New Measures for the Evaluation of Interactive Information Retrieval Systems: Normalized Task Completion Time and Normalized User Effectiveness

Jing Cheng
Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign
501 E. Daniel Street
Champaign, IL 61820, U.S.A.
jingcheng@gmail.com

Xiao Hu
Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign
501 E. Daniel Street
Champaign, IL 61820, U.S.A.
xiaohu@illinois.edu

P. Bryan Heidorn
School of Information Resources & Library Science, The University of Arizona
1515 East First Street
Tucson, AZ 85719, U.S.A.
heidorn@u.arizona.edu

ABSTRACT
User satisfaction, though difficult to measure, is the main goal of Information Retrieval (IR) systems. In recent years, as Interactive Information Retrieval (IIR) systems have become increasingly popular, user effectiveness also has become critical in evaluating IIR systems. However, existing measures in IR evaluation are not particularly suitable for gauging user satisfaction and user effectiveness. In this paper, we propose two new measures to evaluate IIR systems, the Normalized Task Completion Time ($NT$) and the Normalized User Effectiveness ($NUE$). The two measures overcome limitations of existing measures and are efficient to calculate in that they do not need a large pool of search tasks. A user study was conducted to investigate the relationships between the two measures and the user satisfaction and effectiveness of a given IR system. The learning effects described by $NT$, $NUE$, and the task completion time were also studied and compared. The results show that $NT$ is strongly correlated with user satisfaction, $NUE$ is a better indicator of system effectiveness than task completion time, and both new measures are superior to task completion time in describing the learning effect of the given IR system.

Keywords
Measurement, experimentation, human factors.

INTRODUCTION
Information Retrieval (IR) system evaluation has been studied in the IR research community for decades. There is still a heated debate regarding measures of system evaluation (e.g., Borlund & Ingwersen, 1997; Cooper, 1973; Geva, Kamps, Peters, Sakai, Trotman, & Voorhees, 2009). In addition, as there have been more and more emerging Interactive Information Retrieval (IIR) systems. Evaluations of IR systems consequently need to take into account users’ interactive processes of information searching and retrieval (Beaulieu, Robertson, & Rasmussen 1996; Belkin, 2008; Kelly, 2009). The more effective a system is, the less time a user will need to spend on it to fulfill their information needs. The whole interactive information search process is also the process of users learning to utilize IR systems with increasing proficiency. Consequently, for a given IR system, it is important to support non-expert users to improve their search performance over time.

This study focuses on evaluating IIR systems by proposing two new measures -- Normalized Task Completion Time ($NT$) and Normalized User Effectiveness ($NUE$). These measures strive to evaluate the effectiveness, efficiency as well as the learning effect of IR systems. Here we define learning effect as systems’ support for non-expert users improving search performance over a series of search sessions. Using a standard task completion time and a standard user effectiveness measure as normalization factors, the two new measures aim to overcome deficiencies of existing measures such as precision, recall and task completion time. A user study of a real-world IR system was conducted to examine whether the proposed measures were good indicators of system effectiveness, user satisfaction and learning effect.

RELATED WORK
It has been a half-century since the Cranfield model of laboratory-based tests were invented, and it is still the dominant approach to IR evaluation, with precision and recall still widely used as IR evaluation measures (Borlund, 2003). However, the literature in IR evaluation has reported mixed results regarding whether precision or recall is a
significant factor of user satisfaction towards IR systems (e.g., Turpin & Hersh, 2001; Kelly, Fu, & Shah, 2010). In addition, precision and recall are insufficient for evaluating IIR systems, because while users may modify or develop search queries and strategies during search processes, the two measures cannot quantify the “informativeness” of interactions (Borlund & Ingwersen, 1997).

Cooper (1973) and other researchers believe that an ideal evaluation methodology must somehow measure the ultimate worth of a retrieval system to its users in terms of an appropriate unit of utility. But both user satisfaction and utility are difficult to define and quantify.

Cleverdon and Keen (1966) listed time lag (response time of the system to user requests) as one of the six important criteria that could be used to evaluate IR systems. Su (1992) stated efficiency-time as the most important search success category reported by users (also see Jones, 1997).

With the development of IIR systems, researchers then started to pay attention to the dependency between task completion time and interface. Dunlop (1997) proposed a new measure, called the expected search duration and established an interface-based predicted-time model, which measured how long it took to view a set of documents before a number of relevant ones were found. This model used task completion time as a criterion and integrated user interface into IR system evaluation.

