A Technical Note on TREC Evaluation

Peter Pirolli
Xerox PARC

Mark Wilson
University of California, Berkeley

December 1995

INTRODUCTION

The purpose of this note is to address some of the evaluation problems raised by recent discussions concerning the TREC interactive track. These include

1. Assuring *comparability* of system measurements. This is essential to establishing scientific measurement.
2. Assessing and accounting for *measurement error* due to user sampling and topic sampling, as well as site-specific sampling bias effects. This is directly related to the issue of improving the *power* of the experiments.
3. Developing appropriate statistical techniques.

We will not be concerned with two other very important kinds of issues that need to be addressed by the Interactive Track specification. The first of these concerns the representativeness or *validity* of the assessment tasks. Such issues address the degree to which the evaluation tasks are characteristic of tasks in the real world, or some desired state of affairs. Neither will we be concerned with the *heuristic value* of the approved evaluation tasks. Once a competition is defined and a gold medal offered to the winner, we can expect to see evolution in action. It is up to the TREC promoters and community to decide on the evaluation tasks that best move the field forward. This note is strictly concerned with scientific methodology and analysis.

The outline of approach and examples presented here should be taken as a starting point for further discussion and refinement. Much of the work requires additional review since these designs and analyses are nonstandard.

THE PROBLEM SITUATION

Different interactive information retrieval (IR) systems are tested at different sites by their developers. One or more users at each site are selected to participate in the studies and are asked to retrieve as many relevant documents (and as few irrelevant ones) from a common test corpus and in response to a common set of topic queries (25 of them). Each topic takes about 30 minutes to complete. Outcomes for each system and each query (across one or many participants) are submitted to a central authority for score computations. It is tempting to think of this as a design of System × Topic variables. Unfortunately, the problems with this design are many.

First, there are unknown effects of sites. In the past, participant selection methods at individual sites have varied greatly. In an ideal experiment, one would like to have participants randomly sampled from some population pool and then randomly assigned to sites. Even if a standardized solicitation procedure could be adopted at all of the sites (for instance, by advertising at the site institutions), there would remain demographic differences between locales and demographic similarities within locales that prohibit any semblance of true random selection from the population at large. Such random sampling is assumed by textbook approaches to regression and analysis of variance (ANOVA). It has been proposed that covariate measures (such as typing speed, experience with IR systems, etc.) be assessed of participants and used to partition out some of these site effects as well as participant effects. Unfortunately, standard analysis of covariance techniques also assume random assignment to experimental conditions. A method must be developed that addresses these site effects directly.

Second, previous evaluations have not assessed *measurement error* or experimental error. This is the uncontrolled variance that contributes to an observation. This is basically the sum of the variances within each cell of a multi-factor design. One way reduce this error is to introduce more controlled factors.

Many testers have assumed that there is no variation in the performance of participants and have only used a single participant. This ignores the relevant data obtained in TREC4, not to mention over a 100 years of scientific psychology. Other testers have come up against resource limitations due to the requirement of testing 25 topics and have been unable to study many participants. The proposed method must assess error due to user variance and measurement error.

Third, testing 25 topics for about 13 hours to obtain a single observation per topic is very demanding of research resources. Generally, the *power* of a significance test conducted on an experiment (i.e., the probability of rejecting the null hypothesis when there is a true difference) increases as measurement error decreases.[1] Although it is true than increasing the number of observations (on topic queries) is one approach to reducing measurement error, we noted above that the other method is to increase the number of experimentally controlled variables. The experimental categorization of individual topic queries is one way to do this and could increase experimental power while requiring fewer topic queries to be used.

Fourth, the framework ought to permit additional exploratory secondary data analyses to take place, perhaps to investigate similarities among topics, individual differences in strategies, and so on.

THE SOLUTION

The solution we propose combines the following three stages, each involving design elements and analyses of increasing sophistication:

A.1 A variation of what is called the Pretest-Posttest Control Group design (or a Difference Score) design (Campbell & Stanley, 1963).

---

[1]Determination of the power of an experiment and the number of observations required in order to reject the null hypothesis is illustrated in most textbooks on experimental design. The procedure basically involves making the best available guesses about the true differences, the variances, making some simple calculations and consulting standard charts.

A.2 A multilevel or hierarchical design based on approaches developed in educational research to address the nesting of participants within sites (Raudenbush & Bryk, 1986), elaborated to include topic queries as additional random factors.

A.3 The use of covariate measures on participants.

B The use of randomized block designs on topics to improve precision.

C A latent variable method to address the issue of query sampling.

Our presentation will initially adopt an approach that assumes the designs will be analyzed using analysis of variance (or equivalently, linear regression) techniques. We do so mainly because there is a widely developed and well-established literature on these approaches that is familiar to most working scientists. It should not be a problem to develop nonparametric or nonlinear equivalents. Toward the end of our discussion we will raise an alternative approach that is based on Item Response Theory developed in psychometrics.

