

## **Cognitive Fit and an Intelligent Agent for a Word Processor: Should Users Take All That Advice?**

Dennis F. Galletta  
University of Pittsburgh  
galletta@katz.pitt.edu

Alexandra Durcikova  
University of Pittsburgh  
adurciko@katz.pitt.edu

Andrea Everard  
University of Pittsburgh  
aeverard@katz.pitt.edu

Brian Jones  
University of Pittsburgh  
bjones@katz.pitt.edu

### **Abstract**

*While intelligent agents have been developed to provide objective and expert advice to users, most experienced users know that they should not be followed blindly. Cognitive fit theory was developed about ten years ago to support the notion that tools should fit the tasks for which they were designed in light of the user's capabilities. Recently, intelligent agents have been provided to nearly every computer user as part of the Microsoft Office Suite. In nearly all of the applications in the suite, suggestions pop up as the software encounters recognized patterns. Users' capabilities vary widely, however. Some users have noticed anomalies in the advice, and their expertise leads them to override that advice. The computer credibility literature would predict that some users will take that advice without questioning it; this paper asserts that this will occur when there is lack of cognitive fit. In this study, the "Advisor," one particular intelligent agent in Microsoft Word was examined. In this experimental study, 33 undergraduate students were exposed to a passage of text with five repetitions each of three types of error conditions: (1) errors flagged correctly, (2) errors found by the Advisor that were not truly errors, and (3) errors missed by the Advisor. Hypotheses were that (1) the Advisor would in general improve performance, (2) Expertise in English would in general improve performance, and (3) the Advisor would help more those with higher English skills than those with lower English skills. Verbal SAT scores were obtained by permission of the subjects to serve as a measure of English skills. Analysis of the data showed that overall, all three hypotheses were supported in general. The paper also provides more detailed results for each of the error types. The results imply the need for careful use of intelligent agents; agents will not substitute for user expertise and could indeed degrade the performance of non-expert users.*

### **1. Introduction**

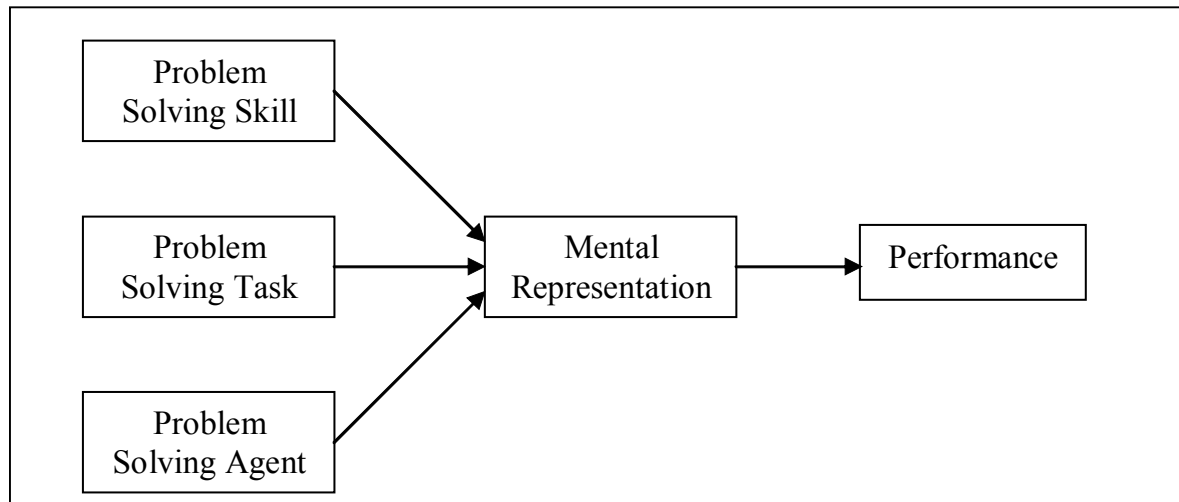
Not long ago, word processing software could barely keep up with users as they typed. As processors moved from 1 MHz devices in 1980 to more than two thousand times that speed today with data paths 8 times larger than before, vendors have never failed to make use of each step in extra speed by providing ever more sophisticated and complex features. Some features are devoted to formatting, providing a preview on screen that emulates the printed page nearly identically, and other features focus on content.

