Development and Evaluation of a Model of Behavioral Representation Techniques

J.D. Chase, Robert S. Schulman, H. Rex Hartson, and Deborah Hix

TR 93-34

Department of Computer Science
Virginia Polytechnic Institute and State University
Blacksburg, Virginia 24061

November 1, 1993

DEVELOPMENT AND EVALUATION OF A MODEL OF BEHAVIORAL REPRESENTATION TECHNIQUES

J. D. Chase[1], Robert S. Schulman[2], H. Rex Hartson[1], and Deborah Hix[1]

Department of Computer Science[1] and Department of Statistics[2], Virginia Tech, Blacksburg VA 24061

internet: chase@csgrad.cs.vt.edu, 703-231-6470

KEYWORDS

Usability, Behavioral Representation Techniques, Interaction Development, Model, Empirical Evaluation

ABSTRACT

A user-centered approach to interactive system development requires a way to represent the behavior of a user interacting with an interface. While a number of behavioral representation techniques exist, not all provide the capabilities necessary to support the interaction development process. Based on observations of existing representation techniques and comments from users of the User Action Notation (UAN), a user- and task-centered behavioral representation technique, we have developed a model of behavioral representation techniques. Our model is an epistemological framework for discussing, analyzing, extending, and comparing existing behavioral representation techniques, as well as being a springboard for developing and evaluating new techniques. We present the model and results of our evaluation demonstrating the model's reliability and utility within the context of behavioral representation techniques.

PROBLEM STATEMENT AND CONTEXT

It has been shown repeatedly that traditional software engineering methods do not necessarily lead to high usability when applied to the development of user interfaces (e.g. 1, 2, 3). This is reasonable since the focus of these methods is the software, not the user. To achieve high usability, the interface development process should be user-centered; i.e., it should focus on the user's tasks, needs, and behavior while interacting with the system. This view, referred to as the *behavioral view* (4), has led to a variety of techniques for representing the design of a user interface in terms of the *behavior of the user*, independently of user interface software and hardware considerations.

One such technique, the User Action Notation (UAN) (4, 5), is a user- and task-oriented notation that describes the behavior of a user and an interface during their cooperative performance of a task. Other behavioral representation techniques include GOMS (6), Command Language Grammar (CLG) (7), keystroke-level model (8), Task Action Grammar (TAG) (9), Reisner Action Language (10), work by Kieras and Polson (11), scenarios or story-boarding, and Task Artifact Framework (12).

The primary abstraction of the UAN is a *user task* — a user action or group of temporally related user actions performed to achieve a work goal. A user interface is represented as a quasi-hierarchical structure of tasks that are asynchronous — the sequencing within each task is independent of that in other tasks. User actions, corresponding interface feedback, and state information are presented at the lowest level. Levels of abstraction are used to hide the details of lower level tasks by combining these tasks under a single task name. At all levels, user actions and user tasks are ordered and combined using temporal relations such as sequencing, interleaving, and concurrency. Since textual notations are not always convenient for specifying all components of an interface, UAN descriptions may include screen pictures (scenarios), and can be supplemented with state transition diagrams to indicate precisely how the user interacts with the interface.

The UAN has progressed from its inception as a research project to a practical approach that has been used extensively in real world development environments. Over the past 3 years, we have observed use of the UAN in a substantial portion of the more than 50 industrial and government sites that are using the UAN for user interface development. We have interacted with these users of the UAN to obtain their feedback for improving and extending the UAN.

However, as we began making extensions to the UAN based on data from its users, we found that there was no structured method for determining whether our changes were actually improving the notation. This lack of either a practical or theoretical yardstick for determining improvement in the UAN led us to develop and evaluate a model of behavioral representation techniques. The model is an epistemological framework for discussing, analyzing, extending, and comparing existing behavioral representation techniques, as well as being a springboard for developing and evaluating new techniques.

APPROACH

The traditional scientific method consists of *observation*, *theorization*, and *evaluation*, where a model is often used to aid in expression of the theory. These steps, when applied to the development of a theoretical model of behavioral representation techniques, translate to:

1. observation of behavioral representation techniques and their use, as well as study of related existing models found in the literature;

2. construction of a model of behavioral techniques based on those observations;

3. evaluation of the proposed model of behavioral representation techniques.

