Cross Evaluation—A Pilot Application of A New Evaluation Mechanism

Ying Sun and Paul Kantor
Rutgers University. 4 Huntington St. New Brunswick, NJ 08903.
Email: {ysun, kantor}@scils.rutgers.edu

Tomek Strzalkowski
SUNY Albany, Western Avenue, Albany, NY 12222. Email: tomek@csc.albany.edu

Robert Rittman and Nina Wacholder
Rutgers University. 4 Huntington St. New Brunswick, NJ 08903.
Email: {rritt, nina}@scils.rutgers.edu.

The work reports some initial success in extending the Rutgers Paradigm of IR evaluation to the realm of concrete measurement, not in information retrieval per se, but in the arguably more complex domain of Question Answering. Crucial to the paradigm are two components: cross evaluation, and an analytical model that controls for the potential problems of cross evaluation. We describe the experimental design and analytical models. In the models, interaction effects are examined and found not to be important. After eliminating the interaction effects, we are able to extract meaningful and useful results from a very small study involving just three analysts, five topics, and two "systems".

Introduction

In the evaluation of information retrieval systems, there are two paradigms that deserve our attention. The first, which might be called the Cranford Paradigm, emphasizes controlled experimentation in a laboratory type setting. This has had very great impact leading to the established standards at the Text Retrieval Conferences (TREC) (Voorhees & Harmon, 2000), and the near-universal homage paid to the concepts of precision and recall. The Cranford Paradigm has perpetuated itself in important conferences on other topics, such as the Cross Language Evaluation Forum (CLEF) which is sponsored by the European Commission (Smeaton & Harmon, 1997), and the NII-NACSIS Test Collection for IR Systems (NTCIR) project, a workshop on Asian language text retrieval, question answering and text summarization (NTCIR, 2004). In fact, the notions of precision, recall, and additional (in some sense more fundamental) notions of utility have spread from these conferences to evaluation in other fields such as natural language processing and speech processing.

The second approach, the Rutgers Paradigm, has had a different path. It has primarily been seen in research-oriented papers, where the method of evaluation is itself the subject of the research. This began with studies by Saracevic and Kantor in the 1980s (Saracevic & Kantor, 1988, 1988a, 1988b). The underlying principle of the Rutgers Paradigm could be called the "three realities". This refers to the insistence on doing studies that involve "real users, real problems, and real systems". The compelling logic of this approach has made it attractive. At the same time, it contains within itself, some potential for its own destruction.

The destructive potential of the Rutgers Paradigm lies in the fact that while "real systems" are something that we know how to deal with, particularly in the world of information retrieval research, where the systems are usually studied by those who have built them, the other two components: users and problems, are open-ended. The study of users can carry us into the study of cognitive science, the study of personal psychology, the study of educational theory, the study of socially constructed behavior, and so-forth. These represent an "attractive nuisance" to the working scholar.

The study of "real problems" also poses enormous potential hazards. Problems, like questions, can reflect the entire range of human knowledge or ignorance. The classification of problems or questions then becomes no simpler than the long-standing problems of classifying all of knowledge. These have occupied librarians for generations, and anyone familiar with the available solutions knows that they are driven more by the pressures of practical reality than by any compelling logic. Thus, people who have taken up the issue of "classifying problems" find themselves looking at the distressingly large and elusive notions of "task", and "context". To make matters worse (or better, from the perspective of the aspiring academic) the notion of context easily works its way back to connect to all of complexities related to the notion of the "user".

As a result, the typical discussion about developing rigorous measures, which takes into account these three realities, quickly degenerates into speculation, counter speculation, and a despairing conclusion that the interactions are too complex and that the problem would exhaust all reasonable resources. This conversation of despair has played out for a number of years in the TREC interactive track (Over, 2001), which has struggled to pin down some of the

proposed concepts with, it could be argued, relatively little success.

Given all of this it may be surprising that in the present note we report some initial success in extending the Rutgers Paradigm to the realm of concrete measurement, not in information retrieval per se, but in the arguably more complex domain of Question Answering. The paradigm reported here has been developed for the study of collaborative information seeking (Kantor, Sun & Rittman, 2002), but is here reported publicly for the first time in an application to a study of the HITIQA Interactive Question Answering System developed at SUNY- Albany, with sponsorship from the intelligence community's Advanced Research and Development Activity (ARDA).

