More Statistical Power Needed: The OHSU TREC 2002 Interactive Track Experiments

William Hersh
Susan Moy
Dale Kraemer
Lynetta Sacherek
Daniel Olson

{hersh, moys, kraemerd, sacherek, olsondan }@ohsu.edu
Division of Medical Informatics & Outcomes Research
Oregon Health & Science University
Portland, OR, USA

The original goal for the Oregon Health & Science University (OHSU) TREC 2002 Interactive Track experiments was to perform some preliminary experiments comparing searching on tablet devices versus ordinary personal computers. Unfortunately, the vendor who had promised the devices we planned to use was unable to deliver them in time for the experiments. We therefore shifted our experimental focus to assessing user factors found in previous experiments to be associated with success, with a particular desire to assess the role of spatial visualization.

A variety of studies have demonstrated that spatial visualization is associated with successful computer use. Egan and Gomez have shown that spatial visualization is associated with two processes in text editing: finding the location of characters to be edited and generating a syntactically correct sequence of actions to complete the task [1]. Similarly, Vincente et al. have found that the ability to use a hierarchical file system is associated with spatial visualization as well as vocabulary skills [2]. In addition, Allen has demonstrated that this trait is associated with the appropriate selection of key words in searching [3]. We have previously found that the ability of medical and nurse practitioner students to answer clinical questions found spatial visualization to be highly predictive of success [4]. (Spatial visualization actually demonstrated multicollinearity with whether a searcher was a medical or nurse practitioner student, which may have been the actual predictive factor.) In previous TREC Interactive Track experiments, our results showed a trend towards spatial visualization being predictive of searching success in instance recall tasks, although they did not achieve statistical significance [5].

Methods

Our methods employed the consensus approach agreed upon by track participants and posted on the track Web site (http://www-nlpir.nist.gov/projects/t11i/guidelines.html). We used the .GOV Web collection created for the TREC 2002 Web Track as the searching data. The collection was accessed by using the Panoptic search engine. We followed the experimental protocol developed for the TREC-9 Interactive Track, which was designed to allow the comparison of two systems or system variants using a minimum of 16 searchers. Each searcher performed eight tasks, which are listed in Table 1.

Table 1 - Eight searching tasks in TREC 2002 Interactive Track.

1. You are traveling from the Netherlands, and want to bring some typical food products as gifts for your friends. What are three kinds of food products from the Netherlands that you are not allowed to bring into the US? [Government Regulation]
2. You are concerned with privacy issues related to electronic information and would like to know what laws have been passed by the US Congress regarding these issues. Identify three such laws. [Government Regulation]
3. A friend has a private well which is the family's only source of drinking water. Locate a US publication, which contains guidelines for the maintenance of safe water standards for private well use. [Health]
4. You are not sure about the safety of genetically engineered foods, and would like to find more information and research on this topic. Name four potential types of safety problems that have been raised. [Health or Project]
5. You are interested in learning more about what measures the US government has taken since 2001 to prevent Mad-Cow Disease. Identify three such measures. [Health or Project]
6. Name/find three research programs/projects that investigate the treatment/causes of dwarfism. [Project]
7. You are planning a cycling expedition along the Silk Road in Central Asia. Find a website that is a good source information about health precautions should you take. [Travel]
8. You are planning to travel to the northeast territories of India and wonder if there are any problems/restrictions for tourists. Find a website that is a good source of information about such problems/restrictions. [Travel]

The data collection on each searcher included:

- Pre-Experiment questionnaire - measuring gender, age, educational level, searching experience, and general computer usage
- Paper-Folding Test (VZ-1) - measuring spatial visualization trait
- Pre-searching answer and certainty
- Post-searching with answer and certainty
- Exit questionnaire - measuring understanding of and satisfaction with experimental process
- Questionnaire for User Interface Satisfaction (QUIS) - validated questionnaire of satisfaction with a computer user interface [6]

Successful completion of the task was determined by evaluating the user answer. Since some questions required more than one answer (e.g., food products from the Netherlands), each user's search was assigned a score, with two points for a complete answer, one point for a partial answer, and zero points for wrong answer. The grading of results was done by an OHSU graduate student.

