**To be published in the Proceedings of ACM Symposium on Empirical Software Engineering and Measurement (ESEM) Lake Buena Vista, FL, 2009**

# Usability Testing with Total-Effort Metrics

Liam Feldman, Carl J. Mueller, Oleg V. Komogortsev, Dan Tamir
Texas State University—San Marcos
{lf1081, cm58, ok11, dt19}@txstate.edu

## Abstract

Usability testing activities have numerous benefits in theory, yet they are often overlooked or disregarded in practice. A testing paradigm which yields objective, quantitative results would likely lead to more widespread adoption of usability evaluation activities. Total-Effort Metrics is such a novel framework. This paper describes a usability study conducted using a total-effort metrics approach. In this study, subjects interact with three interfaces which have varying element layout proximities. The time and effort measures of time-on-task, total keystrokes, correctional keystrokes, saccade amplitude (point-to-point eye movement) and gaze-path traversal are recorded and analyzed. The findings of the study demonstrate a correlation between the intrinsic effort of an interface and its usability as predicted by extant interface layout guidelines.

## 1. Introduction

A survey conducted in 2007 determined that software engineers routinely neglect to perform usability testing as part of their development process. A majority of developers regard usability evaluations as being unhelpful, while a minority find them to be valuable—yet let them fall by the wayside anyway [14]. This is a curious state of affairs given that usability is a fundamental characteristic of software quality, one which affects not only overall quality, but safety concerns as well [4]. Usability strongly correlates with a product's perceived salability, reputation, supportability, training and documentation expenses, and potential for adverse legal action [10].

A vast body of literature exists pertaining to usability design best practices, but metrics which provide insight into the usability of an interface are rarer [10]. Data derived from usability evaluation is by-and-large of a qualitative nature. As it is currently practiced, usability testing consists of activities like heuristic (i.e. expert) evaluation, walkthroughs, and predictive modeling. These methods yield valuable insights, but suffer from a lack of objectivity. Even logging-actual-use methods, which record user interactions with software systems in the field, require expert interpretation [2].

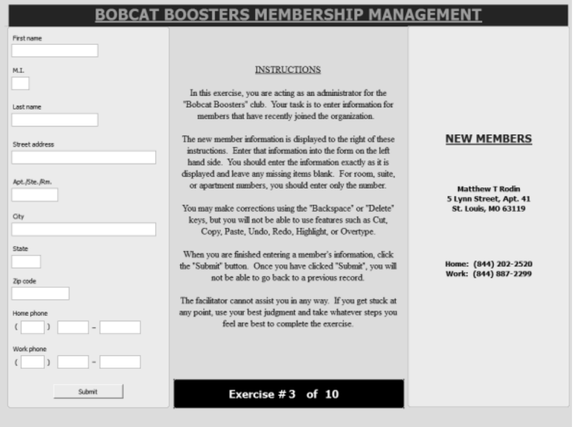
Total-Effort Metrics (TEM) is a novel usability testing methodology that yields elegant, quantitatively-expressed insights into the usability characteristics of software systems. It is an analysis framework that integrates commonplace methods with measurements that are not yet universally utilized, but have high applicability to usability testing.

This paper presents one of a series of studies demonstrating the use of a total-effort measurement methodology as a usability verification and validation tool. The current study examines how variance in interface element placement affects the total effort necessary to enter data into form fill-in interfaces. This study is in effect an effort-based validation of the "Law of Proximity." Derived from Gestalt psychology, this design guideline dictates that closely-grouped elements are perceived by users as belonging to a single unit, therefore related elements in an interface ought to be placed in proximity to each other [9].