Measurement and Comparability

In the natural sciences, it is not uncommon to begin textbooks with a chapter on measurement (e.g., (Halliday & Resnick, 1970)). This is mimicked by both editions of *Steven's Handbook of Experimental Psychology* (Atkinson, Herrnstein, Lindzey, & Luce, 1988; Stevens, 1951). The information retrieval community, and TREC in particular, has sought to *measure* system performance. This is a very powerful goal to hold, since direct quantitative measurement of theoretical variables is a strong marker of understanding and control over the phenomena of interest. Measurement, however, is founded on experimental control and comparability:

> Quantification in science is inseparable from the experimental method. The hypothesis that a particular variable is quantitative is a substantive hypothesis. It requires that the values of that variable manifest a definite kind of structure. With the possible exception of some quantitative measurements open to extensive measurement, evidence for or against the quantitative measurement must be gained by experiment. The gathering of such evidence requires a high degree of experimental control and, often, sophisticated apparatus and methods of observation. It is no accident that the extension of quantification in physics from geometry and statics to dynamics, thermodynamics, and electrical phenomena went hand in hand with experimental techniques and apparatus. ((Michell, 1990), p. 86).

The marks of quantity are established by *ordinal relations* and *additive structure* among the variables of interest (Campbell, 1928; Krantz, 1964; Luce & Tukey, 1964). In the case of extensive properties, such as length, specific ordinal relations are clearly manifest in comparisons of objects of different lengths and additivity is manifest in the manner in which lengths can be concatenated to produce new lengths. For example, with respect to the extensive property "length", note that two particular objects can be compared and ordered ("longer than") and they can be concatenated in a particular way to yield a new length that can be tested against other lengths. These ordering, comparison, and concatenation operations must meet certain conditions in order to be quantified in a coherent and meaningful way (Michell, 1990).

The TREC interactive evaluations have proceeded as if systems combined with people could be measured on ruler-scales as if they were lengths of rope. The problem, however, is that the evaluation tasks measure many variables simultaneously. This raises special issues regarding the separation and scaling of the different variables.

TREC testers get observations that simultaneously measure many variables, including variables associated with the systems, users, and topic queries. The contribution of each of these variables to the observed performance has to be separated from the mix. The observations have to be conceptually arranged in a structure that allows one to separate the effects of the component variables and to understand how they combine to produce the observations. The essence of such structure involves being able to compare observations (order them) and being able to perform addition and subtraction (perhaps after applying a transformation to the data). The additive factors logic underlying experimental designs aims to do this, and such logic will be essential to separating the effects of systems from the effects of sites, users, and topics. As Campbell and Stanley (1963) note, the issue of experimental design is paramount and separable from the intricacies of appropriate statistical analysis:

> Good experimental design is separable from the use of statistical tests of significance. It is the art of achieving interpretable comparisons and as such would be required even if the end product were to be graphed percentages, parallel prose case studies, photographs of groups in action, etc. In all such cases the interpretability of the "results" depends upon control over the factors we have been describing...Use of significance tests presumes but does not prove or supply the comparability of the comparison groups or the interpretability of the differences found. (p. 192)

Random vs Fixed Factors

Some issues concerning the proper framework for analysis of evaluations will depend on whether we view the effects of topic queries as *random* or *fixed* effects . It is worth reviewing the meaning of those terms in the context of designs based on analysis of variance.

A *factor* (sometimes called a *treatment*, *experimental variable*, *independent variable*) consists of a series of related *levels*. The levels of a factor may be quantitative variations on an essentially quantitative variable (e.g., the word length of topic queries) or qualitatively different categories within an essentially qualitative variable (e.g., the IR system used among a set of IR systems).

An experiment draws upon *P potential* levels of a factor, a_1, a_2, ...a_P. A *fixed factor is* used when

1. The number of factor levels included in the experiment is *P*, or
2. The selection of factor levels is done by a nonrandom procedure (for instance, constructing an experimental factor for query-word length in which the levels are: word-length = 4, word-length = 8, word-length = 12).

A *random factor* is used when the levels of the factor are a random sample from the *P* potential levels. Typically, psychology experiments attempt to randomly sample a participant population and so participants are a random factor in such designs.

The questions that will arise (from the analysis of variance perspective) will concern whether we think of specific topic queries as fixed factors constituting the universe of queries or as random samples from broader populations.

Use of a Common Control Task in a Hierarchical Difference Score Design

A basic experimental design solution can be modeled after education research (Campbell & Stanley, 1963) involving the use of "difference scores" or repeated measures obtained from all users working on a common control (C) IR system and the site-specific experimental (E) IR system. The difference between the C and E IR system performances is the basic measurement of interest. To address the effects of nesting participants within sites, we can elaborate upon a multilevel structural or hierarchical models, similar to those developed in educational research (Raudenbush & Bryk, 1986). Such models have been used to examine effects of school sites in relation to other variables thought to affect academic achievement.

Example 1: An Elementary TREC Design and Analysis.

The first example will be used to introduce the basic scheme for comparing effects and to raise some issues for discussion. Several simplifying assumptions will be made for the purposes of illustration and so this design should be considered unrealistic.