Each of these steps is discussed in later sections.

A RELATED MODEL

At least one previous attempt has been made to produce a theoretical model of techniques to represent interface designs from the user's perspective. Simon (13) developed a "trade-off" space of user models, the purpose of which was to create a theoretical view of the strengths and weaknesses of any given representation technique. The three dimensions of this model were: Processing Resources (e.g., actions, cognition, perception, and sensation), Knowledge Representation (e.g., from high to low continuously), and Level of Representation (e.g., from actual operations specified to no operations specified continuously). This space was to help a designer/analyst choose among available techniques according to these dimensions.

While this trade-off space raises many interesting questions about relationships among various techniques, it has some limitations. First, the space has not been evaluated in any way. Second, the use of continuous axes for two of the dimensions raises questions about where a given technique should reside. For example, the Level of Representation axis refers to the specificity of the technique with regard to user operations, ranging from the top of the axis where operations are not specified at all to the bottom of the axis where operations are specified in detail. For a technique that is very explicit about cognitive operations but not about physical ones, the model is not clear about whether that technique reside near the top or the bottom of the Level of Representation axis?

STEP 1: OBSERVING USERS OF UAN AND COLLECTING DATA FOR A MODEL OF BEHAVIORAL REPRESENTATION TECHNIQUES

We believe ours to be the first *model* to provide a general framework for discussing and evaluating behavioral representation techniques. During model formulation, we collected data through observations from several sources. In addition to observing use of the UAN, we also studied related models, and approaches to their development, found in the literature. For example, in the area of input devices and device selection, researchers have found it useful to build a model of devices in terms of their attributes and classes of attributes (14, 15). Similarly, in our research, the behavioral representation techniques to be modeled have a number of attributes based on the needs of their users. With the goal of identifying and classifying these attributes, users of the UAN were asked to contribute a list of their needs for a behavioral representation technique. The lists contained suggestions such as representing hierarchical relationships among user tasks, providing analytical support for complexity and performance, and providing support for interaction documentation.

Further, existing behavioral representation techniques, such as GOMS, CLG, TAG, etc., were analyzed to determine requirements for which they were designed. We found items such as performance prediction, cognitive modeling, and object and artifact definitions.

STEP 2: DEVELOPING A MODEL OF BEHAVIORAL REPRESENTATION TECHNIQUES

A model of behavioral representation techniques was then developed based on these observations and data. Because our model is taxonomical, it is a scientific metaphor for the systemization of attributes of a group or class, in the style of Kaplan (16). The model, shown in Figure 1, contains three dimensions, each of which has several discrete attributes:

- *Scope* — activities within the interface development process that can use the technique. These activities (attributes) include task analysis, design, and so on.

- *Content* — interface or interaction components being represented using the technique. These components (attributes) include user definition, cognitive processes, and so on,

- *Requirements* — qualities of the representation. These qualities (attributes) include facility, expressiveness, and so on.

STEP 3: EVALUATING THE MODEL OF BEHAVIORAL REPRESENTATION TECHNIQUES

The model was evaluated over a variety of criteria. The first criterion we examined was that of *completeness*, defined to be the model's coverage of all aspects of behavioral representation techniques. Developers of the device models mentioned earlier (14, 15) found it virtually impossible to test completeness of their model; we, too, found this to be the case in evaluating our model's coverage of behavioral representation techniques. However, we postulate that our model should provide ample coverage at least of existing techniques since it was derived from an analysis of them.

Evaluation of the device models centered on the criterion of *utility* (14, 15), meaning that the models provided some

useful function in working with input devices. However, another important criterion is that of *reliability*, meaning that the model must produce similar results across those who use it to compare and evaluate behavioral representation techniques. The model of behavioral representation techniques was also evaluated under these two criteria, as described in the following sections.

Demonstrating Reliability of the Model

As described previously, the purpose of our model is to support development and evaluation of new behavioral representation techniques, as well as analysis, extension, and comparison of existing techniques. Thus, this model supports two basic functions:

- mapping problems or critical incidents observed in the use of a specific behavioral representation technique to a missing or failing activity (from the Scope dimension), component (Content dimension), and quality (Requirements dimension); and
- analyzing an existing behavioral representation technique for its ability to cover all activities, components, and requirements identified in the model.