Researchers also began to realize the influence of learning in the IR process. Allen (1994) studied the relationship between “perceptual speed,” “learning,” and “information retrieval performance” by end users. He found that information system usability was determined by interactions between characteristics of the users and system features. Borlund (2003) also argued that IR evaluation should pay attention to the learning process of users and take into account modifications made by users during the IR process. Therefore, it is desirable to have measures that can gauge the learning effect of an IR system (i.e., the support for non-expert users becoming proficient users) and its effectiveness at the same time.

**Deficiencies of Existing Measures**

*Precision* and *recall* are the most popular measures of IR evaluation, but researchers (e.g., Su, 1992; Hearst, 2009) have found that users may not always care about precision or recall. For example, depending on their information seeking tasks, users may not be concerned about retrieving all the documents relevant to their search tasks; for many users, they are happy if they can get a good answer in a short amount of time. In other cases, users do not only care about precision and recall; other attributes of the system, user and task may also play important roles in the search process. Therefore, additional measures are needed to better understand the interactive nature of IIR.

It is commonly agreed that user satisfaction is the main goal of IR systems. However, it is difficult to precisely define “satisfaction” or objectively evaluate systems by subjective feelings reported by users (Belkin & Vickery, 1985).

According to Su (1992), users consider task completion time as critical to successful information retrieval, and time-based measures are often used in Web searches (Xu & Mease, 2009) and IIR evaluations (Kelly, 2009). Here we formally define the task completion time as the time period from when the user begins a search to the point he or she stops the search. The task completion time is not only determined by an IR system’s parameters, e.g., the system’s response speed, its user interface, and ranking of results, but it is also influenced by the users and the complexity of search tasks (also called “task difficulty” (Gwizdka, 2008)). This makes the task completion time incomparable across different systems, users and search tasks.

To reduce users’ influence on the task completion time, researchers normally divide users into different groups, based on their knowledge background. To reduce the influence of task difficulty, researchers often permute search tasks through a Latin-Square arrangement to ensure all search tasks were performed in different orders by the users. Note that this method only works when the sample pool of search tasks is large. Should we have a standard task completion time, e.g., the time taken by an expert, as a reference, the influence of task difficulty could be eliminated without requiring a large task sample pool. In this way, the cost of IR evaluation can be greatly reduced. The two measures proposed in this study are based on a standard task completion time and thus have the merit of being more cost-effective than task completion time.

**NEW MEASURES AND HYPOTHESES**

**Definition of NT and NUE**

The new measures—NT and NUE—are proposed to solve the aforementioned deficiencies of existing measures, using a standard task completion time and a standard user effectiveness measure as normalization factors. NT is defined as:

\[
NT_i = \frac{t_e}{t_{ni}} \quad (1)
\]

Where \(t_e\) is the average time expert users spend on a search task, \(t_{ni}\) is the time that a novice user, \(i\), spends on the search task.

Unlike the task completion time, NT is not an absolute value of completing a task, but the proportion of time that a “perfect” user spends on one single search task relative to the time that a novice user spends on the same search task. We define the “perfect” user using three criteria: 1) being good at searching for information in general and thus able
to correctly interpret search tasks and employ sound search strategies; 2) having the best knowledge of a given IR system and thus be able to choose the optimal way to do a search using this system; 3) having the best expertise in the domain of a given search task and thus be able to create good queries and assess relevance for the task. In the real world, the “perfect” user does not exist, but the time spent by a “perfect” user can be approximated by aggregating the time spent by a number of expert users. We use the mean as the aggregation function in order to give each expert an equal weight. Therefore, supposing there are m expert users, formula (1) can then be written as:

\[ NT_i = \frac{\sum_{k=1}^{m} t_{ik} / m}{t_{ni}} \]  

(2)

For example, let us compare two systems. If a novice user spends 2 minutes conducting one search task by using system 1, while the average time of expert users is 1 minute, then the NT of system 1 would be 0.5; if the novice user spends 1.5 minutes by using system 2, and the average time of expert users is still 1 minute, the NT of system 2 would be 0.67. Based on the sample data, we may conclude that system 2 is better than system 1 for the novice user.