Let us suppose that there are three test sites, Site A, Site B, and Site C, each testing their own IR systems, System 1, System 2, and System 3, respectively. A control IR system, System 0, is available over the net at all sites. For the sake of concreteness we might imagine that the dependent variable of interest is number of relevant documents retrieved (as assessed by the TREC experts), but the response measure could equally well be precision or recall. One also might want to analyze all three response measures.

Let us also suppose that there are four participants at each site. Each participant answers $q = 1, 2, \ldots Q$ topics on the control system (System 0) and on the experimental system (System 1, 2, or 3). That is, each topic will be tested $r = 1, 2$ times. Assume that the sequencing of topic presentation and the sequencing of system testing are appropriately counterbalanced across participants such that practice effects and other transfer effects are eliminated.[2] For the q^{th} topic, we might arrange the data as depicted in the four leftmost columns of Table 1.

At this point we will propose structural model for this design that might be used to develop an analysis of variance, then we will present a regression equation that could be used to estimate the relevant effects. The response y_{ijsq} for $i = 1, \ldots, n_j$ participants at $j = 1, \ldots, J$ sites, on $s = 1, 2, \ldots S$ systems, on $q = 1, 2, \ldots Q$ topic queries would be

$$y_{ijsq} = \mu + \alpha_j + \beta_{s(j)} + \varepsilon_{i(s)} + \varepsilon_q \tag{1}$$

where

$\mu =$ the grand mean,

$\alpha_j =$ the effect of site,

$\beta_{s(j)} =$ the difference effect of system within site,

$\varepsilon_{i(s)} =$ error associated with subjects within sites,

$\varepsilon_q =$ error associated with query topics.

The notation $\beta_{s(j)}$ indicates that the effect of systems (s) is nested within the effects of site (j).

---

[2]This is probably unrealistic, but let us make the assumption in order to develop some of the other elements of this illustration.

Table 1

Conceptual layout of data for the q^{th} topic tested on four users at three sites.

Participant i	*Site* j	*System*	*Observation* y_{ijrq}	X_{1q}	X_{2q}	X_{3q}	X_{Bq}	X_{Cq}
1	*A*	0	y_{1A1q}	0	0	0	0	0
1	*A*	1	y_{1A2q}	1	0	0	0	0
2	*A*	0	y_{2A1q}	0	0	0	0	0
2	*A*	1	y_{2A2q}	1	0	0	0	0
3	*A*	0	$\vdots$	0	0	0	0	0
3	*A*	1	$\vdots$	1	0	0	0	0
4	*A*	0	$\vdots$	0	0	0	0	0
4	*A*	1	$\vdots$	1	0	0	0	0
5	*B*	0	$\vdots$	0	0	0	1	0
5	*B*	2	$\vdots$	0	1	0	1	0
6	*B*	0	$\vdots$	0	0	0	1	0
6	*B*	2	$\vdots$	0	1	0	1	0
7	*B*	0	$\vdots$	0	0	0	1	0
7	*B*	2	$\vdots$	0	1	0	1	0
8	*B*	0	$\vdots$	0	0	0	1	0
8	*B*	2	$\vdots$	0	1	0	1	0
9	*C*	0	$\vdots$	0	0	0	0	1
9	*C*	3	$\vdots$	0	0	1	0	1
10	*C*	0	$\vdots$	0	0	0	0	1
10	*C*	3	$\vdots$	0	0	1	0	1
11	*C*	0	$\vdots$	0	0	0	0	1
11	*C*	3	$\vdots$	0	0	1	0	1
12	*C*	0	$y_{12,C,1,q}$	0	0	0	0	1
12	*C*	3	$y_{12,C,2,q}$	0	0	1	0	1

These data in Table 1 could be entered into an ordinary least squares (OLS) regression, using "dummy" variables. One set of variables will be used to code the sites and another to code the systems. The regression equation for each topic q would be

$$\hat{y}_q = \beta_{0q} + \beta_{1q}X_{1q} + \beta_{2q}X_{2q} + \beta_{3q}X_{3q} + \beta_{Bq}X_{Bq} + \beta_{Cq}X_{Cq}, \tag{2}$$

where

$$X_{1q} = \begin{cases} 1, & \text{if System 1} \\ 0, & \text{otherwise} \end{cases},$$

$$X_{2q} = \begin{cases} 1, & \text{if System 2} \\ 0, & \text{otherwise} \end{cases},$$

$$X_{3q} = \begin{cases} 1, & \text{if System 3} \\ 0, & \text{otherwise} \end{cases},$$

$$X_{Bq} = \begin{cases} 1, & \text{if Site B} \\ 0, & \text{otherwise} \end{cases},$$

and

$$X_{Cq} = \begin{cases} 1, & \text{if Site C} \\ 0, & \text{otherwise} \end{cases}.$$

The values of these variables for the observations for the q^{th} topic are presented in the five rightmost columns of Table 1.

Note that one could apply the analysis described above to an aggregate over the individual query outcomes (say, the simple sum, or an aggregate of transformed outcomes). This would give some indication of general trends, but would not allow one to examine the results for differential effects by query topic.