Moran (17) raised the need for the first of these two functions in describing the need to categorize and analyze hundreds of "nits" or critical incidents involving interface development at Xerox (17). Jacob (18) raised the need for the second of these functions in discussing the need to analyze and compare new and existing user models and design representations.

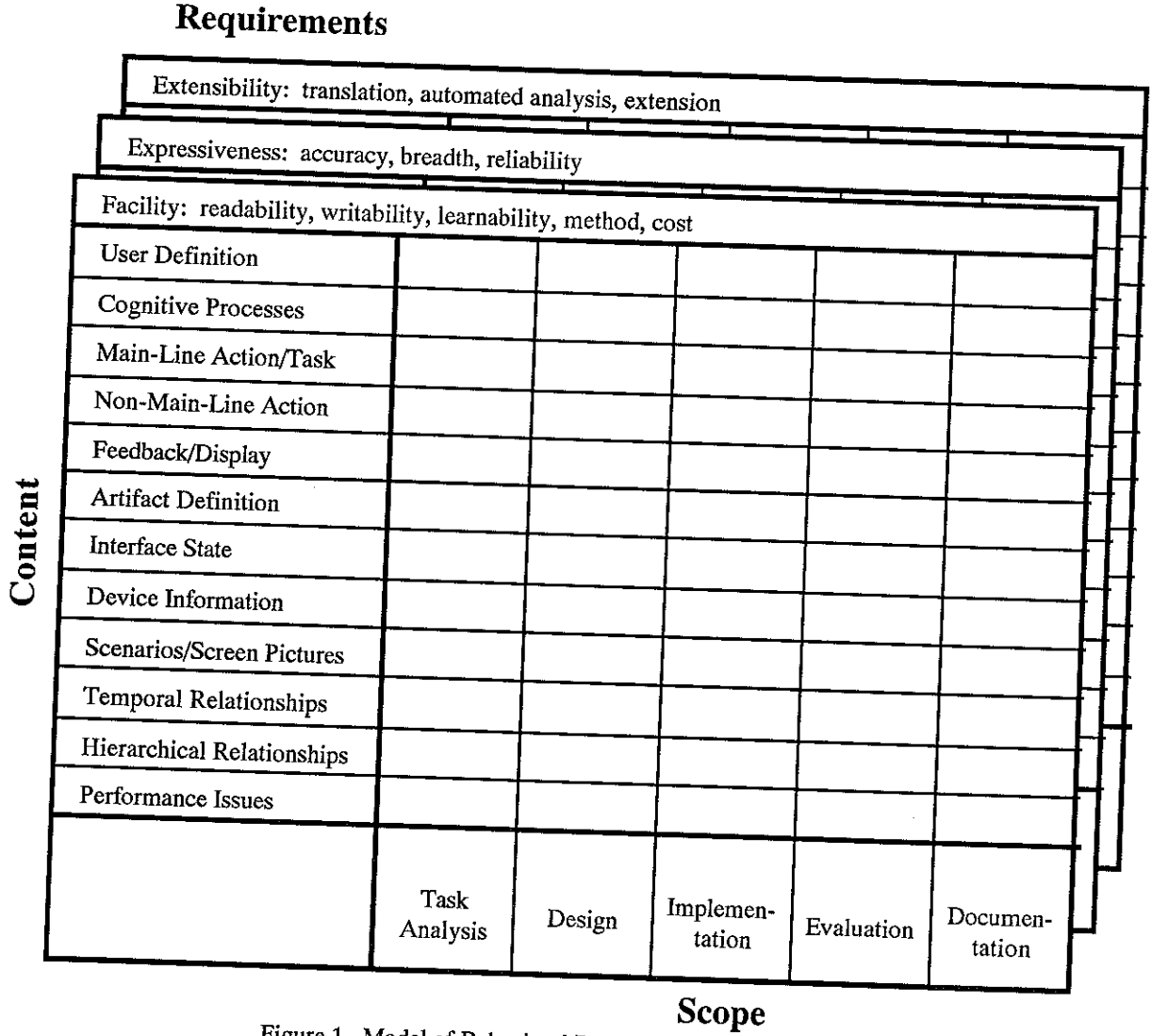

Figure 1. Model of Behavioral Representation Techniques

The model was evaluated for *reliability* by having a group of subjects perform several tasks and then performing a statistical test of the similarity of their results. In the first task, involving incident mapping, we selected 40 critical incidents commonly reported by UAN users and had five subjects map these incidents into the model. For purposes of this mapping, we defined critical incidents to be:

- an encounter with an interface component that the

UAN does not represent;

- a difficulty using the existing UAN notation;
- a variation in notation or method from the current UAN standard; or
- an incident in which the UAN provides a notation or a view of the interface that could not have occurred otherwise.

We selected the 40 critical incidents to provide the broadest possible coverage of the model, i.e., so that they would not all map to a small number of cells. The five subjects were selected on the basis of their varying levels of experience with the UAN. We intentionally chose a mix of subjects from UAN novices to UAN experts in order to demonstrate that the model would produce reliable results across experience levels. Subjects were shown two example mappings and were given a brief description of the model. Each subject then independently mapped each of the 40 incidents to the cell in the model in which they thought it best fit. For example, if an incident was reported in which two UAN users created different hierarchical task structures during task analysis, then that incident might be mapped to the cell of *task analysis* on the *Scope* axis, *hierarchical relationships* on the *Content* axis, and *expressiveness* (which includes the idea of reliability of the technique) on the *Requirements* axis.

The reliability of the model is the extent to which the five subjects agree on their mappings of the 40 critical incidents. Of course *some* agreement in mappings would be expected by chance alone: that is, even if the subjects assigned the incidents to the cells at random. Cohen's kappa (19) is a measure of the proportion of agreement above and beyond what would be expected on the basis of chance. Kappa is scaled between 0 and 1, with 0 corresponding to only chance agreement and 1 corresponding to perfect agreement. Kappa is approximately normally distributed and can be used to test the null hypothesis of no agreement beyond the chance level.

While kappa is limited to assessing the agreement between two subjects, an extension (20) permits measuring agreement among several subjects. As with the two-subject version, the extension also produces a kappa value between 0 and 1, and also allows testing for agreement by reference to the normal distribution. In this study we employed the extended version to measure agreement among our five subjects. To gauge the reliability of each dimension in the model individually, separate tests were conducted for each of the three axes. Table 1 shows the proportion of agreement, kappa, z-value (standard normal scaling of kappa), and p-value for each of the three dimensions. For each test, the p-value is the chance of obtaining the observed level of agreement by chance alone. The extremely small p-values shown in Table 1 provide convincing evidence that each dimension of the model is reliable.

Table 1. Results of analysis of incident mapping

Dimension	Proportion of Agreement	Kappa	z-value	p-value
Scope	0.8725	0.8362	28.56	<0.00001
Content	0.6575	0.6224	36.04	<0.00001
Requirements	0.8925	0.8327	20.10	<0.00001

While these results show highly significant levels of agreement (i.e., well beyond chance) among the five subjects, the level of agreement on the Content axis appears lower, since the proportion of agreement and kappa values for the Content axis are well below those of the other two axes. However, since there are 12 levels on the Content axis but only 3 and 5 levels, respectively, on the Scope and Requirements axes, it is less likely that all subjects would agree on the best fit of an incident along this axis. As indicated by the larger z-value for the Content axis, the obtained level of agreement provides even more convincing evidence of reliability for this dimension than for either the Scope or Requirements axes.

We should note that there are two domains to which we might like to generalize our results: the domain of critical incidents and the domain of subjects. For statistical purposes, the sample size of 40 for each test (the number of critical incidents) implies that our model is reliable for the domain of all critical incidents. Although our five subjects were chosen to span a wide range of experience with the UAN, the statistical results do not speak to generality in this domain; we can only speculate that similar consistency would be obtained among any reasonable set of subjects.

These results suggest that two or more persons working to modify and/or extend a behavioral representation technique would most likely agree on the ramifications or meaning of a particular critical incident, i.e., the Scope, Content, and Requirements components of the technique which need to be addressed.

