Using HCI Experiments to Validate Intelligent Multimedia Cue Generation

Nancy Green

Department of Mathematical Sciences
University of North Carolina at Greensboro
Greensboro, NC 27402 USA
nlgreen@uncg.edu

Abstract

This paper presents a case study of our experience conducting HCI experiments to inform multimedia cue generation. We analyze the interrelated HCI and NLG issues that a complete multimedia cue generator should address, and the limitations of current empirical approaches. We describe the two experiments we performed. Given their limited success, the cost-benefit of adopting this approach remains an open issue

Introduction

Let us consider the choices that a text generator faces in producing phrases that we shall refer to as *multimedia cues*, e.g., *as shown in Figure 2*. Such phrases are used to direct the reader's attention to a graphic in the same document, where a *document* may be presented in print or on-screen. A generator must decide *whether* and, if so, *how often* to explicitly direct the reader's attention to a graphic, *where* to place multimedia cues in the text, and *what* to say in each cue.

To illustrate the *how often* and *where* decision problems, we present an excerpt from a print journal article in column one of Table 1. (In the article, the text was displayed in a normal paragraph format; we present it in a tabular format for discussion purposes.) Most of the paragraph describes a figure, called *Figure 2*, that was presented on a different page than the paragraph. Also, that figure contained two line graphs, one on the left and one on the right. However, the text contains only one multimedia cue and the cue does not distinguish which of the two graphs is described. In column two of Table 1, we list which graph (*left* or *right*) that the text in the same row describes. Also, the first time a sentence about each graph occurs is indicated.

This example shows that a generator must select from among many candidate sites for multimedia cue placement (*where*), e.g. at the beginning/end of the paragraph, before/in/after the first/last sentence about a graphic, etc. This example also illustrates the interrelated problem of *how often*, since discussion about the two graphs switches back and forth, but a simple heuristic such as "use each site" would result in too many cues, and the simple heuristic, "use only one site", may not be sufficient. (Note in the article only one site is used.) As for *what* to say, options range from incomplete identification of the graphic, as used in this example, to providing more specific identifying features, e.g., *(Figure 2, left)*, and interpretive information, e.g., *The blue line in Figure 2 (left) shows U.S. traffic crashes*.

Clause(s)	Graph
In the United States the number of traffic deaths has remained relatively constant at about 41,000 per year for the last decade.	left (1^{st})
In the early years of the 20^{th} century, few people were killed in U.S. traffic crashes because there were few motorized vehicles (Figure 2).	left
As car ownership increased rapidly, so did traffic deaths, peaking in 1972 at 54,598.	left
In nations that are less thoroughly motorized, for example in the world's most populous nation, China, the number of fatalities per year continues to increase.	left
But in all countries, the number of deaths per registered vehicle has declined over time.	right (1^{st})
Since the late 1930s, the U.S. rate has declined by about 2.7 percent per year, or by half every 25 years.	right
If the 1935 rate were to apply to the present U.S. vehicle population, annual U.S. traffic fatalities would exceed a quarter of a million.	neither
Traffic fatalities continue to increase in China,	left
but the rate per vehicle is declining even faster -- by 10 percent per year, or by half in 7 years.	right
Arguably, a higher rate of decrease is to be expected when initial rates are higher.	neither

Table 1. Excerpt from Leonard Evans, "Traffic Crashes", *American Scientist*, May-June 2002, Vol. 90, 244-253.

A text generator may maintain a rich multi-level internal representation of a text to which a multimedia cue generator may have access. Possibly included in the internal representation are factors with a potential bearing on the *where/how often* decision such as discourse structure (e.g., rhetorical or intentional structure, presence of other cue phrases); semantic factors such as equivalence of text content and graphics content; syntactic factors such as type of syntactic constituent and level of embedding of a candidate cue phrase; and/or display features such as sentence length and paragraph length.

Psychologists are investigating issues related to the design of effective multimedia documents. For example, according to Mayer's Spatial Contiguity Principle (2001), "students learn better when corresponding words and pictures are presented near rather than far from each other on the page or screen" (p. 184). While applied psychology and human-computer interaction (HCI) research has provided useful heuristics, the results are targeted more for human designers than for supporting intelligent multimedia generation. Unfortunately, those experiments are not typically designed to answer fine-grained questions about the role of the above text-related factors in multimedia cue generation.