If we carry out separate regressions on the $q = 1, 2, \ldots Q$ topics we would get a set of parameter estimates for the β's in Equation 4 that could be arranged as in Table 2. The means of the parameter estimates are along the bottom row margin. Relative to the reference value of $\bar{\beta}_0$, which is the predicted response on System 0 at Site A, the remaining mean parameter values give the effect of System 1, System 2, and System 3, and Sites B and C. Furthermore, the distribution of the values in each of the columns could be investigated directly by inspection (or plotted as histograms) to reveal the distributions of topics effects.

Table 2

Arrangement of predictors obtained from the regression on each topic.

Topic	β_{0q}	β_{1q}	β_{2q}	β_{3q}	β_{Bq}	β_{Cq}
1						
2						
$\vdots$						
Q						
	$\bar{\beta}_0$	$\bar{\beta}_1$	$\bar{\beta}_2$	$\bar{\beta}_3$	$\bar{\beta}_B$	$\bar{\beta}_C$

The design above is based on the assumption that sites are closely linked to particular systems. By using just one system as a universal comparison, we are making a minimal assumption that at least one system can be implemented without great difficulty at each site. This is, of course, not the ultimate aim--each system should be easily implementable at a broad range of sites. In such a situation, one could have more freedom in design. For example, a completely random allocation of systems to sites would allow of somewhat simpler analyses. Or, one could create more efficient designs than those described here.

Note that the nesting of some factors within others means that certain interactions cannot be tested. This is a drawback to the design that is necessitated by practical considerations. For instance, we would not be able to test any interactions involving both sites and systems. We have also not discussed how to test the significance of effects, which will be addressed in Example 2.

Topics-as-Fixed-Effect Fallacy

Some of the recent TREC e-mail discussions have suggested that there are substantial System × Topic interactions, although no generalizations could be made about such interactions. This leads us to believe that some testers think of the topics as "fixed effects" rather than samples from some population(s) of queries. Similar fixed-effect assumptions plagued the analysis of psycholinguistic data before they were diagnosed (Clark, 1973). Methods that solve this problem include selecting appropriate experimental designs, sampling procedures, and statistical techniques. The "topics-as-fixed-effect" assumption is problematic for at least two reasons: (1) it may seriously undermine inferences about the generalizability of results and motivate inappropriate statistical analyses and (2) it may seriously undermine the power of the studies. For these reasons it is worth reviewing some of the recommendations on the "language-as-fixed-effect fallacy" (Clark, 1973), as well as reviewing how to increase experimental power through what are known as blocked designs.

Some of the issues worth consideration are:

- Topics should be treated as random factors just as participants are treated as random factors in the experimental design. It may be possible to categorize topics into different classes for the purposes of experimental study, but within those classes the individual query items must be treated as random samples from the category of interest. In Equation 1, the topic queries were treated as a random factor.
- Just as within-subject designs improve experimental power, so do within-topic designs. Example 1 tested topic queries across systems to achieve a within-topic subdesign.
- An experimental design is only as sensitive as the least sensitive of the Participants × Systems or Topics × Systems subdesign. The measurement error associated with each has to be considered and appropriately analyzed (see (Clark, 1973)). One such test statistic is the *quasi-F* test.
- Generalizability and power are contingent on systematic repeatable sampling: (1) define populations of topics, (2) sample without bias, (3) sample by replicable procedure.
- The error variance can be reduced if it is possible to remove the contribution of individual difference variables among topics (or participants). This makes it easier to detect the effects of the independent variable of interest (i.e., systems). This can be achieved by blocking topics (or participants) on the

basis of their scores on some concomitant variable that is thought to be correlated with the performance measure. As one gains knowledge of the relevant variances and correlations, one can design the optimal number of blocks and items within blocks (Meyers, 1966).

- Rather than using blocking as a method for improving experimental power, one might be interested in experimentally studying the performance effects of different categories of topic queries. If systematic topic category effects are found, then one has gets a gain in power (for the same reasons as blocking) as well as corroboration of a substantive hypothesis about topics.

To illustrate a possible design, imagine that a total of Q topics are divided into two types on the basic of some feature thought to have an effect on performance: Q1 topics and Q2 topics (each type having Q/2 topic queries). The Example 1 design would be modified as follows. Each participant would now be tested on Q1 topics and Q2 topics on both System 0 and their experimental system. If the performance difference between Q1 and Q2 topics is sufficiently strong, then fewer than Q/2 topics could be used for each topic condition. This also allows more systematic investigation of Topic Type X System interactions. The regression Equation 2 would be modified to include variables coding Topic Type, as well as the relevant interaction terms.

Blocking and Counterbalancing

Blocked experimental designs are often constructed in conjunction with counterbalancing schemes., so it is worth reviewing these also. Counterbalancing schemes usually refer to what are called *Latin Square* designs (which have close and more sophisticated relatives called *Greco-Latin Square* designs), which are very common in psychological research. We used a Latin Square design on an earlier Scatter/Gather experiment that will be presented at CHI-96, so let me use that as a concrete example of the more general idea.