The philosophical essence of the model reported here is to use realistic (not quite real, in the sense that they were not in fact owned and generated by the users involved) problems, which imply a context and a time limitation. These were coupled with realistic users (that is, users drawn from the community of users for whom the system is intended to serve). In addition, the system itself, the HITIQA system, is a realistic system which is under development. In fact, the experiments described here are a part of the process of improving the system so that it can be moved into the real setting. Since we have a single system, but we want to establish a paradigm for the comparison of systems, we report here on two separate sessions involving use of the system, between which some changes, which were hoped to be improvements, were made. Thus the variable described in this paper as the "workshop" is a surrogate for changes in the system.

We have to point out that this variable is also the surrogate for two other factors which are inextricably confounded with the system changes in the pilot study reported here. First, with the passage of time the participants may well have reflected on their interaction with the system and changed the ways in which they seek to use it. Second, a change in the instructions to participants which made them aware of the Cross Evaluation Paradigm to be described below and may well have influenced their behavior. These however are conceptually unimportant artifacts for the main point of this paper, which is that it is possible to design experiments and analytical methods which extract the effect of differences between two instances of the real users on a set of problems. In other applications, these differences will be differences between systems. In this particular instance they also involve changes in the presentation of the task and the passage of a month of elapsed time.

Crucial to the paradigm described here are two components: cross evaluation, and an analytical model that controls for the potential problems of cross evaluation. In cross evaluation we ask the same individuals who have conducted the sessions using the system to provide judgments of the results of those sessions. We do this using a special purpose tool built by the first author in connection with the Antworld studies cited above. This tool (in its full version it actually permits comparison of the work of teams as well as comparison of the work of individuals) makes it very easy for all of the individuals to judge cited sources (a capability not used in the research reported here) and to judge the finished products of the analytical task. Details of the system and its workings are given in section 2 below.

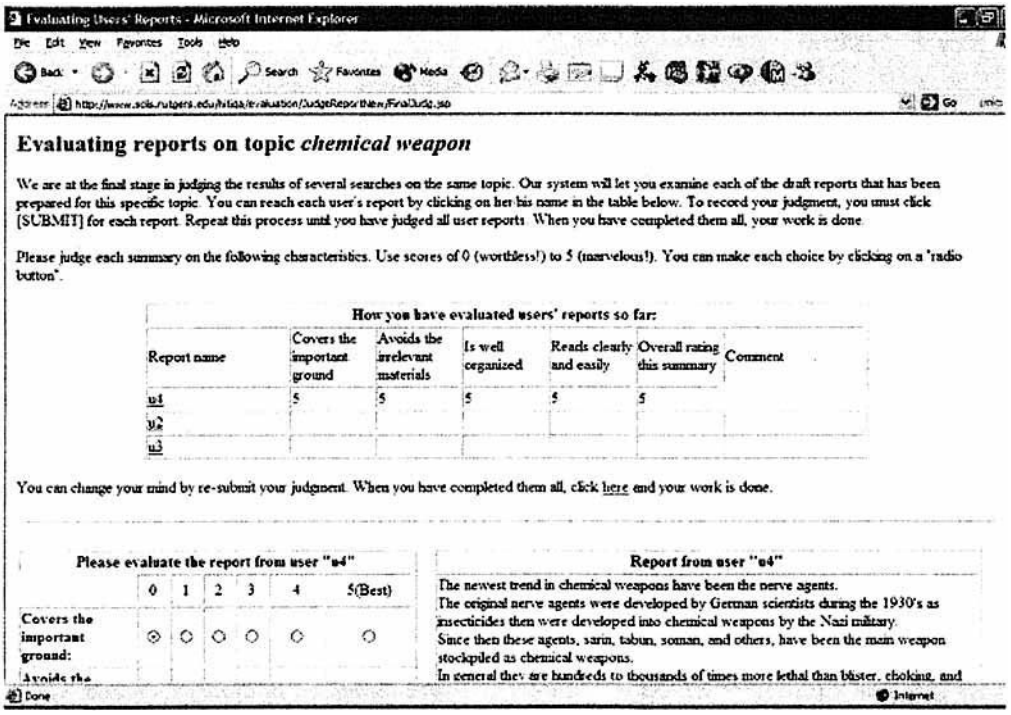
The other key feature of the paradigm is an analytical model based on the analysis of variance. In this particular study we selected five scale-based measures of the quality of an analytical report. Thus for each report we have five measures of its quality. In our analysis we identify these measures as having been (1) produced by a particular judge for (2) the work of a particular participant, on (3) a particular topic, and (4) using a particular instance of the system. As described below, we introduced four parameters representing the main effects of these variables, which are treated as fixed effect factors shifting the mean of the scores for the cases to which they apply. We also introduced a self-evaluation factor, which takes the value one if an analyst is judging his or her own work and takes the value zero otherwise. This permits us to control for the expected effects of bias in favor of ones own work.