Natural language generation (NLG) researchers attempting to identify text-related factors determining optimal placement and selection of *discourse* cue phrases (e.g., *although*, *thus*) have applied machine learning to annotated print corpora (DiEugenio et al. 1997). However, analyzing multimedia cue usage in print corpora may not provide sound information for multimedia cue generation in on-screen presentations, since HCI studies have shown that there are significant differences between reading from print and electronic documents (e.g., Dillon 1992, Muter 1996).

Another limitation of using corpus studies to develop models for on-screen multimedia cue generation is that the layout of text in the print corpus may differ from the layout that a system must use for on-screen presentation due to medium-related constraints. Furthermore, layout should be designed with the reader's goals in mind (Wright 1999). For example, a reader of a non-user-tailored healthcare document may scan it in a nonlinear manner to find the answer to a particular question. Thus, it is plausible that the effectiveness of on-screen multimedia cues depends not only on properties of the text itself, but also on layout and the user's tasks.

As an alternative to corpus studies, NLG researchers have done ablation experiments on implemented intelligent multimedia generation systems to evaluate alternative strategies (e.g. Carenini and Moore 2000). A practical problem with that approach is that after expending effort to develop a system providing multiple strategies, one may discover that none of the strategies work well enough.

Our goal is to develop a multimedia cue generator, as part of an intelligent generation system, addressing the *where/how often* issue for on-screen presentations. After considering the methodological problems noted above, we decided to conduct HCI experiments using experimenter-constructed presentations (i.e. instead of presentations generated by an implemented system). The empirical results would be used to justify strategies that could be implemented, not just by our system, but other intelligent systems as well. Our plan was to start by confirming *whether* providing multimedia cues in on-screen presentations is worthwhile to users for certain tasks and types of screen layout, using an *intuitively reasonable* approach to cue placement. Our plan was, after having confirmed this basic premise, to perform successive experiments to answer more fine-grained questions of interest to NLG researchers on what text-related factors to consider for *optimal* cue placement.

This paper presents two experiments that we performed to answer the first question (*whether*). Surprisingly, the main results were negative, casting doubt on whether it is worthwhile to pursue the more fine-grained experiments on *where/how often* questions. Thus, although it is rare for NLG researchers to report negative results, we feel that it is worthwhile to present this work, as argued in (Reiter, Robertson, and Osman 2003).

Other Related Work

Although NLG researchers have not investigated multimedia cue placement (*whether/where/how often*), they have investigated other questions relevant to a complete model of multimedia cue generation. Discourse cue and multimedia cue phrases are subsumed by Hyland's category of the *endophoric marker* (2000). NLG research has addressed part of the *what* question for endophoric markers that refer to text or graphics within the same document, namely, how to generate referring expressions employing deictic descriptions; e.g., *the next chapter* or *the above figure* (Paraboni and van Deemter 1999). Other NLG research relevant to the *what* issue includes generation of multimodal referring expressions (McKeown et al. 1992, André and Rist 1994), and generation of captions (Mittal et al.. 1998, Fasciano and Lapalme 1999). Recent NLG research on layout has not addressed the role of multimedia cues in effective layout (Bateman et al. 2001, Power et al. 2003), but clearly, layout needs to be addressed in a complete multimedia cue model.

Experiments

We now present two experiments whose goal was to confirm the premise that providing multimedia cues in on-screen presentations (using intuitively reasonable cue placement) is worthwhile to users, at least for certain types of tasks and screen layout. The first experiment was designed to evaluate the usefulness of multimedia cues for the task of skimming through text "to locate specific information or gain the gist", a reading strategy often used

by web page readers (Dyson and Haselgrove 2001). Subjects were shown presentations in which a paragraph of text was accompanied by two or three figures aligned horizontally below the paragraph. Each presentation screen was followed by a mouse-input multiple-choice test screen. The main hypothesis was that average task time and test score would be superior for subjects shown presentations containing multimedia cues (as shown in Figure 1a) compared to subjects shown presentations in a layout that was the same except not including the multimedia cues. The results of the first experiment were inconclusive although subjects' subjective ratings favored providing multimedia cues.