By dividing the time spent by the “perfect” user (i.e., a standard task completion time), NT is expected to be more stable than the original task completion time when the difficulty level of the search tasks varies. For instance, if one search task is very difficult, both expert users and novice users may spend a long time searching for it, so the task completion time will be longer than that of conducting other search tasks using the same system. Here, NT will not change as much as the task completion time, because the numerator and the denominator are both increasing. The same applies if a search task is easy. The task completion time may be very small, while the NT will still be relatively stable. This feature of NT (as well as NUE, see below) focuses on the evaluation of IR systems and avoids distractions by the difficulty levels of search tasks, which cannot be easily controlled.

Despite the aforementioned advantages, NT is, in fact, a simplified calculation. Although a novice user spends more time than an expert user, the quality of the former’s search results may still not be as good as that of the latter’s. The proportion of relevant documents retrieved by the two users can indicate the difference of the effectiveness of their searches. In the above example, using system 1, the novice user may have retrieved 5 relevant documents while the expert user retrieved 10; and using system 2, the novice user may have retrieved 3 relevant documents, while the expert user retrieved 8. In this case, the difference in search quality between the two users cannot be reflected by NT. To solve this problem, NUE is proposed:

\[ NUE_i = NT_i \times \frac{\sum_{k=1}^{m} I_{ek} \times I_{ni}}{\sum_{k=1}^{m} I_{ek} / m} \times t_{ni} \sum_{k=1}^{m} I_{ek} \]  

(3)

where \( I_{ek} \) stands for the number of relevant results retrieved by the novice user, \( i \), and \( I_{ek} \) stands for the number of relevant results retrieved by the expert user, \( k \).

The name of NUE comes from user effectiveness. Al-Maskari & Sanderson (2010) defined user effectiveness as “the accuracy and completeness with which users achieve certain goals” (p. 860) and measured it by the number of relevant documents retrieved and the time users take to complete the task. NUE is the multiplication of NT and the number of relevant documents retrieved by a user normalized by the number of relevant documents retrieved by a “perfect” user. Again, the “perfect” user is represented by the average of a number of expert users. In this study, the notion of binary relevance is adopted, i.e., a document is either relevant or irrelevant to a search task. According to the definition, in the above example, the NUE of system 1 is 0.25, and the NUE of the system 2 is also 0.25. In terms of NUE, the two systems are the same.

NUE contains more information than NT, as it also measures the effectiveness of retrieval by calculating the normalized number of retrieved relevant documents. NUE is specifically useful when recall is critical for users, i.e., when users need to retrieve as many relevant documents as possible (e.g., when doctors may want to find all prescriptions for one disease).

As mentioned above, because NT and NUE are not affected by the complexities of different search tasks as much as the task completion time is, calculating them only requires a small sample pool of search tasks. This efficiency in calculation is also an advantage over task completion time.

**Hypotheses on NT and NUE**

This study aims to find metrics that can evaluate both the effectiveness and the efficiency of IR systems, while being a good indicator of user satisfaction. Meanwhile, as learning is an indispensable part of an IR process, the metrics also aim to measure systems’ learning effect (i.e., their support for non-expert users improving search performance over search sessions). This section states the four hypotheses examined in this study.

**Hypothesis 1:** NT and NUE are highly correlated to user satisfaction.

As user satisfaction is the main goal of IR systems, it is desirable to have measures with strong correlations to user satisfaction. We expect NT and NUE are such measures.
**Hypothesis 2**: $NT$ and $NUE$ are better measures of user satisfaction than the task completion time, as defined by fit to regression models.

Due to the aforementioned advantages of $NT$ and $NUE$ over the task completion time, we also expect that $NT$ and $NUE$ will have a higher correlation to user satisfaction than the task completion time.

**Hypothesis 3**: $NT$ and $NUE$ are highly correlated to the $F$ measure.

To investigate whether $NT$ and $NUE$ are good measures for system effectiveness, we examine the correlations between each of the two new measures and the $F$ measure, a combination of precision and recall that is commonly used to measure the effectiveness of IR systems.

The $F$ measure is formally defined as:

$$F = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}.$$ 

If $NT$ and $NUE$ can evaluate the effectiveness of an IR system, they should be strongly correlated with the $F$ measure.

**Hypothesis 4**: $NT$ and $NUE$ are better than the task completion time in measuring the learning effect of an IR system.