Cross Evaluation

In the first part of the simulation task, all of the participants use the experimental system to gather information on a particular topic. Specifically, each analyst was asked to select among the information pieces collected and organize them into a report.

Then, for each topic, we asked all the users who worked on the topic to score all of the reports prepared by each of the users (including herself/himself) as a whole. For this purpose we developed a tool which provides an evaluation form, and also displays each report. The tool is written in JSP and the collected information is stored in a MySQL database. Figure 1 (a, b) shows the interface of the tool.

The interface is divided into two parts. The upper part of the interface (Figure 1a) shows the topic the user is working on, followed by instructions of the task. The table following the instructions shows the user's judgment of all the reports on the topic so far. For example, we can see in Figure 1a that there are three reports on the topic "chemical weapon". They are named u2, u3 and u4 (which reflect the authors of the reports). The user has judged u4's report and the scores are shown in the table. The user has not made judgments on u2 and u3's reports yet. The name of each report is a hyperlink. Users can view reports by clicking on them. The corresponding report will show up in the right hand side of the lower part of the interface. The left hand side of the lower part is the evaluation form.



a. upper part.

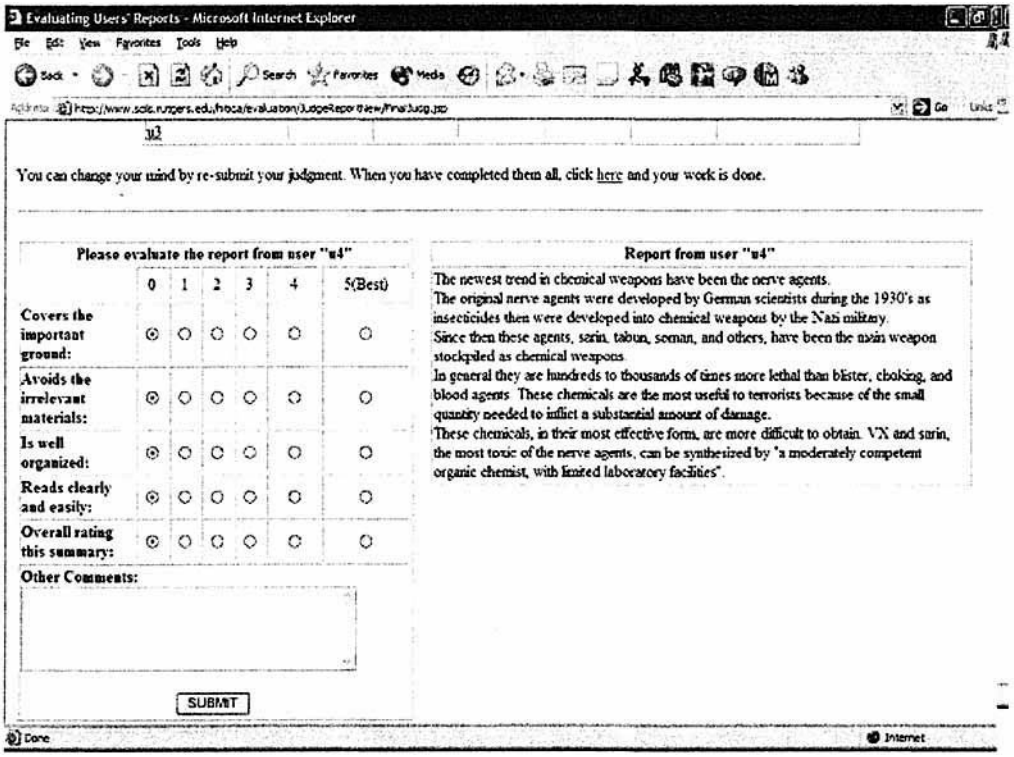


Figure 1. Cross-evaluation Interface

Users are instructed to judge each report according to 5 characteristics:

Covers the important ground

Avoids the irrelevant materials

Is well organized

Reads clearly and easily

Overall rating this report

Each characteristic is scored in a range from 0 (worthless!) to 5 (marvelous!)