Since we were focused on user factors, our analysis was carried out on the level of searcher, not individual questions. A correlation matrix for the four major measurements (VZ-1 score, pre-searching score, post-searching score, and QUIS score) was built using Pearson's correlation coefficient with two-tailed testing for statistical significance.

Results

We recruited the minimum 16 searchers from students in the computer science program at Portland State University and the medical informatics program at OHSU. Experiments were carried out at a computer laboratory at OHSU using PCs running Microsoft Windows 2000 and the Internet Explorer version 5.5 Web browser, connected to the campus computer network, which was in turn connected to the Internet.

The general characteristics of the searchers are shown in Table 2. This group was highly experienced in computer use and Web searching, although they had lesser experience searching on-line public access catalogs and bibliographic indexes. They were highly experienced with searching related to their work, but less experienced searching in the searching tasks for this study, such as health, shopping, and government policy. The group unanimously reported Google as their preferred search engine.

Table 2 - General characteristics of searchers.

Characteristic	**Average result**
Gender	9 male, 7 female
Age	median 18-27
Experience using computers (1-none to 4-some to 7-great deal)	6.8
Experience using Web (1-none to 4-some to 7-great deal)	6.5
Frequency of use for work tasks (1-never to 4-monthly to 7-daily)	6.8
Frequency of use for academic tasks (1-never to 4-monthly to 7-daily)	6.7
Frequency of use for personal tasks (1-never to 4-monthly to 7-daily)	6.7
Level of expertise with computers (1-novice to 7-expert)	6.0
Experience with Web search engines (1-none to 4-some to 7-great deal)	6.3
Experience with OPACs (1-none to 4-some to 7-great deal)	4.8
Experience with indexes (1-none to 4-some to 7-great deal)	2.8
Usually find what looking for when searching Web (1-rarely to 4-sometimes to 7-often)	6.3
Frequency of searching for work (1-never to 4-monthly to 7-daily)	6.3
Frequency of searching for shopping (1-never to 4-monthly to 7-daily)	4.6
Frequency of searching for traveling (1-never to 4-monthly to 7-daily)	4.4
Frequency of searching for medical/health (1-never to 4-monthly to 7-daily)	4.0
Frequency of searching for government policy (1-never to 4-monthly to 7-daily)	2.7
Frequency of searching for entertainment (1-never to 4-monthly to 7-daily)	4.8
Overall expertise with searching (1-novice to 7-expert)	5.7
Years searching	5.2 years
Favorite search engine	All 16 - Google

Table 3 shows the results of the major searcher-related variables. Table 4 shows the correlation matrix for those variables, with Figures 1-3 showing VZ-1, pre-searching score, and QUIS score

plotted versus post-searching score. The largest correlation was 0.405 for pre-searching and post-searching scores, with an associated p-value of 0.12. Thus, no variable would enter a regression model and therefore additional analyses were not warranted. In order for a correlation coefficient of 0.40 to be significant with 80% power and a two-sided 5% significance level, a sample size of 47 would be required. The sample size was thus too small for any meaningful interpretation. The data were similarly analyzed using a non-parametric approach and comparable results were obtained.

Table 3 - Searcher-level analysis of major characteristics measured.

Searcher	**VZ-1 Score**	**Pre-searching Score**	**Post-searching Score**	**QUIS Score**
1	17.8	4	11	5.0
2	5.8	0	12	4.0
3	15	4	11	7.0
4	10	0	4	5.4
5	13	0	10	6.7
6	13.8	0	11	6.0
7	14	0	10	4.4
8	10	2	10	8.0
9	8.5	0	10	7.3
10	4.3	0	8	6.2
11	14.8	1	9	4.1
12	8.5	5	11	4.7
13	15	0	7	5.4
14	11	1	8	6.6
15	7	1	6	5.3
16	12.8	0	7	7.4
Average	11.3	1.1	9.1	5.9

Table 4 - Correlation matrix for searcher-level results.