Analysis of content requires more sophistication. Earlier spelling and grammar checking were "batch" operations, required to be run as a separate operation after the document was completed. Today's content-oriented functions run in real-time, and do not even need to be invoked; they start by default and are therefore already likely to be taken for granted by most users just as the formatting functions have been for some time.

The software employs an intelligent agent that consults a set of grammar rules, searching for common usage errors such as the use of fragments, run-on sentences, subject-verb disagreement, passive voice, double words, and split infinitives. Not only must the agent find the individual words in its list, but they must be arranged in conformance to its rules.

While its sophistication is indeed more advanced than ever before, there are two important problems that most writers find with these tools. First, false positives often occur (errors that incorrectly pass the agent's tests). Based on an analysis of the twenty most common grammar errors [1], Kies [9] found that Word 2000 uncovered none of them. The same samples run through Word XP show some improvement, with 6 of the 20 errors found.

The second problem is the opposite case of "false negatives," where some items are flagged as errors when they are not. Often, suggestions are made that would satisfy the agent, but would either distort the true



**Figure 1. Cognitive Fit Model Applied to the Word Agent (Adapted from Vessey & Galletta, 1991)**

meaning or create an error that would not be flagged. For example, the agent analyzes the sentence “Multiple regression was run,” underlines “regression,” and suggests it be changed to “regressions.” If that advice is followed, the word “was” is then underlined and the suggestion is made to change the word to “were.” The final sentence reads “Multiple regressions were run,” which is a distortion of the true meaning of the original sentence.

It is therefore clear that the agent cannot perform all error-checking; users are not “let off the hook” completely by using the spelling and grammar tools. While there indeed needs to be a fit between task and technology [5], there are sometimes difficulties caused by cognitive limitations. The word processing task is one such situation, requiring language expertise of users. Therefore, there sometimes needs to be more than a fit between task and technology alone, but a “cognitive fit” [14] as well. Our model is shown in Figure 1.

This experimental study addresses that fit by asking users to review a letter for spelling and grammar. The agent, called the “Advisor” in this study, was disabled for half of the users, and verbal SAT scores were used to categorize users into language “experts” and “novices.” The rest of this paper provides a review of related literature, hypotheses, our methodology, and our results.

## 2. Prior Research

Two bodies of literature provide guidance in this area. First, the area of Cognitive Fit provides deeper understanding of how task and technology are augmented by a user’s cognitive abilities. The second area is that of

research in Computer Credibility, suggesting that a person might be unduly influenced by advice received from an agent.

### 2.1. Cognitive Fit

The paradigm of Cognitive Fit [14][15] was first applied to the “graphs versus tables” literature. The primary thrust is that there should be consonance among three factors: the user’s cognitive skills, the task, and the representation of the task (as presented to the user). If there is a match, the proper mental representation will be formed, and proper solution of the problem will be possible. In the case of the Advisor, the task representation would be the indicators of necessary refinement, the “wavy” underlines. In theory, a person who intends to produce an error-free document would notice the underlined text, develop an understanding of what is wrong, and fix the error, resulting in clean text. Flaws in grammar-checking software require that a user possess higher levels of verbal skill than the user might believe initially.

The cognitive fit approach has been studied in many different domains, such as multiattribute data [13], verbal geographic instructions [7], spatial maps [2], programming [6], and even accounting [3]. This paper provides an extension into word processing for what we believe is the first time.

### 2.2 Computer Credibility

Computers are sometimes perceived as ‘infallible’, ‘faultless’, ‘awesome thinking machines’ that have

		Advisor	
		On	Off
User familiarity with subject matter	Novice		
	Expert		

**Figure 2. Experimental Design**

‘superior wisdom’. That is, of course, not the case, as discussed above. Computers themselves nearly always perform precisely as instructed, but the problem is that the instructions have flaws because all possible situations have not been anticipated fully.

There have been several streams of research that have looked at an array of issues pertaining to computer credibility. First described by Sheridan et al. [12], others studied how Computer Credibility is gained, lost, and regained [8][10]. The context in which the computer is being used affects computer credibility [11], and individual characteristics of the user can affect this credibility [4].

Credibility is made up of multiple dimensions, such as trustworthiness and expertise. Credibility is often equated with believability [4]; if a computer is deemed believable, it is thought to be credible. Terms such as “accepting the advice,” “trusting,” and “quality of information provided” are seen as conveying computer credibility ([4], p.81).