The Experimental Design and Analytical Model

Experimental Design

The first evaluation was conducted in September 2003. Four analysts participated. They used the HITIQA system to collection information on four Scenarios and were instructed to use the information to prepare brief reports. The four scenarios include two long scenarios:

- South Africa's WMD Program
- al-Qaida Terrorist

and two short scenarios:

- Nuclear Arms Relationship: Russia and Iraq
- Chemical Weapon: sarin

In table 1, the Sarin scenario is shown as an example.

Table 1. Scenario Example

Short Scenario #2: Chemical Weapon: sarin
The Department of Homeland Security has requested a complete report on the chemical weapon, sarin. This report is due in 5 hours. In your report, include its potency and potential impact on a community, what countries and organizations have been involved in producing it, where these locations are, the production method and how it has developed, who possesses it now, who distributed it (if through trade, what was traded for it?), potential means of use, how can this be integrated into warheads, any known defenses against it, and who is at the greatest threat. Provide any other information that you see relevant.

The analysts have 4 hours to use the HITIQA system and prepare report for a long scenario. For a short scenario, they have one hour.

Based on the experience and comments made by the analysts, we made some changes to the HITIQA system. At the end of October, the analysts were invited to come back and try the system again. Only three of the analysts could attend. A new long scenario on North Korea's nuclear weapons program was added. All three analysts worked on the new scenario. In addition, they worked on scenarios they had not try in September. In the end, each of the three analysts completed five scenarios (between workshop I and workshop II). Since there were five analysts, this means they produced a total of 15 reports (those reports from each analyst on five scenarios) (see Table 2). All reports were judged at the second workshop.

Table 2 analysts by Scenario and treatment.

Scenario	Analysts		Total number of reports
	Phase I	Phase II	
Sarin	u3, u4	u2	3
South Africa	u4	u2, u3	3
Russia/Iraq	u2	u3, u4	3
Al-Quada	u2, u3	u4	3
North Korea		u2, u3, u4	3

Analytical Model

As discussed in the first section, we want to explore the main effects of five independent variables;

Workshop difference. The goal of the evaluation mechanism is to compare systems. We only have one system, so we compared two versions of the same system. Because two versions of the HITIQA system were tested at different times with the same group of analysts, a learning effect may cause the analysts to perform differently at the second workshop. Also, differences between the instructions given in the two workshops may cause the measured changes in performance. Analysts were not told that the reports would be evaluated at the first workshop. We did observe that at the second workshop, the analysts spent more time on preparing the reports compared to the first workshop. We understand that we cannot measure the difference between two versions of the HITIQA system directly. So we call the first factor Workshop Difference instead of System difference.

Self-judgment. A key feature of the experiment design is that participants evaluate each other's work, including her/his own. The hypothesis is that they may give their own reports higher scores.

Author difference. We also realized that there are individual difference among analysts. A highly scored report may not relate to a better system. An analyst's reports may be always good no matter which system is used.

Scenario. Some scenarios may be easier than others.

Judge difference. Some judges may tend to give high scores. All judging work was done at the second workshop, no matter when the reports were created. There is no learning effect or time difference effect on the judges.

The independent variables in the current study are author, judge, scenario, self-judgment, and workshop. First, we examine the 2nd order analytical models. There are 10 such models, depending on which of the two-way effects we include. We examined them one at a time, as described below.

$$s_c(a,j,s,self,w) = \lambda^0 + \lambda^a + \lambda^j + \lambda^s + \lambda^{self} + \lambda^w + \lambda^{f_1,f_2} + e$$

where:

s_c: a report's score on characteristic c,

a: author of the report,

j: who made the judgment,

s: scenario of the report,

self: if this is a self-judgment,

w: the report is created at which workshop,

λ^i: coordinates of factor i, determines the contribution of the independent variable i,

e: random error.

$$f_1, f_2 \in \{a, j, s, self, w\}$$
$$f_1 \neq f_2$$

Equation 1. 2nd Order Analytical Model

We expect to obtain a model as Equation2, in which no interaction effects.

$$s_c(a,j,s,self,w) = \lambda^0 + \lambda^a + \lambda^j + \lambda^s + \lambda^{self} + \lambda^w + e$$

Equation 2. 1st Order Analytical Model.