	VZ-1 Score	**Pre-searching Score**	**Post-searching Score**	**QUIS Score**
VZ-1 Score	1	.23 .39	.16 .56	-.03 .91
Pre-searching Score		1	.41 .12	-.05 .87
Post-searching Score			1	-.09 .75
QUIS Score				1

Figure 1 - Scatter plot of VZ-1 score versus post-searching score.

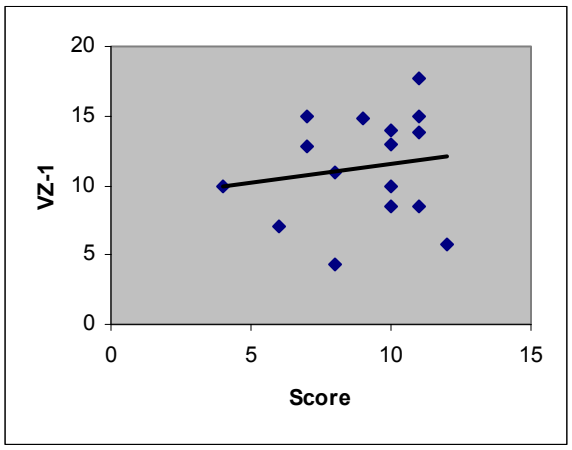


Figure 2 - Scatter plot of pre-searching score versus post-searching score.

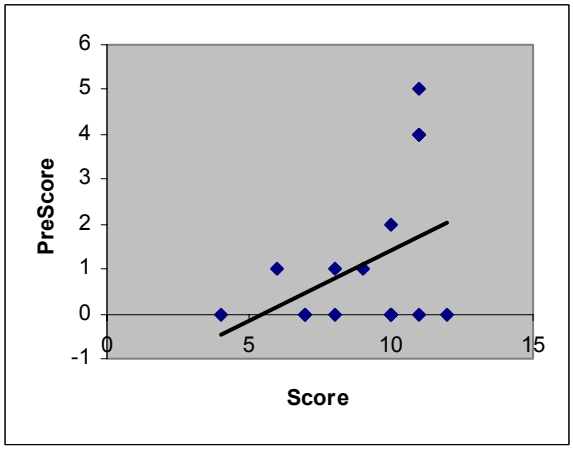
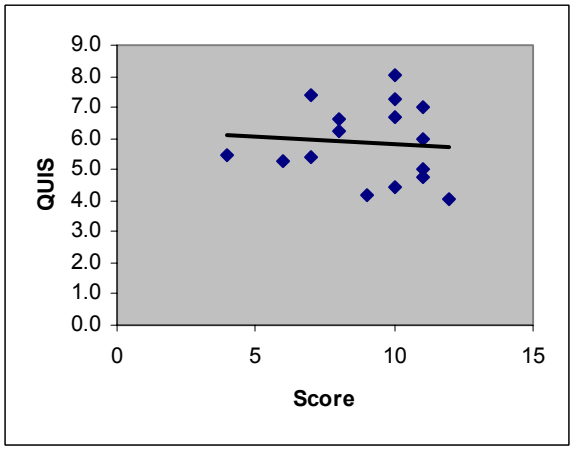


Figure 3 - Scatter plot of QUIS score versus post-searching score.



Conclusions

The OHSU results for the TREC 2002 Interactive Track showed some possible correlation between various user measures that did not reach statistical significance. Whether these measures were truly important could only have been assessed with a much larger sample size.