There are several instances in which the credibility of computers is especially important, such as when computers act as (1) decision aids, (2) knowledge sources, and (3) tutors or instructors [4]. In fact, Fogg & Tseng found that credibility is increased with these uses. Our research considers the computer as a source of knowledge or a decision aid and therefore we need to be concerned with over-reliance on agents such as the Advisor. It is possible that even individuals who are competent in the area in question will still allow the technology to make incorrect decisions for them.

### 2.3 Research Expectations

It is expected that there will be direct effects of expertise and also the Advisor. The most interesting question, however, is how the two factors will interact.

- H1: The Advisor will, in general, affect performance on the task.*
- H2: Expertise will, in general, affect performance on the task.*
- H3 The Advisor will affect experts differently than novices, that is, there will be an interaction effect.*

### 3. Method

In order to test the research question, namely whether intelligent agents designed to support the user can in fact do more harm than good, we employed a 2 x 2 factorial design (see Figure 1 below).

A between-subjects design was used. Participants were asked to edit a business letter using Word under one of two conditions: with the intelligent agent on or off. The outcome measure is the performance on the task, calculated by using a weighted count of errors on the completed document. (a lower score is better). The calculation is described later.

Data for model testing were gathered from multiple random samples of undergraduate business students enrolled in an introductory MIS class in a major University in the Northeastern U.S. The division between “novices” and “experts” was identified through standardized testing scores (SAT scores) obtained from the Dean’s office after receiving permission from the subjects. The scores were dichotomized by splitting along the median into experts and novices. The Advisor was either activated (On) or not (Off). The performance was measured in terms of the four types of mistakes incorporated in the text; there were five instances of each type of mistake. The four types of errors were:

**Error Type 1** – errors flagged correctly: For example, a spelling error.

**Error Type 2** – errors found by the Advisor that were not truly errors: For example, Advisor suggests “multiple regressions” when in fact “multiple regression” is correct.

**Error Type 3** – errors missed by the Advisor: For example, the Advisor does not pick up that “their” should be used instead of “there” given the context.

**Error Type 4** – suggestions given by the Advisor: for example, changing from the passive voice to the active voice.

The main study included 33 volunteers (15 males and 18 females). Materials were refined in a pilot study of 20 students conducted earlier. Subjects were given verbal instructions and 15 minutes to complete the task.

The participants were graded on their performance on each of the four error types. The three more egregious errors (types 1, 2, and 3) were given 2 points each and the suggestion error (type 4) was only assigned one point. The goal of each participant was to find all errors and take appropriate action thereby earning zero points. For purposes of this experiment, the larger the number earned by the participant the poorer the performance. Three graders individually assessed the performance, and

**Table 1. ANOVA For Overall Performance**

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
<b>Main Effects</b>					
Expertise	125.029	1	125.029	7.96	0.0086
Advisor	463.697	1	463.697	29.5	0
<b>Interactions</b>					
Expertise x Advisor	81.387	1	81.387	5.18	0.0304
Residual	455.784	29	15.7167		
Total (corrected)	994.061	32			

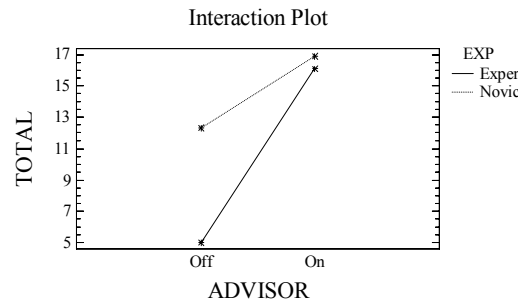
discrepancies between differences in grading were resolved through discussion.

At the end of the experiment, subjects were asked to fill out an exit questionnaire to determine whether the Advisor helped them accomplish the task, i.e., whether they realized that the Advisor had been either turned on or off. The question asked subjects to indicate their agreement with the following statement: "MS Word helped me with my spelling and grammar while accomplishing this task" and was scored with a 5-point Likert scale.

