of increasing reasoning performance. This will be further investigated.

The next step will be to improve ERST’s sophistication by enabling it to generate ER-to-task matching situations (contrived query trials) more proactively than it currently is capable of i.e. to systematically ‘probe’ an individual users’ knowledge of ERs and his/her ER-to-task matching skills.

Then, in the case of a user manifesting high error rates for particular task/ER combinations, ERST could offer tutorial clarification, e.g. about the functionality of a particular ER, its cognitive and semantic properties, provide examples of good practice in its use, etc.

References

1. Brusilovsky, P., Eklund, J.: A study of user model based link annotation in educational hypermedia. *Journal of Universal Computer Science*, special issue on assessment issues for educational software **4** (1998) 429-448
2. Cheng, P.C.-H.: Functional roles for the cognitive analysis of diagrams in problem solving. In: Cottrell, G.W., eds.: *Proceedings of the 18th Annual Conference of the Cognitive Science Society*. Mahweh NJ, Lawrence Erlbaum Associates (1996) 207-212
3. C. Conati, A. Gertner and K. VanLehn: Using Bayesian networks to manage uncertainty in student modeling. *User Modeling & User-Adapted Interaction* **12** (2002) 371-417
4. Cox, R., Brna, P.: Supporting the use of external representations in problem solving: The need for flexible learning environments. *Journal of Artificial Intelligence in Education* **6** (1995) 239-302
5. Cox, R., Stenning, K., Oberlander, J.: The effect of graphical and sentential logic teaching on spontaneous external representation. *Cognitive Studies: Bulletin of the Japanese Cognitive Science Society* **2** (1995) 5-75
6. Cox, R.: Representation construction, externalised cognition and individual differences. *Learning and Instruction* **9** (1999) 343-363
7. Cox, R., Grawemeyer, B.: The mental organisation of external representations. *European Cognitive Science Conference (EuroCogSci)*. Osnabrück (2003)
8. Cox, R., Romero, P., du Boulay, B., Lutz, R.: A cognitive processing perspective on student programmers' 'graphicacy'. In: Blackwell, A., Marriott, K., Shimojima, A., eds.: *Diagrammatic Representation & Inference*. Volume 2980 of *Lecture Notes in Artificial Intelligence*. Springer (2004) 344-346
9. Day, R.: Alternative representations. In: Bower, G., eds.: *The Psychology of Learning and Motivation* **22** (1988) 261-305
10. Grawemeyer, B., Cox, R.: The effects of knowledge of external representations and display selection upon database query performance. *Proceedings of the Second International Workshop on Interactive Graphical Communication (IGC2003)*. Queen Mary, University of London (2003)
11. Grawemeyer, B., Cox, R.: A Bayesian approach to modelling user's information display preferences. In: Ardissono, L., Brna, P., Mitrovic, T., eds.: *UM 2005: The Proceeding of the Tenth International Conference on User Modeling*. Volume 3538 of *Lecture Notes in Artificial Intelligence*. Springer (2005) 233-238
12. Grawemeyer, B., Cox, R.: Graphical data displays and database queries: Helping users select the right display for the task. In: Butz, A., Fisher, B., Krüger, A., Olivier, P., eds.: *Smart Graphics, 5th International Symposium on Smart Graphics, SG 2005*. Volume 3638 of *Lecture Notes in Computer Science*. Springer (2005) 53-64
13. Humphreys, G.W., Riddoch, M.J.: *Visual object processing: A cognitive neuropsychological approach*. Lawrence Erlbaum Associates, Hillsdale NJ (1987)
14. Jameson, A., Gromann-Hutter, B., March L., Rummer, R.: Creating an empirical basis for adaptation decisions. In: Lieberman, H., eds.: *IUI 2000: International Conference on Intelligent User Interfaces*. (2000)
15. Kirby, J.R., Moore, P.J., Schofield, N.J.: Verbal and visual learning styles. *Contemporary educational psychology* **13** (1988) 169-184
16. M. Kozhevnikov, M. Hegarty and R.E. Mayer: Visual/spatial abilities in problem solving in physics. In Anderson, M., Mayer, B. and Olivier, P., eds.: *Diagrammatic representation and reasoning*. Springer (2002)

17. Mitchell, T.M.: Machine learning. McGraw Hill, New York (1997)
18. Norman, D.A.: Things that make us smart. Addison-Wesley, MA (1993)
19. Novick, L.R., Hurley, S.M., Francis, M.: Evidence for abstract, schematic knowledge of three spatial diagram representations. *Memory & Cognition* **27** (1999) 288-308
20. Novick, L.R., Hurley, S.M.: To Matrix, Network, or Hierarchy: That Is the Question. *Cognitive Psychology* **42** (2001) 158-216
21. Pearl, J.: Probabilistic reasoning in intelligent systems: Networks of Plausible Inference. Morgan Kaufmann (1988)
22. Shute, V.J., Gawlick-Grendell, L.A., Young R.K., Burnham, C.A.: An experimental system for learning probability: Stat Lady description and evaluation. *Instructional Science* **24** (1996) 25-46
23. Stenning, K. , Cox, R., Oberlander, J.: Contrasting the cognitive effects of graphical and sentential logic teaching: Reasoning, representation and individual differences. *Language and Cognitive Processes* **10** (1995) 333-354
24. Vessey, I.: Cognitive fit: A theory-based analysis of the graphs versus tables literature. *Decision Sciences* **22** (1991) 219-241
25. Weibelzahl, S.: Problems and Pitfalls in Evaluating Adaptive Systems. In Chen, S. and Magoulas, G., eds.: *Adaptable and Adaptive Hypermedia Systems*. Heshy, PA: IRM Press (2005) 285-299