Results and Analysis

Cross Tabulation Results

The overall results are summarized in the following table. Each block (with 9 cells in it) represents the analysts' judgments of reports on one scenario on one criterion. The column is the author of the report, while the row is the analyst who made the judgments. So, each cell is a ***Score$_c$(j,a,s)***: the score assigned to ***a***'s report on ***s*** scenario and on ***c*** criterion by analyst ***j***.

Due to a miscommunication, we did not receive one analyst's judgments on the Russia/Iraq scenario at all.

Table 3. Cross tabulation of judgment results

Criteria / **Authors** / **Scenario** / **Judges**		Covers the important ground			Avoids the irrelevant materials			Is well organized			Reads clearly and easily			Overall rating		
		U2	U3	U4	U2	U3	U4	U2	U3	U4	U2	U3	U4	U2	U3	U4
North Korea	U2	5	4	3	4	3	1	3	5	3	4	3	3	4	4	3
	U3	4	3	4	4	3	3	4	2	2	4	1	3	4	2	3
	U4	4	4	4	4	4	4	5	4	3	4	5	3	4	5	3
Sarin	U2	5	4	3	4	4	0	4	4	1	4	4	1	4	4	1
	U3	5	4	3	2	2	4	4	4	2	4	4	2	4	4	2
	U4	5	4	3	5	4	1	5	3	1	5	4	1	5	4	2
Al-Quiada	U2	4	4	4	4	3	3	4	4	4	3	4	3	4	4	4
	U3	4	3	4	3	2	2	4	3	3	4	3	4	4	3	4
	U4	4	4	5	4	4	4	3	4	5	3	4	5	3	4	5
Russia/ Iraq	U2	2	3	3	2	2	3	2	3	3	3	3	3	2	3	3
	U3															
	U4	3	3	4	4	4	3	3	4	4	4	4	4	4	4	4
South Africa	U2	4	4	1	5	3	1	4	4	1	4	4	1	4	4	1
	U3	4	4	1	4	3	0	5	4	1	3	3	1	4	4	1
	U4	5	5	2	5	5	1	5	5	1	5	5	1	5	5	1

Difference Between Workshop I and Workshop II

As the final goal of the study is to evaluate the HITIQA system, we would like to see if there are significant differences between the reports created at the first and second workshops. There are 17 cases from workshop I. And there are total 25 cases from workshop II.

A one-way analysis of variance (one-way ANOVA) was conducted to evaluate whether the means of the workshop I reports' quality are significantly different from those of the workshop II reports on five criteria. The results are summarized in Table 4. The difference between workshop I and workshop II was significant on all five criteria.

The results are good, but we must be careful about drawing conclusions because there are four other factors involved that may affect the dependent variables.

Table 4. Results of the ANOVA analysis.

Judgment Criteria	Means Workshop I	Means Workshop II	Differences	Significance
Covers the important ground	3.17	4.05	.88	.01*
Avoids the irrelevant materials	2.67	3.62	.95	.04**
Is well organized	2.58	3.86	1.27	.00*
Reads clearly and easily	2.75	3.76	1.01	.02**
Overall rating	2.83	3.90	1.07	.01*

* Statistically significant at .99 level.

** Statistically significant at .95 level.

GLM Results

The General Linear Model (GLM) allows us to include an enormous amount of information and to fully explain the variation in the dependent variables. As all five independent variables in our study are categorical variables, GLM is used to examine their individual as well as interaction effects on report judgment scores. The report quality is measured by five criteria: Coverage, Avoiding Irrelevant Materials, Organization, Clearness, and Overall.

Interaction effects. As usual in postulating a model with main effects we are obliged to establish that there are not very strong interaction effects which might lead to error in interpretation of the main effects. We have examined all the interaction effects and we find almost no evidence for strong interactions. We examined the 10 two-way interactions among the main effects. We tested 10 analytical models described by Equation 1. Table 5 shows each two-way interaction effect's F value and P value in the corresponding analytical models. We found that none of the two-way interactions have significant effect on the dependent group (five report quality criteria).