**Learning effect** can be depicted by a learning curve, which is a graphical representation of search performances over a series of search sessions. In the aeronautical industry, learning curves are based on repeating tasks performed by a worker, and task completion time is a common measure to study learning curves (e.g., Wright, 1936). However, this is not the same situation in IR research where there is no point asking a user to repeatedly search for the same task. Thus, different search tasks are assigned to the same users to test a system. Therefore, the task completion time is very likely to be influenced by the difficulty levels of search tasks. It was hypothesized that $NT$ and $NUE$ can measure the learning effect of an IR system better than the task completion time, because, as stated in last section, $NT$ and $NUE$ are expected to be more stable than the task completion time when the difficulty level of search tasks varies.

**EXPERIMENTS**

**Research Settings**

An experiment was conducted to examine the proposed hypotheses by comparing $NT$ and $NUE$ with other selected measures. This experiment was conducted on a state university campus from January 2007 to March 2007. The details of the experimental settings are explained in the following sections.

**Testing System**

In this study, LexisNexis Academic was the selected testing system. As a commercial database, though it is accessible to novice users, it is mostly used by experienced users (e.g., information professionals) because this powerful database is not as easy to use as Web search engines (e.g., Google) for the general public. Therefore, this system is desirable for recruiting both novice and expert users. LexisNexis Academic provides comprehensive coverage of five sections: news, business, legal information, medical information and reference publications, with full text and abstracts.

The search tasks assigned to participants covered all of the above five sections. Subjects were asked to use the Guided News Search, rather than the Quick News Search option, because the latter one is too general and could return too many documents.

**Subjects**

There were a total of 20 subjects, including 15 novice users and 5 experts who voluntarily participated in this experiment on the campus of the state university. The subjects were selected based on their familiarity with the LexisNexis Academic search engine. The expert users were university librarians who used LexisNexis Academic extensively. The novice users were staff, undergraduate and graduate students at the university. Their majors varied from science and social science to humanities. Among the 15 novice subjects, 7 were male and 8 were female. The subjects were given an introduction letter specifying the experimental procedure, and the researchers reiterated the procedure to the subjects before starting the experiment(s).

Subjects were assigned an ID code as their identity. Before the novice subjects started searching, they were asked to answer an entry-questionnaire to make sure they were qualified for this research, i.e., they had some knowledge of the Internet, but had not used the testing system before. The novice users all received basic training on how to use this system prior to performing searches. The training lasted for about 10 minutes. They were then asked to study the instructions for LexisNexis Academic by themselves. The researchers answered their questions if they had any.

**Search Tasks**

Each participant was assigned 10 search tasks. These search tasks were based on real-life reference questions from the university library. The 10 search tasks covered the five major sections of LexisNexis Academic (see the list of search tasks below). The search tasks were not domain-specific, and thus the last criterion of a “perfect” user was relaxed in this experiment. Specifically, the “perfect” user was represented by 5 experts (librarians) who were experienced in searching for information, were very familiar with the testing system, but were not required to be experts in any specific domain.
The researchers ran a pilot study before finalizing the search tasks to make sure that LexisNexis Academic could retrieve the relevant documents for all these search tasks. Each search task was assigned an ID code, and the 10 tasks were given to the subjects in different orders distributed by a Latin-Square arrangement to reduce the influence of the search tasks’ difficulty levels on the task completion time. For example, the following are the search tasks given to one subject, who conducted the searches in the listed order from 1 to 10. The numbers in parentheses were the search task IDs, e.g., the first search task given to this subject was search task 4 (Q4).

1. Who are the announced Democratic 2008 Presidential candidates? (Q4)
2. Find an obituary for Princess Diana from the Washington Post. (Q9)
3. Find reviews of the movie Brokeback Mountain. (Q10)
4. According to recent research, which vitamin is implicated in skin cancer development with the intake of folate? (Q3)
5. Find data on the murder rates for each state in the USA after 2000. (Q7)
6. How many points did Yao Ming score in the game after his return from his most recent injury? (Q6)
7. Who was the auditor of Best Buy Co. in 2004? (Q1)
8. Find articles about President Bush’s most recent nominee after his return from his most recent injury? (Q6)
9. What are keywords that may result in a webpage being blocked by the Chinese government? (Q2)
10. What is the phone number for Google’s main corporate office? (Q8)

**Measures**

To test the proposed hypotheses, this study compared the following seven measures:

1. **Precision (P):** The proportion of retrieved documents that are relevant, where
   \[ P = \frac{\text{the number of retrieved relevant documents}}{\text{the number of retrieved documents}} \]
2. **Recall (R):** The proportion of relevant documents that are retrieved, where
   \[ R = \frac{\text{the number of retrieved relevant documents}}{\text{the number of relevant documents}} \]