Next, we designed a second experiment using a different style of layout and tasks. The reason for changing layout was that presenting three figures in horizontal alignment on a 19-inch screen limits the resolution of the images. A more practical design, given current screen constraints, is to put no more than two in horizontal alignment (Figure 1a), or to present each figure on a separate screen accessed by hyperlinks (e.g. Figure 1b). The task was changed because the type of application that our research addressed had changed. In our current application, the multimedia presentation should facilitate comprehension of user-tailored medical information. Thus, the task was changed to reading for comprehension in addition to reading to locate specific information. Since in this application users would be able to take as much time as desired, we measured test score but not task time. (Also, since our intended users would be able to review a presentation as often as wished, we decided it was not necessary to test long-term retention.) Although the results of the second experiment were not statistically significant, there was a practically significant (about 10%) improvement in average test score for the case where figures were presented on separate screens and hyperlinks were integrated with the text as multimedia cues (Figure 1b).

Experiment 1

Experiment Design. The first experiment used a between-group design with subjects, undergraduate volunteers, randomly assigned to one of three groups of 10 subjects each. (Data for four of the subjects was excluded from the final analysis because those subjects were ESL speakers.) All subjects were shown a sequence of four presentations. All groups received the same content, but each group was shown a different version of the layout shown in Figure 1a. In version 1, the text above the figures contained no multimedia cues and each figure was followed by a two to four sentence caption (not shown in Figure 1a). In version 2, the text of the captions was integrated into the block of text above the graphics. Version 3 (shown in Figure 1a) was identical to version 2 except that, for each figure, a multimedia cue was integrated with the text above the figures; the cues were placed at the end of the text that came from each caption. The independent variable was version. The dependent variables were the time to complete the tests and test score. Average time and score were compared between groups. The main hypothesis was that average time and test score would be better for group/version 3 compared to group/version 2. Version 1 was tested for exploratory purposes.

Materials and Procedure. The one-page presentations were constructed by the experimenter by selecting excerpts and related data graphics (line graphs, bar graphs, and pie charts) from four sources representing different authors, genre, topics (airline profits, college enrollment, US teacher salaries, global warming), and audiences. The excerpts were approximately the same word-length (including captions). Except for the first presentation, which was used only for practice and included two graphics, each presentation included three graphics. In constructing the three different versions of each presentation, differences other than layout style described above (e.g., line length, color scheme, font style, and font size) were minimized by the experimenter as much as possible.

Each of the presentation screens was followed by a mouse-input multiple-choice question screen containing four questions. Scrolling was disabled but subjects could use a button to return to the preceding presentation screen as often as necessary to answer the questions about it, which were designed to test how well subjects could get the gist of the presentation and locate specific information. The experiment's instrument was displayed on a desktop PC with a 19-inch color monitor. The instrument was implemented by a computer program written in Javascript and was run by a web browser. The program recorded the subject's answers and times. After subjects finished the experiment, they were given a short paper questionnaire including an ease-of-use question (*How easy was it to get the information you needed to answer the questions?*) with choice of answers on a scale from 1 (*Very Difficult*) to 5 (*Very Easy*).

Results and Discussion. The experiment did not show any significant or practical difference in objective results between any groups, and there was wide variation within each group. A possible explanation is that the predicted effect was obscured due to problems in the instrument or setting: Some subjects reported difficulty in interpreting some of the graphs, due to the screen resolution and/or due to the design of the graph. The setting was a public computer room where noise and other distractions were not controlled. The only positive results were subjective; subjects' responses on the ease-of-use question was dependent on version (chi-squared=7.369578, P-value < 0.005). As shown in Figure 2, the responses to the ease-of-use question tended to be more favorable for version 3 (i.e. with cues) than for version 2 (without cues). (In addition, users preferred version 1, with captions but no cues, to version 2.)

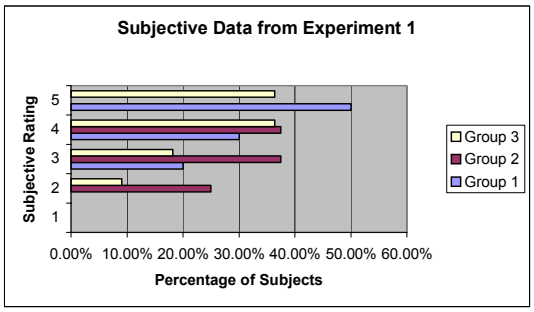

Figure 2: Subjective Ease-of-Use Rating in Experiment 1

Experiment 2

Summary of Changes. As described above, the layout style and task were modified to be relevant to our new application. In addition, we changed the presentation content to address the problems reported with interpreting the graphs in Experiment 1, and moved the experiment to a distraction-free environment. Since we had few responses to an initial advertisement for volunteers this time, we recruited undergraduate subjects by offering $5.00 for participation. Also, ESL speakers who responded to the advertisement were screened out before participating in the experiment.