We defined three types of query: easy (E), medium (M), and hard (H), with 4 queries in each type (i.e., 12 altogether). We then created four blocks A, B, C, D, each containing an E, M, and H type queries. Two systems being tested were I1 and I2. We wanted the testing to proceed in the order I1, I2, I1, I2. The first step in creating a Latin Square is to create an N x N "standard square" where each row represents a participant and each column some treatment, task, material, or presentation order, or whatever aspect of the design for which you want to reduce measurement error. So the standard square is

	I1	I2	I1	I2
S1	A	B	C	D
S2	B	C	D	A
S3	C	D	A	B
S4	D	A	B	C

Note the constraints: Each participant is presented with all blocks, each column contains all blocks.

The rows then the columns are permuted to get something lik

	I1	I2	I1	I2
S1	C	B	A	D
S2	D	C	B	A
S3	B	A	D	C
S4	A	D	C	B

which maintains the constraints but introduces a randomization. This is a Latin Square. More complicated things can be done for higher levels of combinations (e.g., 2 x 2 x 2).

Example 2: A Design Including Topic Types and its Analysis

Suppose that the existing bank of TREC topic queries is examined and categorized into $t = 1, 2, ...T$ topic types that are thought to produce differential effects. The layout of each participant's data in Table 1 would now become,

Participant	*Site*	*System*	*TopicType*	*Query*	*Observation*
i	j	0	1	1	y_{ij011}
⋮	⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	Q	⋮
⋮	⋮	⋮	2	1	⋮
⋮	⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	Q	⋮
⋮	⋮	⋮	T	1	⋮
⋮	⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	Q	⋮
⋮	⋮	s	1	1	⋮
⋮	⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	Q	⋮
⋮	⋮	⋮	2	1	⋮
⋮	⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	Q	⋮
⋮	⋮	⋮	T	1	⋮
⋮	⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	Q	y_{ijktq}

That is, each subject would be tested on T topic types, each type consisting of Q queries, on both the control System 0 and the experimental System s.

The structural model for an analysis of variance would be an elaboration of Equation 1,

$$y_{ijsq} = \mu + \alpha_j + \beta_{s(j)} + \gamma_t + \alpha\gamma_{jt} + \beta\gamma_{s(j)t} + \varepsilon_{i(s)} + \varepsilon_{q(t)}, \tag{3}$$

where we have now added a factor γ_t for topic types. Terms for the topic type by site interaction $\alpha\gamma_{jt}$, and the topic type by system within site interaction $\beta\gamma_{s(j)t}$, have now been added. Because of nesting in the hierarchical design, no other interactions can be tested.

Regression equations could also be developed by analogy to Equation 2. For instance, suppose we the situation illustrated in Table 1 is elaborated such that we now have two topic types {0, 1}. Let

$$X_I = \begin{cases} 0, & \text{if topic type 0} \\ 1, & \text{if topic type 1} \end{cases}.$$

For instance, regression of

$$\hat{y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_B X_B + \beta_C X_C + \beta_I X_I, \tag{4}$$

would provide estimates of the site, system, and topic types effects.

To perform statistical tests based on analysis of variance we will need to do two things that are nonstandard. First, our designs are nonstandard because they contain nested and crossed-nested (interaction) terms. These deviate from the fully factorial ANOVAs that are performed on commonly available statistical software. Second, we have multiple random factors that block the use of the standard *F* statistic, so we will use what are known as *quasi F* ratios (Winer, 1962).

Fully general, rule-based procedures for determining the appropriate sums of squares, degrees of freedom, and mean square error terms are presented in most textbooks on analysis of variance and experimental design (e.g., (Winer, 1962)). More expedient methods making use of standard ANOVA programs for fully factorial designs are also possible (Montgomery, 1984). The basic procedure is to submit the data to an ANOVA as if it were a fully factorial design then, for each nested factor, to pool certain sums of squares and degrees of freedom indicated in the analysis output. Specifically, one pools the nested factor, plus all interactions involving the factor and its nesting factor. Whatever the approach, the essential sources of variance can be laid out as in Table 3.

In Table 3, each sum of squares is the sum of the squared deviations of each element of the source of variation about the mean for that source. For instance, the sum of squares for topic types computed by squaring the deviation of each topic type mean about the grand mean of topic types and summing those squares. The degrees of freedom are computed as the number of independent observations in the source of variation less the number of independent parameters that are estimated. Again, the details of computing the sum of squares and degrees of freedom can be found in many textbooks. The mean square is the sum of squares divided by degrees of freedom for a source.

Common practice is to test hypotheses by constructing an *F* statistic. The value of *F* for a particular treatment (a factor or interaction) is the ratio of the mean square for that factor divided by an appropriate error term:

$$F(df_{treatment}, df_{error}) = MS_{treatment}/MS_{error}.$$

One then can look up the significance of the *F* statistic value in standard tables arranged by the degrees of freedom $df_{treatment}$, df_{error}. Table 4 shows the test statistics for the various factors and interaction in Table 3. Only the first effect (of sites) can be tested with a simple *F* statistic.