It is difficult to know how many relevant documents for a search task exist in a large-scale IR system like LexisNexis Academic. Because of this well-known difficulty of calculating recall, a modified relative recall was used to substitute. We followed the “pooling” method used in the Text Retrieval Conference (TREC) (Voorhees & Harman, 2001). For a given search task, all documents retrieved by the 20 users were merged into a single set, and only those documents were judged on relevance. Documents that were not retrieved by any of the subjects were not considered. We employed one assessor to make relevance judgments, because the experience in TREC indicated that judgments made by a single judge and group judgments did not affect the relative performance of the retrieval systems (Voorhees, 1998). The assessor in this experiment was an IR researcher in the library school at the state university and was familiar with the nature and issues of relevance judgments.

3. **F measure (F):** The weighted harmonic mean of precision and recall, also called the traditional F-measure or the balanced F-score, where
   \[ F = \frac{2 \times (\text{precision} \times \text{recall})}{\text{precision} + \text{recall}} \]
4. **Task completion time (T):** The time from the start until the completion of a retrieval task. The task completion time is recorded in seconds, but note that the minute was the unit in the calculations. The subjects stopped each search session by themselves, either once their search needs were satisfied or when they gave up.
5. **User satisfaction (S):** An ordinal number indicating the level of user satisfaction. A questionnaire was provided to the users after each search was complete. It asked the subjects to rate their satisfaction level towards the system in regard to supporting the accomplishment of the search task, using a scale from 1 (not satisfied at all) to 5 (extremely satisfied).

6. **NT (see the “Definition of NT and NUE” section)**
7. **NUE (see the “Definition of NT and NUE” section)**

**Data Collection**

The whole experimental process lasted between 45 minutes to 1.5 hours for each subject. For each search task, the subject stopped the search session when either the information need was satisfied or he or she gave up the task. The researchers observed the subjects performing their searches in order to ensure that the correct procedures were being followed. The researchers also recorded the task completion time of each task, and asked the subjects to answer a questionnaire after each task. As a subject may modify queries during a search session, the retrieved documents were defined as those displayed on the screen in response to the last query in a search session. As required in the instructions of the experiment, the subject sent the researchers his or her search results by e-mail at the end of each search session. After gathering the e-mails from all subjects, the relevance of all the retrieved documents was judged by the independent assessor.
RESULTS
The 10 novice users and the 5 expert users each performed the 10 search tasks, and thus there were 150 search sessions. Among all the 150 search sessions, 7 were unsuccessful in that no relevant documents were retrieved. In these cases, the users gave up the task after a few futile attempts. As the tasks were never completed, the task completion time in these cases would be infinite, and thus the values of NT and NUE would be 0. For 18 other sessions, the users completed their search when they believed they had retrieved relevant documents, but in fact, they did not retrieve any relevant documents. For these cases, we still used the time they spent on the sessions as their task completion time and calculated NT accordingly. However, the value of NUE would be 0 because the number of retrieved relevant results was 0. For a comparison of statistics of novice users and expert users, Table 1 lists the average task completion time and the numbers of retrieved relevant documents of the two user groups on each search task.

Statistics Analysis on User Satisfaction
To test the first hypothesis, NT and NUE are correlated with user satisfaction; correlation analysis and regression were applied. The correlations between user satisfaction and other measures, such as precision, recall and task completion time, were also studied as comparisons to NT and NUE. Here, user satisfaction is the dependent variable, while task completion time, precision, recall, NT and NUE are the independent variables. The sample size for the analysis is 150, where each of the 15 novice users conducted searches on each of the 10 search tasks.

Search Tasks Correlation Matrix
The degree of user satisfaction has five levels, from 1 (not satisfied at all) to 5 (extremely satisfied). Spearman’s rank correlation coefficient was selected for the correlation analysis because it can be applied to ordinal variables such as user satisfaction. Unlike the more commonly used Pearson correlation coefficient, the Spearman’s rank correlation coefficient does not require the assumption that the relationship between the variables is linear, nor does it require the variables to be measured on interval scales.