In the second task used to evaluate reliability of the model of behavioral representation techniques, we had the same five subjects from the previous experiment use the model for analysis of the UAN. This process involved going through each cell in the model and rating the UAN on a scale of 1 to 3, 1 being the lowest rating and 3 being the highest rating, with respect to that cell. Each subject was given a copy of a UAN tutorial. Using the previous mapping example, i.e., the cell of task analysis, hierarchical relationships, and expressiveness, the UAN might be assigned a score of 3 for this cell since it is very effective at representing decomposition of larger tasks into a hierarchical task structure.

While perfect agreement among all subjects in the rating of

a cell would provide the most convincing evidence of reliability, it would be unreasonable to ignore partial agreements. Accordingly, we considered two possible criteria for "agreement," each allowing for some minor degree of discrepancy among the five ratings for an intersection. For the first criterion, which we refer to as *3 or more agreement*, three subjects had to agree exactly on the score and the other two subjects could differ by at most one from that score, e.g., three subjects give a score of 3, and the other two both give scores of 2. For the second criterion, which we refer to as *4 or more agreement*, four subjects had to agree exactly on the score and the other could differ by at most one, e.g., four subjects select 1 and the other subject selects 2. For each criterion we compared the observed proportion of agreements with the proportion that would be expected on the basis of chance alone. These proportions, along with the resulting z-values and their associated p-values, are shown in Table 2.

Table 2. Results of analysis of ratings comparisons

Criterion	Probability of Agreement by Chance	Proportion of Agreement	z-value	p-value
3 or more	0.3416	0.7795	12.89	< 0.00001
4 or more	0.0947	0.5641	22.39	< 0.00001

Again, as with the incident mapping task, these results indicate a highly significant level of agreement among the five subjects. While the proportion of agreement is somewhat lower for the *4 or more agreement* criterion, the probability of obtaining such agreement by chance is extremely small and therefore the resulting significance is actually higher than that of the *3 or more agreement* method. We can conclude from these results that the model is reliable for the analysis/rating of an existing behavioral representation technique. This type of analysis may provide insight into areas of needed improvement for a given technique. Further, this analysis may provide a means of comparing existing behavioral representation techniques to determine which is best for the task at hand.

Demonstrating Utility of the Model

The other criterion for evaluation of the model of behavioral representation techniques was *utility* — to show that it can provide useful functionality in some measurable way. To do this, our model was applied to the problem of UAN development and extension. If a new version of the UAN based on the model could be shown to be an improvement over the old version, then the conclusion could be drawn that the model of behavioral representation techniques does, indeed, provide useful functionality. There were two parts to this evaluation of model utility.

In order to perform the first part of the experiment, we wrote a UAN tutorial for the original version of the UAN, providing basic information of how to use the UAN in representing an interface design. This original version of the UAN was the one used by most of the real world development sites, mentioned earlier. Using this tutorial, the original UAN was introduced (or in some cases re-introduced) to approximately 30 industrial and academic interface development sites. These sites were asked to report to us critical incidents in using the UAN. We mapped reported incidents onto the model of behavioral representation techniques to identify areas of needed improvement. We used these mappings, along with the ratings of the UAN based on the model, and other observations and literature study (as discussed previously and detailed in (21)) to create an improved, extended version of the UAN. This new version was documented with a new version of the UAN tutorial.

In order to evaluate utility of the model, we selected six subjects from a graduate human-computer interaction course in which all six had equal exposure to the original UAN and other human-computer interaction development concepts. These subjects were given a single simple interface task to describe using the original UAN. The subjects were also asked to supply their grade point average. Based on the quality of their UAN descriptions and their grade point averages, the six subjects were divided into two groups of three subjects each, roughly comparable in both UAN skills and grade point average. One group was given the old UAN tutorial and the other group was given the new tutorial.