For most of the factors and interactions of interest in our example, there is more than one source of error appropriate for analysis. We have error due to both subjects and queries. Winer (1962) has recommended the general approach of using quasi *F* statistics, F', and Clark (1973) has recommended the use in language research. Table 4 contains what seems to be the appropriate F' statistics for the remaining factors and interactions.

Table 3.
Layout of the Analysis of Variance for Example 2

Source	Sum of Squares	Degrees of Freedom	Mean Square (SS/df)
Between Subjects			
Site (A)	SS_A	df_A	MS_A
Subjects within Site (S w. S)	SS_{SwS}	df_{SwS}	MS_{SwS}
Between Topics			
Topic Types(B)	SS_B	df_B	MS_B
Queries within Topic Types (Q w. T)	$SS_{Q(T)}$	$df_{Q(T)}$	$MS_{Q(T)}$
Subjects by Queries			
S w. S × Q w. T	$SS_{S(S)Q(T)}$	$df_{S(S)Q(T)}$	$MS_{S(S)Q(T)}$
Within Subjects			
Topic Type (B)	SS_B	df_B	MS_B
A × B	SS_{AB}	df_{AB}	MS_{AB}
B × S w. S	$SS_{BS(S)}$	$df_{BS(S)}$	$MS_{BS(S)}$
System (C)	SS_C	df_C	MS_C
A × C	SS_{AC}	df_{AC}	MS_{AC}
C × S w. S	$SS_{CS(S)}$	$df_{CS(S)}$	$MS_{CS(S)}$
B × C	SS_{BC}	df_{BC}	MS_{BC}
A × B × C	SS_{ABC}	df_{ABC}	MS_{ABC}
B × C × S w S	$SS_{BCS(S)}$	$df_{BCS(S)}$	$MS_{BCS(S)}$

Table 4
Test statistics for Example 2.

Factor or Interaction	Test Statistic
A	$F(df_A, df_{S(S)}) = \dfrac{MS_A}{MS_{S(S)}}$
B	$F'(i,j) = \dfrac{(MS_B + MS_{S(S)Q(T)})}{(MS_{BS(S)} + MS_{Q(T)})}$
A × B	$F'(i,j) = \dfrac{(MS_{AB} + MS_{S(S)Q(T)})}{(MS_{BS(S)} + MS_{Q(T)})}$
C	$F'(i,j) = \dfrac{(MS_C + MS_{S(S)Q(T)})}{(MS_{CS(S)} + MS_{Q(T)})}$
A × C	$F'(i,j) = \dfrac{(MS_{AC} + MS_{S(S)Q(T)})}{(MS_{CS(S)} + MS_{Q(T)})}$
B × C	$F'(i,j) = \dfrac{(MS_{BC} + MS_{S(S)Q(T)})}{(MS_{BCS(S)} + MS_{Q(T)})}$
A × B × C	$F'(i,j) = \dfrac{(MS_{ABC} + MS_{S(S)Q(T)})}{(MS_{BCS(S)} + MS_{Q(T)})}$

It remains to illustrate the computation of the i and j degrees of freedom for the F' statistics. Clark (1973) suggests that i is the nearest integer value to

$$i' = \frac{(MS_1 + MS_2)^2}{\left(\dfrac{MS_1^2}{n_1} + \dfrac{MS_2^2}{n_2}\right)}$$

where MS_1 and MS_2 are the two mean squares in the numerator of the F' statistic and n_1 and n_2 are their associated degrees of freedom. Similarly j is computed by the same formula except that MS_1 and MS_2 are the means squares for the F' denominator.

Example 3: Use of Covariates

The regressions and analysis models could also be elaborated to include additional covariate measures on participants or on topics. For instance, years of schooling, years of experience with IR systems, typing speed, etc. could be assessed of participants prior to testing. Analogously, measures on the topic queries could be assessed prior to testing. Terms containing variables encoding this measures could be added to the regressions and analysis models prior to fitting the data. Again, error variance will be reduced to the degree that the covariate variables correlate with performance, and this will improve our ability to resolve the effects of different systems.

GENERAL DISCUSSION

We have sketched out some possible solutions to the TREC evaluation problems. We conclude with some remarks about recent analysis approaches that may be more powerful and appropriate, but would require work by experts in the field of measurement.

Hierarchical Linear Modeling

Bryk and Raudenbush have develop a general approach called hierarchical linear modeling (HLM) and software (about $400) that works for some designs. The basic model was developed for the analysis of variables affecting student performance at different school sites. The general structural model for within-site effects on a participants response would be

$$y_{ij} = \beta_{jo} + \beta_{j1} X_{ij1} + \cdots + \beta_{jK-1} X_{ijK-1} + R_{ij} \tag{5}$$

for $i = 1, ..., n_j$ participants at $j = 1, ..., J$ sites, with $k = 0, ..., K$-1 independent variables in the within-site model. For our purposes, the X_{ijk} would be independent variables on participants, such as level of familiarity with the system or topic. The B_{jk} are model parameters (to be estimated) which indicate the dependence of the outcome on the predictors, and the R_{ij} is the error term.