Statistical power is actually an overlooked challenge to evaluation of information retrieval systems. Even those carrying out batch-style evaluations must be concerned about it. Tague-Sutcliffe analyzed the results of the TREC-3 ad hoc experiments and found that the top half of the runs ranked by mean average precision had statistically insignificant differences with each other [7]. While Voorhees [8] has reassuringly found that results tend to keep their order even when different relevance judgments are substituted, Zobel has determined total recall is likely overestimated (i.e., additional relevant documents are likely to be found) [9]. In a related vein, Buckley and Voorhees analyzed the "stability" of results in batch studies and found a minimum of 25-50 queries needed to achieve it [10].

While most researchers who carry out evaluations in any field are probably familiar with statistical significance, which measures *alpha* error, and its meaning. Fewer research papers, however, report statistical power, which measures the minimization of *beta* error. Adequate statistical power is important in research, as an intervention in an experimental study may be of benefit, but the sample size is too small to tell. In fields such as medicine, the performance of "underpowered" clinical trials has been criticized, yet in many other experimental endeavors, researchers stop at reporting that results are "not statistically significant" [11]. Underpowered studies are a concern in TREC, especially given the nature of the venue, i.e., experiments performed on an annual cycle by research groups that do not generally have resources to carry out large-scale studies.

Another statistical challenge often overlooked in user-oriented IR evaluation studies is the non-independence of results from individual questions or topics. That is, analyses carried out at the level of individual question must take into account the fact that such questions are not completely independent, in that users search on multiple questions. In our previous experiments, we had to employ more complex statistical analyses to when evaluating factors at the level of the individual question [4, 5].

The results of our TREC 2002 Interactive Track experiments demonstrate that many measurable factors do influence the outcome of searching, but that sample sizes must be large enough to assess them well. The nature of the TREC experiments, with its short cycle for experimentation, can be at odds with adequately powered experiments. We hope to continue analyzing searchers once the .GOV collection has become stabilized.

References

1. Egan DE and Gomez LM, *Assaying, isloating, and accomodating individual differences in learning a complex skill*, in *Individual Differences in Cognition, Vol. 2*, Dillon R, Editor. 1985, Academic Press: New York.
2. Vincente KJ, Leske JS, and Williges RC, *Assaying and isolating individual differences in searching a hierarchical file system.* Human Factors, 1987. 29: 349-359.
3. Allen BL. *Cognitive differences in end-user searching of a CD-ROM index*. *Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1992. Copenhagen, Denmark: ACM Press. 298-309.
4. Hersh WR, et al., *Factors associated with success for searching MEDLINE and applying evidence to answer clinical questions.* Journal of the American Medical Informatics Association, 2002. 9: 283-293.
5. Hersh W, et al., *Challenging conventional assumptions of automated information retrieval with real users: Boolean searching and batch retrieval evaluations.* Information Processing and Management, 2001. 37: 383-402.
6. Chin JP, Diehl VA, and Norman KL. *Development of an instrument measuring user satisfaction of the human-computer interface*. *Proceedings of CHI '88 - Human Factors in Computing Systems*. 1988. New York: ACM Press. 213-218.
7. Tague-Sutcliffe J and Blustein J. *A statistical analysis of the TREC-3 data*. *Overview of the Third Text REtrieval Conference (TREC-3)*. 1994. Gaithersburg, MD: National Institute of Standards and Technology. 385-398.
8. Voorhees EM. *Variations in relevance judgments and the measurement of retrieval effectiveness*. *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1998. Melbourne, Australia: ACM Press. 315-323.
9. Zobel J. *How reliable are the results of large-scale information retrieval experiments? Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1998. Melbourne, Australia: ACM Press. 307-314.
10. Buckley C and Voorhees E. *Evaluating evaluation measure stability*. *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2000. Athens, Greece: ACM Press. 33-40.
11. Halpern SD, Karlawish JHT, and Berlin JA, *The continuing unethical conduct of underpowered clinical trials.* Journal of the American Medical Association, 2002. 288: 358-362.