The means of the two groups (Advisor on and Advisor off) were compared to confirm that subjects realized the presence or absence of the Advisor. Subjects with the Advisor on responded to the question with a mean of 3.2 (.88). Subjects with the Advisor off responded to the question with a mean of 1.9 (1.41). A t-test confirmed that the difference between the two groups was significant (p-value = .0011, not assuming equal variances), confirming that subjects with the Advisor on understood that MS Word helped them with their spelling and grammar while accomplishing the task.

#### 4. Results

Fourteen and nineteen of the subjects performed the task with the Advisor off and on, respectively. The median for use in splitting subjects into categories of expert and novice was 570 and the scores ranged from 460 to 720. Seventeen subjects were categorized as novices and sixteen as experts. Analysis of the various demographics associated with each cell revealed no unexpected significant differences among them that could confound the results obtained. Further, tests of normality revealed no significant deviations in the data collected.

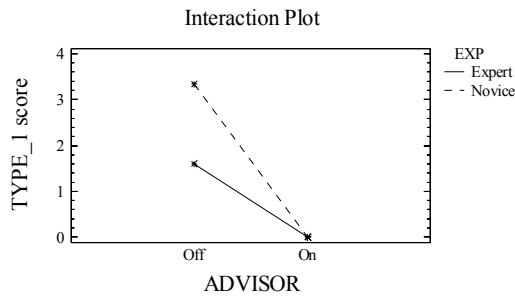


**Figure 3. Interaction Plot for Total Score (All Error Types; Lower Score is Better)**

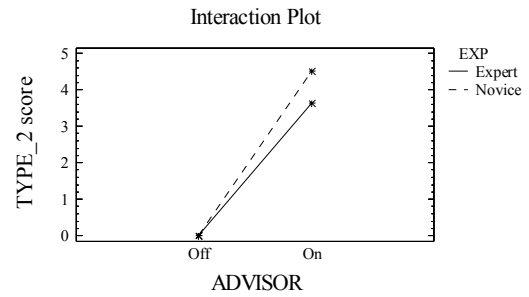
ANOVA was used to test the hypotheses. We tested the hypotheses for (1) the total score that the subjects received and (2) the scores of each type of mistake. Though the results pertaining to the total scores are interesting in and of themselves, deeper understanding of their performance can be obtained by looking at the different types of mistakes on an individual basis, as the types of mistakes are qualitatively different.

Table 1 shows the ANOVA for the Total scores. Because the three p-values are less than .05, these factors have a statistically significant effect on the total score at the 95% confidence level.

Hypothesis H1 predicts that the Advisor would, in general, affect performance. Subjects with the Advisor on performed less well than those with the Advisor off (p-value = .0086). Hypothesis H2 predicts that Expertise would, in general, affect performance. Experts performed better in general than novices (p-value = .0000). Hypothesis H3 predicts that the Advisor would affect experts differently than novices. There is support for H3 at the 95% confidence level (p-value = .0304). Although Table 1 shows support for the hypotheses, the interaction plot in Figure 3 provides pictorial details of the interaction.



**Figure 4. Interaction Plot for Error Type 1: Errors Flagged Correctly (Lower Score is Better)**



**Figure 5. Interaction Plot for Error Type 2: Errors Flagged Incorrectly by the Advisor (Lower Score is Better)**

As Figure 3 suggests, there is an interaction effect between the Advisor and Expertise. Both novices and experts performed less well with the Advisor on than with the Advisor off, and overall novices performed less well than experts. However, the best performance was shown for experts with the Advisor off. Interestingly, turning the Advisor on seems to make experts perform just like novices.

An explanation of these performance differences could be obtained by examining performance scores for each error type, described below.

**4.1. Error type 1: Errors flagged correctly (if the Advisor is on)**

The first error type, errors that were flagged correctly by the Advisor, shows rather mixed results. Table 2 shows that both hypotheses H1 and H3 were supported at the 90% confidence level, and that hypothesis H2 was supported at the 95% confidence level.

From Figure 4 we observe that there was no difference in performance between experts and novices when the Advisor was on and detected actual errors. Both groups corrected all spelling errors flagged. However, as expected when the Advisor was off, experts performed better than novices due to their higher English skills.

**4.2. Error type 2: Errors found by the Advisor (when on) that were not truly errors**

Table 3 shows that hypotheses H1 was supported at the 99% confidence level for errors that were incorrectly flagged by Word.