Each of the within β_{jk} coefficients in Equation 5 has a general between-sites model,

$$\beta_{jk} = \theta_{0k} + \theta_{1k} Z_{1j} + \cdots + \theta_{P-1,k} Z_{P-1,j} + U_{jk} \tag{6}$$

where $p = 0, ..., P$-1 independent variables in the between-site model, which in our case might be something like type of facility (academic vs industry for instance). The Z_{pj} are independent variables on sites and the θ_{pk} are model parameters (to be estimated) which indicate how the site characteristics affect the relationship between the user characteristics and the outcome variable. The U_{jk} are error terms.

As discussed in Raudenbush and Bryk (1986), the hierarchical linear model (HLM) of Equations 5 and 6 can be re-expressed as a single-level model that is very close to our examples,

$$y_{ij} = \theta_{00} + \sum_{k=1}^{K-1} \theta_{0k} X_{ijk} + \sum_{p=1}^{P-1} \theta_{p0} Z_{pj} + \sum_{k=1}^{K-1} \sum_{p=1}^{P-1} \theta_{pk} X_{ijk} Z_{pl} + U_{j0} + \sum_{k=1}^{K-1} U_{jk} X_{ijk} + R_{ij}. \tag{7}$$

If the error terms U_{jk} were all zero then standard ordinary least squares (OLS) would be appropriate. Whwnever that is not the case, as is most commonly the case, then OLS regressions of Equation 7 will be inefficient, the standard errors too small, and there is a lack of resistance to outliers. This means that if the between-sites variables can be assumed to encode all predicted variation between-sites, then OLS can be used.

The multi-level model of Equations 5 and 6 is addressed by HLM (Raudenbush & Bryk, 1986), which uses empirical bayesian estimation techniques and the EM algorithm to address these problems. Most importantly, it separates parameter variance from sampling variance, and "shrinks" the outliers towards the means. The HLM analysis typically gains substantial power over OLS techniques.

The problem with the HLM analysis programs, however, is that they would not be able to deal with the two sources of error (participants and topic queries). The development of such software would require some contracting with outside researchers.

Beyond Hierarchical Analysis of Variance

An alternative way to conceptualize a specific topic set is as a sample from a large "bank" of topics, such as the bank of topic queries already accumulated by TREC. The topic bank can be expanded with time, and can also be designed with structures that represent important sub-topics that may be based on practice and history and/or be the result of empirical investigations of outcome patterns. Instead of making reference to a hypothetical (and unattainable) infinite set of possible queries, this focuses attention on those that we actually have some information about. In testing, this is broadly the distinction between "generalizability theory" and "item response modeling".

The multilevel model described above could be applied to an aggregate of query outcomes, just as was described for the ANOVA approach, or it could be applied to outcomes for individual topics. Both are somewhat unsatisfactory, however, leaving one either ignorant of how individual topics affect the outcomes, or , alternatively, leaving one unsure how to combine the results across the topics. There are a number of recent developments that may indicate different ways to resolve this.

First, there has been some work in the multilevel modeling literature that attempts to deal with the situation where there are to sorts of clustering (i.e., sites and topics). Although these areas yet somewhat restrictive in their design possibilities, they might be applicable to somewhat different designs compared to the one described above.

Second, if one takes the item response modeling approach described above, then a family of models that can be applied to the design above. These would allow considerable flexibility in the way that the competition were designed, even allowing variation in that design from year to year (say, as one built in a typology of queries) while retaining comparability.

Third, the introduction of "random effects" elements into item response modeling approaches such as that of Adams and Wilson is an ongoing research area, and may soon result in options that avoid the need to introduce concepts such as a "query bank".

To sketch out the approach, we might want to develop a multi-dimensional representation of information retrieval*situations*, which consist of multi-dimensional representations of systems and topics. We might also want to develop multi-dimensional representations of users' *responses* to these situations. We might then use a measurement model developed in the context of item response modelling: the Random Coefficients Multinomial Logit (RCML) model (Adams & Wilson, 1992; Pirolli & Wilson, in press).

We can represent a user by a *D*-dimensional vector of (possibly unknown) user-variables, $\underset{\sim}{\theta}$. These might represent different elements of user background knowledge, retrieval goals, or search strategies, for instance. Also, suppose that the task environment (e.g., system + topics + task definition) in which these variables operate can be represented by a *P*-dimensional vector of (possibly unknown) environment-variables, $\underset{\sim}{\xi}$. The observer will be interested in responses that occur in some specific set of environmental situations which will be indexed by $i = 1, 2, ... I$. In each situation i, the probability of a particular response by a user can be represented as a manifest random variable X_i having M_i possible response alternatives:

$$\Pr(X_i = x_i) = f_i(x_i;\underset{\sim}{\xi}|\underset{\sim}{\theta}). \tag{7}$$

To distinguish between conditioning on specific values of random variables and dependence on known constants or unknown parameter values we use the notation $f_i(x_i;\underset{\sim}{\xi}|\underset{\sim}{\theta})$ to indicate that the distribution depends upon $\underset{\sim}{\xi}$ and has been conditioned on the random variable $\underset{\sim}{\theta}$.