The general notion of effort as a driver of intrinsic usability is not foreign to testing literature. Bevan, for example, cites ISO/IEC 9126 (a predecessor to ISO/IEC 9126-1) in defining software usability as, "A set of attributes that bear on the *effort* needed for use, and on the individual assessment of such use, by a stated or implied set of users [1]." Jones' definition of usability is in accord with Bevan's: "Usability is the *total effort* required to learn, operate and use software or hardware [5]." Tamir, Mueller, and Komogortsev have proposed a total-effort model of usability based on time-to-task accomplishment, direct measures of physical effort, and indirect measures of cognitive effort as indicated by eye movements [12]. The total effort-metric equations used in this study are adapted from work by Komogortsev et al. [6][7].

## 2. Test protocol

In the study presented in this paper, subjects were asked to complete simple form fill-in/data-entry tasks. These tasks consisted of copying various pieces of information for fictitious customers displayed on-screen into corresponding textbox fields. The test application logged keystroke, mouse movement, mouse click, and time-on-task data for each subject and set of tasks. A Tobii X120 eye-tracking camera logged eye-movements.



BOBCAT BOOSTERS MEMBERSHIP MANAGEMENT

INSTRUCTIONS

In this exercise, you are acting as an administrator for the "Bobcat Boosters" club. Your task is to enter information for members that have recently joined the organization.

The new member information is displayed to the right of these instructions. Enter that information into the form on the left hand side. You should enter the information exactly as it is displayed and leave any missing items blank. For room, suite, or apartment numbers, you should enter only the number.

You may make corrections using the "Backspace" or "Delete" keys, but you will not be able to use features such as Cut, Copy, Paste, Undo, Redo, Highlight, or Overtype.

When you are finished entering a member's information, click the "Submit" button. Once you have clicked "Submit", you will not be able to go back to a previous record.

The facilitator cannot assist you in any way. If you get stuck at any point, use your best judgment and take whatever steps you feel are best to complete the exercise.

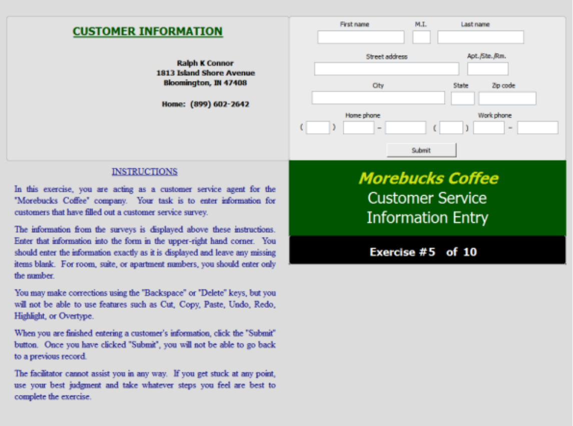
NEW MEMBERS

Matthew T Rodin
5 Lynn Street, Apt. 41
St. Louis, MO 63119

Home: (844) 202-2520
Work: (844) 887-2299

Exercise # 3 of 10

Figure 1. Form A



CUSTOMER INFORMATION

Ralph K Connor
1813 Island Shore Avenue
Bloomington, IN 47408

Home: (899) 602-2642

INSTRUCTIONS

In this exercise, you are acting as a customer service agent for the "Morebucks Coffee" company. Your task is to enter information for customers that have filled out a customer service survey.

The information from the surveys is displayed above these instructions. Enter that information into the form in the upper-right hand corner. You should enter the information exactly as it is displayed and leave any missing items blank. For room, suite, or apartment numbers, you should enter only the number.

You may make corrections using the "Backspace" or "Delete" keys, but you will not be able to use features such as Cut, Copy, Paste, Undo, Redo, Highlight, or Overtype.

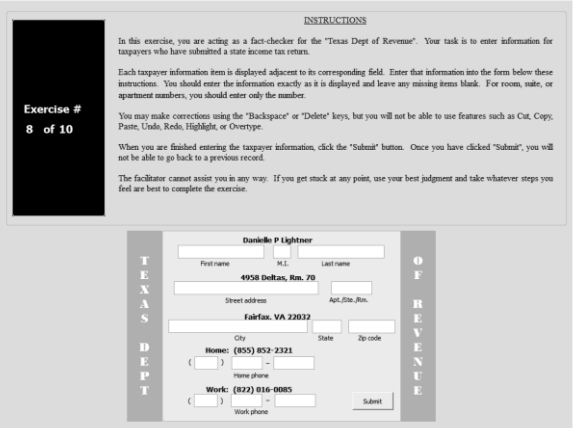
When you are finished entering a customer's information, click the "Submit" button. Once you have clicked "Submit", you will not be able to go back to a previous record.