Experiment Design. This experiment also used a between-group design. Subjects were randomly assigned to one of four groups, with 10 subjects per group. All subjects were shown the same content. Each group was shown the content in a different style of layout determined by the values of two independent binary variables: (1) CUE: presence or not of multimedia cues in the text, and (2) PROXIMITY: whether or not the associated graphics appeared on the same screen as the text. We shall refer to each group by the values of the independent variables for the style presented to the group. Figures 1a and 1b show the (+CUE, +PROX) and (+CUE, -PROX) styles, respectively. In the +CUE condition, a multimedia cue was placed at the end of the first sentence describing a graph (e.g., the sites noted as *1st* in Table 1). Note that in the latter style, (+CUE, -PROX), each multimedia cue in the text functions as a hyperlink (indicated by underlining in Figure 1b) to a screen containing only the graph to which the cue refers. The (-CUE, +PROX) layout was like Figure 1a without the multimedia cues. The (-CUE, -PROX) layout was like Figure 1b except that, instead of providing multimedia cues in the text, hyperlinks at the bottom of the screen (i.e. below the paragraph of text) were provided labeled with the same multimedia cue phrases used in the +CUE versions. The dependent variable was score on a test assessing comprehension and skill in identifying the location of information.

The main hypothesis was that the average test score of the (+CUE, -PROX) group would be better than that of the (-CUE, -PROX) group. In other words, we wished to confirm that, for this type of task and a layout that requires users to navigate to other screens to see related graphics, providing multimedia cues integrated with the text and functioning as hyperlinks would be beneficial to users. We also wished to test the hypothesis (similar to that of Experiment 1) that cues are beneficial even when the graphics are shown on the same screen as the text, i.e., that the (+CUE, +PROX) group would perform better than the (-CUE, +PROX) group. Lastly, based on the Spatial Contiguity Principle, we hypothesized that the (-CUE, +PROX) group would perform better than the (-CUE, -PROX) group and that the (+CUE, +PROX) group would perform better than the (+CUE, -PROX) group.

Materials and Procedure. The subjects were given two presentations each followed by a mouse-input multiple-choice test. Only two presentations were given to make sure that the subjects had as much time as needed for comprehension. The first presentation and test were used for practice only. The presentations were created by the experimenter by modifying the layout of excerpts and related data graphics selected from two different sources. The first presentation was based on an article on college enrollment (also used in Experiment 1). The second presentation consisted of the excerpted text shown in Table 1 (but in paragraph format, and not including the original article's multimedia cue) and the two line graphs from the article's Figure 2, separated into two figures renamed *Figure 1* and *Figure 2* (see Figure 3a and 3b). In constructing the variants for each of the four groups (described in Experiment Design above), care was taken to minimize differences other than the independent variables.

As in Experiment 1, scrolling was disabled but subjects were allowed to navigate between the test screen and the related presentation screen as often as desired. Therefore, it was not possible for a subject to see part of the test and part of the presentation at the same time. Also, subjects in the two -PROX groups were able to navigate between the main text presentation screen and the figure screens as often as they wished. The multiple-choice test's six questions measured comprehension and skill in locating information in the presentation text or graphics. The instrument was implemented by a Javascript program and displayed on a desktop PC with a 19-inch color monitor by a web browser. The program recorded subjects' answers to test questions.

Results. As shown in Table 2, the mean test score for (+CUE, -PROX) was 78.9%, compared to 68.7% for (-CUE, -PROX). The mean test scores for (+CUE, +PROX) and (-CUE, +PROX) were close to the mean score for (+CUE, -PROX). However, a two-sample T-Test showed no statistically significant difference in test scores between (+CUE, -PROX) and (-CUE, -PROX), between (+CUE, +PROX) and (-CUE, +PROX), between (-CUE, +PROX) and (-CUE, -PROX), nor between (+CUE,

+PROX) and (+CUE, -PROX). A two-way ANOVA of test scores found no interaction between PROXIMITY and CUE.