From Figure 5 we can see that there is no statistically significant difference in performance for error type 2 between experts and novices when the Advisor was either on or off. Both groups tended to accept the advice of the Advisor, regardless of their expertise level. In general, when the Advisor was on, performance decreased; this is represented by higher “error” scores. That is, both types

**Table 2. ANOVA for Type 1 Errors (Real Errors Flagged by the Advisor)**

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
<b>Main Effects</b>					
Expertise	5.70081	1	5.70081	3.23	0.0828
Advisor	46.18	1	46.18	26.16	0
<b>Interactions</b>					
Expertise x Advisor	5.70081	1	5.70081	3.23	0.0828
Residual	51.2	29	1.76552		
Total (corrected)	120.242	32			

**Table 3. ANOVA for Type 2 Errors (Flagged Incorrectly by the Advisor)**

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
<b>Main Effects</b>					
Expertise	1.41525	1	1.41525	0.31	0.5822
Advisor	125.613	1	125.613	27.48	0
<b>Interactions</b>					
Expertise x Advisor	1.41525	1	1.41525	0.31	0.5822
Residual	132.545	29	4.57053		
Total (corrected)	264.97	32			

**Table 4. ANOVA for Errors Missed by the Advisor When On**

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
<b>Main Effects</b>					
Expertise	42.3937	1	42.3937	16.05	0.0004
Advisor	54.7828	1	54.7828	20.74	0.0001
<b>Interactions</b>					
Expertise x Advisor	53.6558	1	53.6558	20.32	0.0001
Residual	76.5828	29	2.64079		
Total (corrected)	190.061	32			

of users “fall for” most of what is really incorrect advice. When the Advisor was off, not a single user flagged any of the items as errors. Computer credibility indeed seems alive and well.

#### **4.3 Error type 3: Errors missed by the Advisor (when on)**

Table 4 shows that hypotheses H1, H2 and H3 were supported at the 99% confidence level.

From Figure 6 we can see the strong interaction between the Advisor and Expertise. The best performance was achieved by Experts with the Advisor off. In this case, the subjects relied only on their knowledge, and had no help from the Advisor. On the other hand, experts with the Advisor on actually

performed exactly like Novices with either the Advisor on or off.

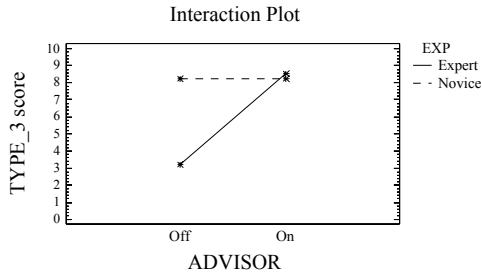
#### **4.4 Error type 4: Suggestions given by the Advisor**

Table 5 shows that hypothesis H1 was supported at the 99% confidence level.

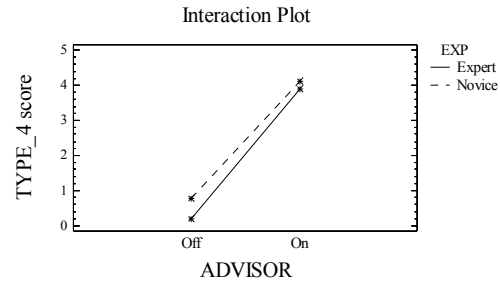
Figure 7 illustrates the striking and significant difference in performance between subjects who performed the task with the Advisor on and those who performed the task with the Advisor off.

### **5. Discussion and Conclusions**

From this study it is apparent that the hypotheses were supported when looking at overall performance. Both



**Figure 6. Interaction Plot for Error Type 3: Errors Missed by the Advisor when on (Lower is Better)**



**Figure 7. Interaction Plot for Suggestions Given by the Advisor**

expertise and the use of a spelling and grammar-checking agent (Advisor) affected the users' performance. Also, the best performance was obtained from experts who had the Advisor turned off. Interestingly, turning on the Advisor made our English language experts perform just like novices, rather than making our novices perform more like experts.

A more detailed look at a variety of error types provides more understanding of this result. When real errors are flagged correctly, novices and experts are about equal at eliminating errors. If the advisor is turned off, novices are disadvantaged but experts are not.

Another interesting finding is that when "false negatives" (non-errors flagged as errors) are considered, it appears that the Advisor indeed impairs performance. When more subtle stylistic suggestions were made by the Advisor, performance improved for both novices and experts. Even the high SAT subjects heeded this kind of assistance.