We then gave both groups the same interface prototype, a graphical interface prototype for a disability assessment system (developed in Hypercard), and asked them to write a description of its interface using *only* their respective versions of the UAN. Each group was cautioned to include in their UAN descriptions just the features and symbology of the UAN contained in their tutorial. The groups met separately and were given four hours to complete the UAN descriptions for the disability assessment interface. The members of each group were then asked to rate their respective versions of the UAN using the model of behavioral representation techniques. They accomplished this by rating their versions at each cell in the model on a scale of 1 to 3 (with 1 being the lowest rating and 3 being the highest rating).

We then compared the old and new versions using two criteria:

- each subject's perception of the version of the UAN he or she used as recorded by model rating results and
- quality of UAN task descriptions for the disability assessment interface as determined by a panel of

human-computer interaction experts.

Our behavioral representation technique model includes a total of 180 cells (5 levels of Scope by 12 levels of Content by 3 levels of Requirements), each of which was rated on the three-point scale by each subject. For each cell two averages were computed, one across the three subjects who used the old version of the UAN, and one across the three subjects using the new version of the UAN. Old and new versions of the UAN were then compared across all 180 cells using a paired t-test. As can be seen in Table 3, we are extremely confident that the new version of the UAN is rated higher than the old version.

Table 3. Results of t-test of user ratings comparison

Number of Observations	Mean Difference in Ratings	t value	p value
180	0.7044	15.20	< 0.00001

The mean rating for the old version across all 180 cells was 1.5930, while the mean rating for the new version was 2.2974. Although these numbers may appear similar, we should bear in mind that our scale allowed ratings only between 1 and 3, so that the observed difference of 0.7044 represents approximately 35% of the entire scale. Looked at from another angle, we could state that the new version produced a rating 44% higher than the old version.

In the second part of the experiment to evaluate improvement in the UAN, a panel of four subjects rated the design representations created by the groups from the first part of the experiment. These subjects were selected based on their expertise in human-computer interaction, as well as their level of experience with the UAN. Each of these subjects compared the two sets of UAN design representations to the interface prototype and rated each set at 36 of the 180 cells of the model. (Since actual design representations were being rated, the only relevant level of the Scope axis was the *design* activity. Each UAN design description was therefore rated for each combination of the 12 levels of Content by 3 levels of Requirements, resulting in 36 ratings.) Presentation order of original and new versions of UAN was balanced across subjects. Again, results were compared using a paired t-test, the results of which are summarized in Table 4.

Table 4. Results of t-test of expert ratings comparison

Number of Observations	Mean Difference in Ratings	t value	p value
36	0.361	3.28	0.0012

Across all 36 cells rated, the mean rating of the UAN design representations created using the original version was 1.3889, while the mean for the UAN design representations based on the new version was 1.75, representing an improvement of 26%.

Since the new version of the UAN was developed using the model of behavioral representation techniques, we can conclude that the model does provide useful function in a statistically measurable way.

SUMMARY AND CONCLUSIONS

Our original goal of improving the User Action Notation (UAN) revealed the need for a taxonomical model for discussing, comparing, and evaluating behavioral user interface design representation techniques. We developed such a model through a process of observation, theory formulation, and evaluation. The model contains three dimensions, each with discrete attributes: Scope (activities within the interface development process that can use the technique), Content (interaction components that can be represented), and Requirements (qualities such as learnability, accuracy, and reliability).

It was not possible to evaluate the model on the basis of completeness, but it was evaluated with regard to reliability and utility. Through two empirical evaluations, we have shown that our model of behavioral representation techniques is *reliable* for incident mapping and technique analysis. We used an extension of Cohen's kappa as a measure of agreement to show that subjects agreed, beyond the level expected by chance, on mappings from critical incidents observed in the use of representation techniques to cells in the model. On average, we achieved more than 75% of the possible agreement beyond chance. We thus conclude that this model should be useful to behavioral representation technique designers/analysts who are iteratively refining an existing behavioral representation technique, even if results are being compared across different individuals/analysts. The fact that the model tested reliably also suggests that there are discrete, discernible attributes of behavioral representation techniques. Our use of a mix of subjects from novices to experts implies that these attributes are recognizable even to novice interface designers/analysts. This reinforces the idea that an interface designer/analyst should be able to recognize these attributes, select the ones that are most important in their environment and choose a technique accordingly using the model and associated ratings of behavioral representation techniques.