Table 5 F and P values of interaction effects in each model

Two-way interaction	**F**	**P**
Author * Workshop	1.80	.08
Author * self judgment	1.53	.16
Author * Scenario	1.02	.45
Judge * self-judgment	1.53	.16
Judge * Scenario	1.09	.37
Judge * Workshop	1.55	.15
Self-judgment * Scenario	.86	.64
Self-judgment * workshop	1.13	.37
Scenario * Workshop	1.21	.29
Judge * Author	1.03	.44

Main effects. Having eliminated interaction, we can proceed to report and discuss the main effects. For each dependent variable, we test the analytical model in the form of Equation 2.

For each dependent variable, we calculate the predicted value using the obtained linear model, and we used a 3X3 scatter plot matrix to examine the correlations among the observed values, predicted values and standardized residuals. The standardized residual is the standardized error of the predicted values compared with the observed values. Figure 2 is the plot of the Coverage criterion. For "observed" by "predicted", we are glad to see a clear pattern, which indicates the model can predict the report's Coverage quality pretty well. We will discuss the details later. For the plots involving standardized residuals, both observed by residual and predicted by residual, the cases are scattered all over the area, there is not an obvious pattern, which indicates that the random errors are independent. The other four criteria have similar patterns.

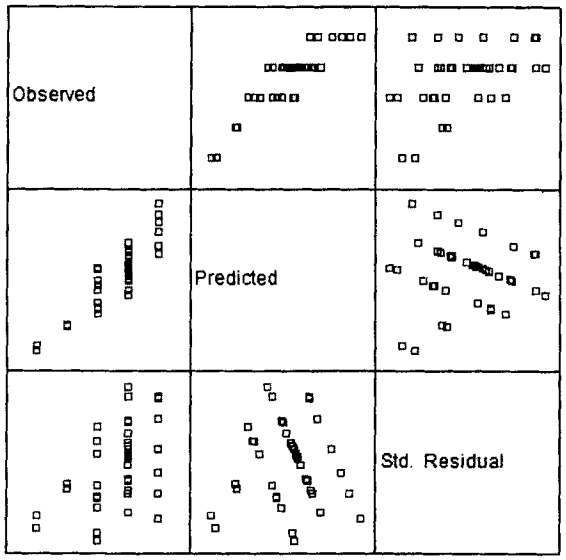


Figure 2. Residual Plot of the Model for the Coverage Criterion

The "multivariate tests" result of the GML simultaneously tests each independent variable's (author, judge, scenario, workshop and self-judgment) effect on the dependent group (five report quality criteria). The results show that three of the five factors have a significant effect on report quality: author difference with $F = 3.42$, $p < .01$; Workshop difference with $F = 12.15$, $p = .00$; Scenario with $F = 2.54$, $p < .01$. There is no significant difference between reports judged by authors themselves or by other analysts. The difference among different judges is not significant.

Table 6 shows each model's R^2. The results show that our models fit the data pretty well. For the Coverage criterion, the R^2 is .79.

Table 6 Five linear models R^2

Models	R^2
Covers the important ground	.79
Avoids the irrelevant materials	.44
Is well organized	.63
Reads clearly and easily	.52
Overall rating this report	.65

The summary of each independent variable's effect on each dependent variable is in Table 7. The results show that different authors' reports have significantly different judgment on all of the five report evaluation criteria. As suggested by our first naïve analysis, reports written at Workshop II are significantly different from those written at workshop I on all five criteria. It is interesting to see that Self-judgment has no significant effect on any of the five criteria.

We are also interested in examining the relationship and significance of the relationship among values of each independent variable. During the two evaluation work shops, we also collected a great amount of other information. We believe that the results may correspond to some phenomenon we observed.

Table 7 Effects on each dependent variable

Source	Dependent Variable	F	Sig.
WORK-SHOP	COVER	57.71	.00*
	NOIRREL	8.67	.01*
	ORGANZ	35.60	.00*
	CLEAR	19.00	.00*
	OVERALL	37.78	.00*
AUTHOR	COVER	11.25	.00*
	NOIRREL	10.02	.00*
	ORGANZ	15.50	.00*
	CLEAR	9.84	.00*
	OVERALL	14.60	.00*
JUDGE	COVER	3.93	.03**
	NOIRREL	5.15	.01*
	ORGANZ	2.09	.14
	CLEAR	3.69	.04**
	OVERALL	4.24	.02**
SELF	COVER	.36	.55
	NOIRREL	.06	.80
	ORGANZ	1.06	.31
	CLEAR	1.32	.26
	OVERALL	.72	.40
SCENE	COVER	12.75	.00*
	NOIRREL	.66	.63
	ORGANZ	3.96	.01*
	CLEAR	2.59	.06
	OVERALL	4.87	.00*

* Statistically significant at .99 level.