Perhaps the most interesting finding is that when errors are missed by the Advisor, performance is impaired for both novices and experts. Our speculation is

that experts tend to be less careful when the Advisor is on, and assume that their text has been checked carefully for them. Users of the Advisor seem to attribute greater power than it really has; they are lulled into a false sense of security.

Future researchers might consider the impact of training and perhaps warning messages for subjects when such an agent is used. Perhaps when levels of confidence are more realistic, and people in general adjust their reliance on software, such agents can have more uniformly beneficial results. This study is one step toward understanding the dynamics of confidence, expertise, and cognitive fit in a document editing task.

## 6. References

- [1] Connors, R.J. and Lunsford, A.A. "Frequency of Formal Errors in Current College Writing, or Ma and Pa Kettle Do Research," *The St. Martin's Guide to Teaching Writing* 2nd ed. Ed. Robert Connors and Cheryl Glenn. New York: St. Martin's, 1992, 398
- [2] Dennis, A., and Carte, T. "Using Geographical Information Systems for decision making: Extending cognitive

**Table 5 – ANOVA for Suggestions Given By the Advisor**

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
<b>Main Effects</b>					
Expertise	1.19528	1	1.19528	1.15	0.2924
Advisor	94.4775	1	94.4775	90.91	0
<b>Interactions</b>					
Expertise x Advisor	0.24847	1	0.24847	0.24	0.6286
Residual	30.1396	29	1.0393		
Total (corrected)	126.182	32			

fit theory to map-based presentations," *Information Systems Research*, June 1998.

[3] Dunn, C., and Grabski, S. "An Investigation of Localization as an Element of Cognitive Fit in Accounting Model Representations," *Decision Sciences*, 32 (1) 55-94.

[4] Fogg, BJ and Hsiang Tseng (1999) "The elements of computer credibility," CHI 99 Conference Proceedings, Pittsburgh, PA, 80-87.

[5] Goodhue, D. L. (1992) "User evaluations of MIS success: What are we really measuring?", in Nunamaker, J. F., & Sprague, R. H. Jr. (Eds.), *Proceedings of the Hawaii International Conference on System Sciences: Vol. IV - Information Systems*, pp. 303-314.

[6] Hayden, M.K., Olfman, L., Gray, P., and Ahituv, N. (1997) "An Experimental Investigation of Visual Enhancements for Programming Environments," *Journal of Information Systems*, Fall 1997, pp. 19-26.

[7] Hubona, G.S., Everett, S., Marsh, E., Wauchope, K. (1998) "Mental Representations of Spatial Language," *International Journal of Human-Computer Studies*, 48, 705-728.

[8] Kantowitz, BH, Hanowski, RJ, and Kantowitz SC (1997) "Driver acceptance of unreliable traffic information in familiar and unfamiliar settings," *Human Factors*, 39(2) 164-176.

[9] Kies, D. (2002). "Evaluating Grammar Checkers" in *Modern English Grammar* (a hypertext book available at [http://papyr.com/hypertextbooks/engl\\_126/gramchek.htm](http://papyr.com/hypertextbooks/engl_126/gramchek.htm) )

[10] Muir, BM and Moray, N (1996) "Trust in Automation: Part II. Experimental studies of trust and human intervention in a process control simulation," *Ergonomics*, 39(3), 429-460.

[11] Pancer, SM, George, M and Gebotys, RJ (1992) "Understanding and predicting attitudes toward computers," *Computers in Human Behavior* 8, 211-222.

[12] Sheridan, TB, Vamos, T and Aida, S (1983) "Adapting automation to man, culture and society," *Automatica*, 19(6), 605-612.

[13] Umanath, N.S. and I. Vessey, "Multiattribute Data Presentation and Human Judgment: A Cognitive Fit Perspective." *Decision Sciences*, Vol. 25, Sept./Dec. 1994, 795-824.

[14] Vessey, I. and Galletta, D.F. (1991) "Cognitive Fit: An Empirical Study of Information Acquisition," *Information Systems Research*, V2, N1, pp. 63-84.

[15] Vessey, I. (1991) "Cognitive Fit: A Theory-Based Analysis of the Graphs versus Tables Literature," *Decision Sciences*, 22 (Spring), 219-241.