The *utility* of the model was also demonstrated in terms of its ability to provide the basis for an improved version of the UAN. Users gave the new UAN 44% higher subjective ratings. Also, the quality of task descriptions made using

the new UAN was judged to be 26% better than those created using the original UAN.

The model provides the ability to predict problems with behavioral representation techniques in areas where none were reported and foresee changes that could be useful. Just as with interfaces themselves, anything that can reduce the amount of empirical testing of a behavioral representation technique is indeed useful.

REFERENCES

1. J. D. Gould & C. Lewis, Designing for Usability: Key Principles and What Designers Think, *Commun. ACM* 28, 300-311 (1985).

2. H. R. Hartson, D. Hix, Toward Empirically Derived Methodologies and Tools for Human-Computer Interface Development,*Int. J. Man-Machine Studies* 31, 477-494 (1989).

3. M. B. Rosson, S. Maass & W. A. Kellogg, Designing for Designers: An Analysis of Design Practice in the Real World, *CHI+GI Conference on Human Factor in Computing Systems* (ACM, New York, 1987), 137-142.

4. D. Hix & H. R. Hartson, *Developing User Interfaces: Ensuring Usability Through Product and Process.* (John Wiley & Sons, Inc., New York, 1993).

5. H. R. Hartson, A. C. Siochi, D. Hix, The UAN: A User-Oriented Representation for Direct Manipulation Interface Designs, *ACM Trans. on Info. Sys.* 8, 181-203 (1990).

6. S. K. Card, T. P. Moran, A. Newell, *The Psychology of Human-Computer Interaction.* (Erlbaum, Hillsdale, NJ, 1983).

7. T. P. Moran, The Command Language Grammar: A Representation for the User Interface of Interactive Computer Systems, *Int. J. Man-Machine Studies* 15, 3-51 (1981).

8. S. K. Card, T. P. Moran, The Keystroke-Level Model for User Performance Time with Interactive Systems,*Commun. ACM* 23, 396-410 (1980).

9. S. J. Payne, T. R. G. Green, Task-Action Grammars: A Model of the Mental Representation of Task Languages, *Human-Computer Interaction* 2, 93-133 (1986).

10. P. Reisner, Formal Grammar and Human Factors Design of an Interactive Graphics System, *IEEE Trans. Soft. Eng.* SE-7, 229-240 (1981).

11. D. Kieras, P. G. Polson, An Approach to the Formal Analysis of User Complexity, *Int. J. Man-Machine Studies* 22, 365-394 (1985).

12. J. M. Carroll, W. A. Kellogg, M. B. Rosson, in *Designing Interaction: Psychology at the Human-Computer Interface* J. M. Carroll, Ed. (Cambridge University Press, New York, 1991) pp. 74-102.

13. T. Simon, Analysing the Scope of Cognitive Models in Human-Computer Interaction: A Trade-Off Approach, D. M. Jones and J. R. Winder, Ed., *People and computers IV: The Fourth Conference of the British Computer Society Human-Computer Interaction Specialist Group* University of Manchester, 1988), vol. 4,

14. T. W. Bleser, Ph. D. Dissertation, George Washington University (1991).

15. J. Mackinlay, S. K. Card, G. G. Robertson, A Semantic Analysis of the Design Space of Input Devices, *Human Computer Interaction* 5, (1990).

16. A. Kaplan, *Conduct of Inquiry : Methodology For Behavioral Science.* (Chandler Pub. Co., 1964).

17. T. P. Moran, in *Methodology of Interaction* Guedj et al, Ed. (North-Holland Publishing Co., 1980) 293-301.

18. R. J. K. Jacob, Using Formal Specification in the Design of Human-Computer Interfaces, *Human Factors in Computer Systems* (ACM Press, Gaithersburg, MD, 1982), pp. 315-321.

19. J. Cohen, A Coefficient of Agreement for Nominal Scales, *Educational and Psychological Measurement* 20, 37-46 (1960).

20. J. L. Fleiss, Measuring Nominal Scale Agreement Among Many Raters, *Psychological Bulletin* 76, 378-382 (1971).

21. J. D. Chase, Ph. D. Dissertation, Virginia Polytechnic Institute and State University (to appear 1994).