** Statistically significant at .95 level.

- Workshop difference effects

Since there are only two values of Workshop variable: workshop I and workshop II. From table 7 we learned that the differences are significant on all five dependent variables. By examining the estimated parameters, we find that the reports written at Workshop II have significantly higher scores than those from Workshop I on all five criteria. Table 8 summarizes the magnitudes of the Workshop effects on five dependent variables.

Table 8 Composite summary of the magnitudes of the Workshop Main Effect

Dependent variables	**Magnitude of the Workshop Effect (II vs. I)**
Covers the important ground	1.45
Avoids the irrelevant materials	1.08
Is well organized	1.65
Reads clearly and easily	1.33
Overall rating this report	1.53

- Author effects

We have three analysts. User "u3" performs best on the Overall criterion. User "u2" performs best on the other four criteria. And both "u2" and "u3" perform significantly better than "u4" on all five criteria. The differences between "u2" and "u3" are not large. The results are in Table 9.

Table 9 Composite summary of the magnitudes of the Author Main Effect

Dependent variables	**Magnitude of the Author Effect**	
	u2 vs. u4	**u3 vs. u4**
Covers the important ground	.90	.65
Avoids the irrelevant materials	1.64	1.14
Is well organized	1.38	1.34
Reads clearly and easily	1.26	1.12
Overall rating this report	1.18	1.20

- Self-judgment effects

Self-judgment has no significant effect on all five dependent variables. In other words, reports judged by the authors themselves do not have significantly different scores from those judged by others. The magnitudes of the effect are in Table 10. However it is interesting to notice that the analysts gave their own reports lower scores on four criteria: Avoiding irrelevant information, Organization, Clearness and Overall. They only judge their own reports better on the Coverage criterion.

Table 10 Composite summary of the magnitudes of the Workshop Main Effect

Dependent variables	Magnitude of the self Effect (self vs. non)
Covers the important ground	-.10
Avoids the irrelevant materials	.08
Is well organized	.25
Reads clearly and easily	.31
Overall rating this report	.19

- Scenario effects

There are a total of five scenarios. On all five criteria, we can divide the scenarios into two groups. Sarin and Al-Quada are relatively "easier". The average scores of reports on these two scenarios are higher than the other three. North Korea, South Africa and Russia/Iraq are relatively "harder".

The reports on the Russian/Iraq scenario got the lowest score on the Coverage criterion. The result is consistent with analysts' comments that they can easily find information on Russia or Iraq separately, but cannot find much information about both Russia and Iraq. And Russian/Iraq is one of the two short scenarios. The analysts had less time on the short scenarios than on the long ones. South Africa and North Korea are the other two "difficult" scenarios in term of the Coverage criterion. The difference between the two groups is significant.

The analysts judged reports on North Korea worst on the Overall criterion. The average score is significantly lower than those of Sarin and Al-Quada. All other differences are not significant.

On the other three criteria, Sarin and Al-Quada are also relatively easy scenarios while the other three are relatively hard. However the difference is not significant.

- Judge effects

Judge "u4" tends to give high judgment scores on all five criteria. Judge "u2" and "u3"'s effects on report judgment are similar. Except for the Organization criterion, "u4" gave significantly higher average score than "u2" and "u3".

Table 11 Composite summary of the magnitudes of the Judge Main Effect

Dependent variables	Magnitude of the Judge Effect	
	u2 vs. u4	u3 vs. u4
Covers the important ground	-.40	-.53
Avoids the irrelevant materials	-.93	-1.13
Is well organized	-.40	-.58
Reads clearly and easily	-.67	-.79
Overall rating this report	-.60	-.68

Summary and Conclusions

The objective of the workshops is to evaluate the HITIQA system under as real condition as possible. The quality of the entire report, which represents the end product of the search task is analyzed. Both the simple ANOVA and the rather complicated GLM analysis all show that "workshop" has significant effects on all five report quality criteria. Reports written at workshop II are much better than reports written at workshop I. As we discussed in the Introduction section, the Workshop Variable is a surrogate for system difference. However because of the learning effect and difference of instructions, we cannot firmly conclude that the HITIQA system is improved in the second workshop directly. The other qualitative and quantitative data we collected may help us to better understand the results.