We use the notion of a *scoring function* to represent the mapping of parameters of the individual onto situation-response pairs and a *design function* for the mapping of parameters of the environment onto situation-response pairs. For technical purposes, these functions are represented as matrices that specify linear combinations of parameters that are associated with situation-response pairs.

We specify f_i as using a design matrix **A** and a scoring matrix **B**: A user in a task environment agent can be characterized by the situation-response probability model:

$$f_i(x_i = h; A, B, \underset{\sim}{\xi}|\underset{\sim}{\theta}) = \frac{\exp[(\underset{\sim}{b}'_{ih}\underset{\sim}{\theta} + \underset{\sim}{a}'_{ih}\underset{\sim}{\xi})]}{\sum_{u=1}^{M_i}\exp(\underset{\sim}{b}'_{iu}\underset{\sim}{\theta} + \underset{\sim}{a}'_{ih}\underset{\sim}{\xi})}. \tag{8}$$

The scoring matrix **B** allows the description of the score or "performance level" that is assigned to each response alternative on each of the *D* dimensions associated with the user. To do this we introduce the notion of a response score b_{ihd} which gives the score level for dimension *d* of the observed response alternative *h* to situation *i*. The b_{ihd} can be collected in a vector as $\underset{\sim}{b}_{ih} = (b_{ih1}, b_{ih2}, ..., b_{ihD})'$, and the vectors can be collected into a matrix B = $(\underset{\sim}{b}'_{11}, \underset{\sim}{b}'_{12}, \ldots, \underset{\sim}{b}'_{1H_1}, \underset{\sim}{b}'_{21}, \ldots, \underset{\sim}{b}'_{2H_2}, \ldots, \underset{\sim}{b}'_{IM_I})$.

Similarly, the design matrix **A** is used to specify the linear combinations of the *P* task environment parameters $\underset{\sim}{\xi} = (\xi_1, \xi_2, ..., \xi_P)'$ used in the response probability model to describe the behavior of the response categories to each situation. These linear combinations are defined by design vectors $\underset{\sim}{a}'_{ih}$, (i=1,...,I; m=1,...,M_i) which can be denoted collectively by the design matrix A = $(\underset{\sim}{a}'_{11}, \underset{\sim}{a}'_{12}, \ldots, \underset{\sim}{a}'_{1H_1}, \underset{\sim}{a}'_{21}, \ldots, \underset{\sim}{a}'_{2H_2}, \ldots, \underset{\sim}{a}'_{IM_I})'$. Such models have been developed in the context of studies of cognitive development and the learning of complex cognitive skills for Lisp programming (Pirolli & Wilson, in press).

As with the HLM analysis, further development of this approach would require additional research and software development.

ACKNOWLEDGMENTS

We thank David Hull and Marti Hearst for comments on earlier drafts of this note.

As with the HLM analysis, further development of this approach would require additional research and software development.

REFERENCES

Adams, R. J. & Wilson, M. (1992, April). *A random coefficients multinomial logit: Generalizing Rasch models*. Paper presented at the Annual Meeting of the American Education Research Association, San Fransisco, CA.

Atkinson, R. C., Herrnstein, R. J., Lindzey, G., & Luce, R. D. (1988). *Stevens' handbook of experimental psychology*. New York: Wiley,

Campbell, D. T. & Stanley, J. C. (1963). Experimental and quasi-experimental designs for research on teaching. In N. L. Gage (Ed.), *Handbook of research on teaching* (pp. 171-246). Chicago: Rand McNally.

Campbell, N. R. (1928). *An account of the principles of measurement and calculation* London: Longmans, Green, and Co.

Clark, H. H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psyhcological research. *Journal of Verbal Learning and Verbal Behavior, 12*, 335-359.

Halliday, D. & Resnick, R. (1970). *Fundamentals of physics* New York: Wiley.

Krantz, D. H. (1964). Conjoint measurement: The Luce-Tukey axiomatization and some extensions. *Journal of Mathematical Psychology, 1*, 248-277.

Luce, R. D. & Tukey, J. W. (1964). Simultaneous conjoint measurement: A new type of fundamental measurement. *Journal of Mathematical Psychology, 1*, 1-27.

Meyers, J. M. (1966). *Fundamentals of experimental design* Boston: Allyn and Bacon.

Michell, J. (1990). *An introduction to the logic of psychological measurement* Hillsdale, NJ: Lawrence Erlbaum Associates.

Montgomery, D. C. (1984). *Design and analysis of experiments (2nd edition)* New York: Wiley and Sons.

Pirolli, P. & Wilson, M. (in press). A theory of the measurement of knowledge content, access, and learning. *Psychological Review,*

Raudenbush, S. & Bryk, A. S. (1986). A hierarchical model for studying school effects. *Sociology of Education, 59*, 1-17.

Stevens, S. S. (1951). *Handbook of experimental psychology* New York: Wiley.

Winer, B. J. (1962). *Statistical principles in experimental design* New York: McGraw-Hill.