Table 2 presents the Spearman’s rank correlation coefficients for all the variables we studied. It shows that NT had the strongest correlation with user satisfaction among all the tested measures, followed by task completion time (T). However, the correlation between NUE and user satisfaction could only be described as moderate at best. This result partially supports the first hypothesis that NT and NUE are highly correlated to user satisfaction. As a comparison, it is noteworthy from Table 2 that precision (P) and recall (R) were weakly correlated with user satisfaction (S).

<table>
<thead>
<tr>
<th>Tasks</th>
<th>Avg. by Novice Users</th>
<th>Avg. by Expert Users</th>
<th>Avg. # of Ret. Docs by Novice Users</th>
<th>Avg. # of Ret. Docs by Expert Users</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q1</td>
<td>470</td>
<td>123</td>
<td>5.6</td>
<td>2.6</td>
</tr>
<tr>
<td>Q2</td>
<td>605</td>
<td>155</td>
<td>2.3</td>
<td>4</td>
</tr>
<tr>
<td>Q3</td>
<td>275</td>
<td>112</td>
<td>15.5</td>
<td>24</td>
</tr>
<tr>
<td>Q4</td>
<td>496</td>
<td>84</td>
<td>3</td>
<td>6.6</td>
</tr>
<tr>
<td>Q5</td>
<td>444</td>
<td>133</td>
<td>3.3</td>
<td>7</td>
</tr>
<tr>
<td>Q6</td>
<td>365</td>
<td>147</td>
<td>1.5</td>
<td>2.2</td>
</tr>
<tr>
<td>Q7</td>
<td>435</td>
<td>63</td>
<td>3.2</td>
<td>2</td>
</tr>
<tr>
<td>Q8</td>
<td>224</td>
<td>47</td>
<td>33.4</td>
<td>47.4</td>
</tr>
<tr>
<td>Q9</td>
<td>511</td>
<td>158</td>
<td>0.4</td>
<td>1</td>
</tr>
<tr>
<td>Q10</td>
<td>117</td>
<td>100</td>
<td>11.6</td>
<td>23.2</td>
</tr>
<tr>
<td>Total</td>
<td>3942</td>
<td>1122</td>
<td>78.9</td>
<td>120</td>
</tr>
</tbody>
</table>

Table 1. Statistics of Two User Groups across Search Tasks

<table>
<thead>
<tr>
<th></th>
<th>S</th>
<th>T</th>
<th>P</th>
<th>R</th>
<th>NT</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>-.61128</td>
<td>&lt;.0001</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P</td>
<td>.25416</td>
<td>.9633</td>
<td>.0160</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R</td>
<td>.22116</td>
<td>-.5786</td>
<td>.36197</td>
<td>&lt;.0001</td>
<td></td>
</tr>
<tr>
<td>NT</td>
<td>.69096</td>
<td>-.86010</td>
<td>.0209</td>
<td>.22974</td>
<td>.0047</td>
</tr>
<tr>
<td>NUE</td>
<td>.44655</td>
<td>-.52272</td>
<td>.40003</td>
<td>.71952</td>
<td>.59704</td>
</tr>
</tbody>
</table>

Note: N = sample size

Table 2. Spearman Correlation Matrix for Covariates of User Satisfaction (N=150)

Regression Models
Five multinomial logistic regression models were run to further study the relationships between user satisfaction and the five measures respectively. As an ordinal variable, user satisfaction does not obey the assumption that the independent and dependent variables have a linear relationship. Thus the multinomial logistic regression model was used, for it assumes neither a linear relationship, nor does it require normally distributed variables. Based on our sample size (N=150), we set the significance level at
0.01, i.e., two variables are not recognized as significantly correlated to each other unless \( p \leq 0.01 \).

A summary of the five regression models is presented in Table 3. Based on these models, NT, NUE and the task completion time were all significantly correlated with user satisfaction at the \( p \leq 0.0001 \) level. According to the log likelihood of the three models, the model of NT had the best fit, i.e., NT was the best predictor of user satisfaction, while NUE was not better than task completion time to predict user satisfaction. Hypothesis 2, which posited that NT and NUE are better measures of user satisfaction than task completion time, was partially supported.

According to Table 3, precision and recall were not significantly correlated with user satisfaction \( (p > 0.01) \), which indicated that the two measures were not good indicators of user satisfaction. This is in accordance with the weak correlations between precision, recall and user satisfaction found in Table 2.

Statistics Analysis on F Measure
To test hypothesis 3, three linear regression models were used to study the relationships between the F measure and each of NT, NUE, and the task completion time (shown in Table 4). Unlike user satisfaction, which is an ordinal variable, F measure, NT, NUE, and task completion time are interval variables, and thus the linear regression model is suitable for them.