	+CUE	-CUE
+PROX	77.9	76.9
-PROX	78.9	68.7

Table 2. Test scores (% correct) in Experiment 2

Evaluation of Usefulness of Approach

Surprisingly, neither experiment provided strong empirical justification for an intelligent presentation system to generate multimedia cues at all. However, the first experiment suggests at least that the subjects' perception of ease of use was positively influenced by presence of multimedia cues; and the second suggests that multimedia cues may be of practical value in an application such as ours, i.e., where any potential means of increasing a user's comprehension of the health-care document is worthy of further investigation. However, given the cost of obtaining these results, it is not clear that it is practical to continue to pursue this approach to address the more fine-grained questions that motivated us in the first place.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Career Grant No. 0132821, and grants from UNCG: New Faculty Grant (2000), Summer Excellence Faculty Research Awards (2000, 2001). We thank the graduate students who implemented the instruments and collected data: Jennifer Brooks, Allison Fowlkes, William Moates, and Karen Jirak. Also, we are grateful for suggestions on the design of Experiment 2 from John Karat of IBM T.J. Watson Research Center, and for assistance with statistical analysis from Scott Richter of the UNCG Statistical Consulting Center. Figures 3a and 3b appeared as Figure 2 in Leonard Evans, "Traffic Crashes", *American Scientist*, May-June 2002, Vol. 90, 244-253; illustration credit: Barbara Aulicino/American Scientist.

References

André, E., and Rist, T. 1994. Referring to World Objects with Text and Pictures. *COLING-94*, 530-534.

Bateman, J., Kamps, T., Kleinz, J. and Reichenberger, K. 2001. Towards constructive text, diagram, and layout generation for information presentation. *Computational Linguistics* 27(3):409-449.

Carenini, G., and Moore, J. 2000. An Empirical Study of the Influence of Argument Conciseness on Argument Effectiveness. In *Proc. of 38th Annual Meeting of ACL*, 150-7.

Dillon, A. 1992. Reading from paper versus screens: a critical review of the empirical literature. *Ergonomics* 35:1297-1326.

Dyson, M.C., and Haselgrove, M. 2001. The influence of reading speed and line length on the effectiveness of reading from screen. *Int. J. of Human-Computer Studies* 54:585-612.

Di Eugenio, B., Moore, J.D., and Paolucci, M. 1997. Learning Features that Predict Cue Usage, *Proc. of 35th Annual Meeting of ACL, 80-7*.

Fasciano, M. and Lapalme, G. 1999. Intentions in the coordinated generation of graphics and text from tabular data. *Knowledge and Information Systems*.

Hyland, K. 2000. *Disciplinary Discourses*. Pearson Education Ltd.

Mayer, R.E. 2002. *Multimedia Learning*. Cambridge U. Press.

McKeown, K. R., Feiner, S. K., Robin, J., Seligmann, D. D., and Tanenblatt, M. 1992. Generating Cross-References for Multimedia Explanation. In *Proc. of AAAI 1992*, 9-16.

Mittal, V., Moore, J., Carenini, G., and Roth, S. 1998. Describing Complex Charts in Natural Language: A Caption Generation System. *Computational Linguistics* 24(3): 431-467.

Muter, P. 1996. Interface design and optimization of reading of continuous text. In Van Oostendorp, H., and DeMul, S. eds., *Cognitive Aspects of Electronic Text Processing*, 161-180. Norwood, NJ: Ablex.

Paraboni, I. and van Deemter, K. 1999. Issues for the Generation of Document Deixis. In André et al. eds., *Deixis, Demonstration and Deictic Belief in Multimedia Contexts,* 43-48.

Power, R., Scott, D., and Bouayad-Agha, N. 2003. Document Structure. *Computational Linguistics* 29(2):211-260.

Reiter, E., Robertson, R., and Osman, L. 2003. Lessons from a Failure: Generating Tailored Smoking Cessation Letters. *Artificial Intelligence* 144:41-58.

Wright, P. 1999. Writing and Information Design of Healthcare Materials. In Candlin, C.N. and Hyland, K. eds., *Writing: Texts, Processes and Practices*, 85-98. London:Longman.