However, the objective of this paper is to test whether such an evaluation can be done, rigorously, with such a small study. The effects of five independent variables are explored. In the linear model we used, besides the workshop, the Author difference is the other main effect that has significant effects on all five dependent variables. It is true that personal differences have effects on the analytical tasks. Our model can identify the effects of author differences, and show us how other factors affect the variance of the dependent variables after excluding the effects of author differences.

We find no self-judgment bias in our data. But it is interesting to notice that analysts gave their own reports lower scores on four criteria (Avoiding irrelevant information, Organization, Clearness and Overall). They seem confident about their own reports' coverage quality. All the differences are not significant.

In a sense, one of the most important results of this pilot study is established as soon as the interaction effects have

been shown to be relatively unimportant. In prior speculation and discussion it has frequently been assumed that interaction effects are so important and so significant that massive studies are needed in order to either control or evaluate them. Precisely because they are found here in practice to be small, we are able to extract meaningful and useful results from a very small study involving just three analysts, five topics, and two "systems". We point out that because of the nature of the analytic question-answering task, these are still not simple studies. Each session or workshop consisted of three eight-hour days of work on the part of the researcher and the analysts. However from the point of view of this paper that is a "mere detail" and the analytical methods described here can be applied to other results which might be obtained more easily in other settings. A more complete discussion of the workshops and the analysis of their other results is given in anther paper (Wacholder et al., 2004).

To sum up, the present study is able to make statistically significant distinctions, but the study is too small to be definitive. The primary value of the study is that it provides a framework which shows what types of data and analysis are needed to reach the conclusions in these kind of complicated evaluation experiments.

ACKNOWLEDGEMENT

The study is part of a large-scale multi-institutional project, named HITIQA, supported by the Advanced Research and Development Activity (ARDA)'s Advanced Question Answering for Intelligence (AQUAINT) Program under contract number 2002-H790400-000.

The cross-evaluation tool is built under the research supported in Part by DARPA under contract number N66001-97-C-8537 to Rutgers University.

REFERENCES

Kantor, P. Sun, Y. & Rittman, R. (2002). A technical Report on Prototype for Evaluating a complex Collaborative Information Finding System for the World-Wide Web: Evaluation of the Antworld System. APLab Technical Report APLab/TR-02/01 Research Supported in Part by DARPA under contract number N66001-97-C-8537 to Rutgers University. http://www.scils.rutgers.edu/antspace/FinalReport/APLabTR-02-01AntWorld.doc

NTCIR Project. Retrieved January 18, 2004 from http://research.nii.ac.jp/ntcir/.

Over. P. (2001) The TREC interactive track: an annotated bibliography. Information Processing and Management, 37(3), 369--381.

Saracevic, T., Kantor. P. et al. (1988). A Study of Information Seeking and Retrieving. I. Background and methodology. Journal of the American Society for Information Science 39(3), 161-176.

Saracevic, T. & Kantor. P. (1988). A Study of Information Seeking and Retrieving. II. Users, questions, and effectiveness. Journal of the American Society for Information Science 39(3), 177-196.

Saracevic, T. & Kantor. P. (1988). A Study of Information Seeking and Retrieving. III. searchers, searches, overlap. Journal of the American Society for Information Science 39(3), 197-216.

Smeaton, A.F. & Harman, D. (1997) The TREC (IR) Experiments and their Impact on Europe. Journal of Information Science, 23, 169-174.

Wacholder, N., Small, S., Bai, B., Kelly, D., Rittman, R., Ryan, S., Salkin, R., Song, P., Sun, Y., Ting, L., Kantor, P., & Strazalkowski, T. (2004). Designing a realistic evaluation of an end-to-end interactive question answering system. To appear in *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC '04)*, Lisbon, Portugal.

Voorhees, E. & Harman, D. (2000). Overview of the sixth Text Retrieval Conference (TREC-6). Information Processing & Management, 36(1), 3-3

Appendix A: General Linear Model Syntax

```
GLM
  cover noirrel organz clear overall  BY workshop t_aut
t_jud self scene
  /METHOD = SSTYPE(3)
  /INTERCEPT = INCLUDE
  /PLOT = RESIDUALS
  /CRITERIA = ALPHA(.05)
  /DESIGN = workshop t_aut t_jud self scene
  /PRINT = PARAMETER .
```