According to the three models, NT and the task completion time were not significantly correlated with the F measure \( (p > 0.01) \), while NUE has a significant correlation with the F measure \( (p < 0.01) \). This finding partially supports hypothesis 3 that states that NUE is correlated with the F measure, but rejects that NT is correlated with the F measure.

Analysis on Learning Curves
The learning curve discloses the relationship between users’ experiences with the system and search performance. As individuals get more experienced at using a system, their search performances usually improve. In IR evaluation, we can compare the slopes of different systems’ learning curves to study the learning effect of IR systems. The steeper the slope, the better the system can support its novice users improving their search performances over time.

A learning curve was drawn using the 15 novice users’ average task completion time of each search task. The x-coordinate is the successive sequence of the 10 search tasks conducted by the novice users. As mentioned in the experiment design section, the search tasks were permuted through a Latin-Square arrangement, such that each subject had a unique sequence of search tasks. Figure 1-3 show the learning curves explained by the task completion time, NT, and NUE respectively.
and the remaining 51.58% could be explained by unknown lurking variables or inherent variability. For example, the fluctuation in the middle part of the curve might be due to changes of interest or concentration level of the human subjects in the middle of the experiment. In any case, the curve shows task completion time is a weak measure for learning effect.

The learning curves on NT and NUE (Figures 2, 3) show that the more search tasks the users practiced, the higher the NT and NUE were, i.e., the novice users became closer to the experts. This also meets the deduction of learning curves. According to $R^2$, 74.32% of the variation in the response variable can be explained by NT, and 76.11% of it can be explained by NUE. NT and NT explain the learning curves much better than the task completion time. Therefore, hypothesis 4 was supported: NT and NUT are better measures for describing systems’ support for novice users improving search performance over search sessions.

DISCUSSION ON RESULTS
Based on the experiment results, we conclude that: NT was strongly correlated with user satisfaction and was a better indicator of user satisfaction than task completion time; NUE was a better indicator of system effectiveness (represented by the F measure) than task completion time; and the two proposed measures were better than the task completion time in depicting the learning effect over the search sessions. The experiment results also confirmed that precision and recall had weak correlations with user satisfaction and were not good indicators of user satisfaction. This is consistent with previous studies that found no substantial relationship between system effectiveness and user satisfaction (e.g., Turpin & Hersh, 2001). Also, the task completion time was not significantly correlated with system effectiveness (F measure), based on our linear regression models. In addition, NT and NUE described the system’s learning effect much better than task completion time. Therefore, we conclude that, for factual search tasks in the general domain (such as the ones in this experiment), NT and NUE can be used together as overall measures in evaluating IR systems in terms of user satisfaction, effectiveness and learning effect.

Calculating NT and NUE requires participation of expert users, but it only requires a small sample pool of search tasks. This is because NT and NUE are not affected by the complexities of different search tasks as much as the task completion time is. Therefore, NT and NUE are cost effective measures compared to task completion time.

CONCLUSIONS AND FUTURE WORK
User satisfaction, system and user effectiveness, as well as support for users’ improvement are all critical to IIR systems evaluation. However, existing measures such as precision, recall and task completion time are not sufficient for measuring these critical aspects. In particular, precision and recall completely leave users out of the picture while task completion time, to a large extent, depends on search tasks and users. This study proposes two new measures, NT and NUE, which can overcome these limitations of existing measures by comparing novice users’ searches to those of experts. A user study was conducted to confirm that NT is good for measuring user satisfaction of IR systems, NUE is a good measure for system effectiveness, and both are good for describing systems’ learning effect. A single NT or NUE can evaluate the user effectiveness of a system, and a series of successive NT or NUE can examine the system’s support for novice users improving search performances by
observing the learning curve slope (i.e., how fast a novice user can catch up to an expert user when using the same system). In summary, we believe that NT and NUE are complementary to existing measures.

It is noteworthy that the search tasks in this experiment are domain general, and thus future studies are needed to validate whether the conclusions are applicable to other domain-oriented search tasks.

To confirm the generalizability of the two measures, a comparison of two or more IR systems is needed, and a larger sample of testing search tasks and subjects is also desirable in the future.

ACKNOWLEDGMENTS
We thank the anonymous reviewers for their suggestions and help on improving this paper.

REFERENCES