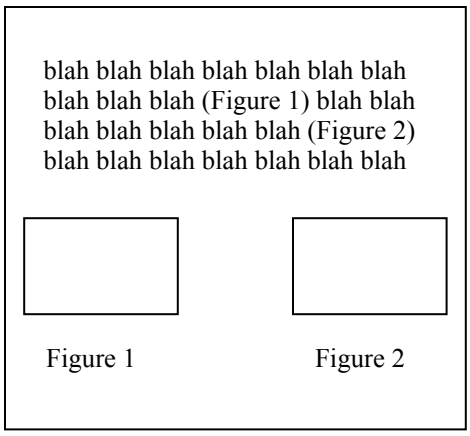

Figure 1a. Layout of screen used for +CUE, +PROXIMITY group in second experiment and similar to layout used for group 3 in first experiment.

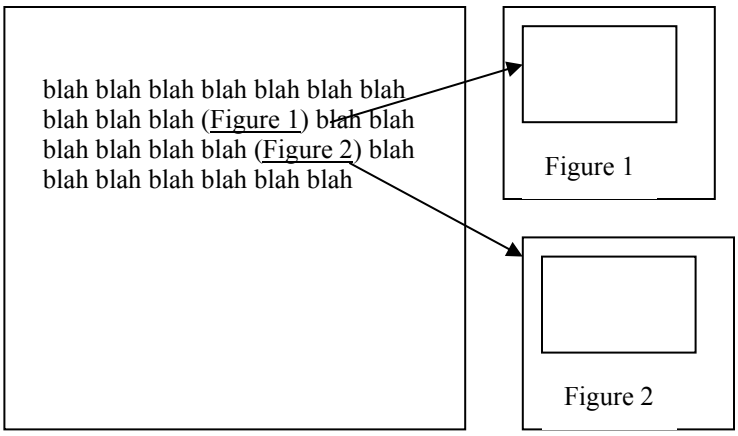

Figure 1b. Layout of three screens used for +CUE, -PROXIMITY group in second experiment. Arrows indicate navigation from screen containing text to screens containing figures.

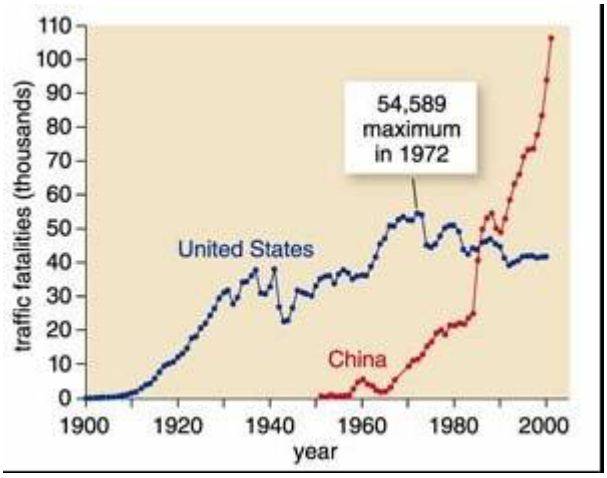

Figure 3a. Figure 1 of Experiment 2. Copyright Sigma Xi, The Scientific Research Society. Reprinted by permission.

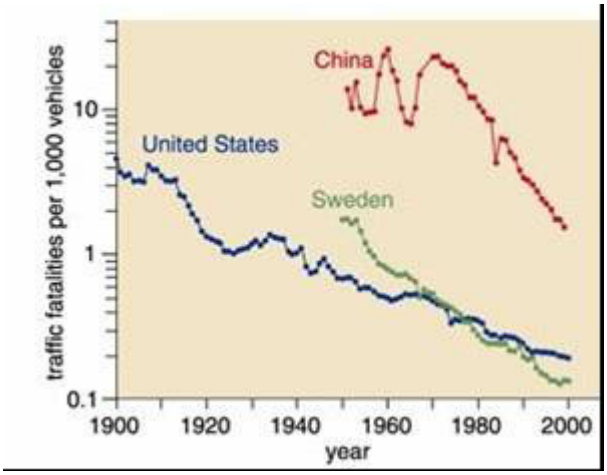

Figure 3b. Figure 2 of Experiment 2. Copyright Sigma Xi, The Scientific Research Society. Reprinted by permission.