The facilitator cannot assist you in any way. If you get stuck at any point, use your best judgment and take whatever steps you feel are best to complete the exercise.

Morebucks Coffee
Customer Service
Information Entry

Exercise #5 of 10

Figure 2. Form B



Exercise #
8 of 10

INSTRUCTIONS

In this exercise, you are acting as a fact-checker for the "Texas Dept of Revenue". Your task is to enter information for taxpayers who have submitted a state income tax return.

Each taxpayer information item is displayed adjacent to its corresponding field. Enter that information into the form below these instructions. You should enter the information exactly as it is displayed and leave any missing items blank. For room, suite, or apartment numbers, you should enter only the number.

You may make corrections using the "Backspace" or "Delete" keys, but you will not be able to use features such as Cut, Copy, Paste, Undo, Redo, Highlight, or Overtype.

When you are finished entering the taxpayer information, click the "Submit" button. Once you have clicked "Submit", you will not be able to go back to a previous record.

The facilitator cannot assist you in any way. If you get stuck at any point, use your best judgment and take whatever steps you feel are best to complete the exercise.

Danielle P Lightner
4958 Deltas, Rm. 70
Fairfax, VA 22032
Home: (855) 852-2321
Work: (822) 016-0085

Figure 3. Form C

Each subject interacted with three different interface form factors. Elements in the “Form A” interface, as shown in Figure 1, were placed so as to maximize the distance between the display of data to be entered and the actual data-entry fields. “Form B,” shown in Figure 2, placed the data-entry display a short distance away from the data-entry fields. “Form C,” shown in Figure 3, interleaved the display of each data element with its corresponding entry field. The order of form factors presented to Group I was reversed from Group II so that each group served as a control for the other, particularly with regard to factors of fatigue, motivation, and learning.

Subjects for this study were volunteers recruited from a population of undergraduate and graduate students in the Computer Science/Software Engineering program at Texas State University–San Marcos. 11 subjects in total completed test sessions: Nine men and two women ranging in age from 20 to 29 years old, with an average age of 24.6 years old, standard deviation ±2.8 years. Test subjects as a whole reported weekly computer usage averaging 45.2 ±17.3 hours and mean weekly Internet/WWW usage of 28 ±17.1 hours. Stated word-processor usage averaged 11.8 +12.3/-11.8 hours per week, while database and spreadsheet usage had a mean of 5.1 +10.1/-5.1 hours per week. Eight subjects indicated that they are “touch typists” (i.e. able to type without looking down at the keyboard). Four subjects reported having learned English as a secondary language.

## 3. Results and analysis

Looked at in isolation, each category of data captured by this study – qualitative, time-on-task, keystroke count, correctional keystrokes, and eye movements – provides useful but limited insight into the usability aspects of the interfaces tested. The timing data captured by the test application, in combination with the qualitative information gathered, indicate that Form A has some indeterminate efficiency issue, while Form C is superior in terms of efficiency-of-use. Logged keystroke data further indicate that Form A inhibits usage effectiveness whereas Form C allows tasks to be accomplished more effectively. Keystroke and time-on-task (i.e. time required by subject to complete a given task) data indicate the presence of a usability issue, but provide no indication as to precisely what the nature of the issue is. On the other hand, when the eye-tracker data is added into the picture, an explanation for the underlying usability issues of Form A becomes clear.

The qualitative data gathered in the course of testing do not by themselves provide a clear picture as to which of the three evaluated interfaces are most usable, much less why one is more or less usable than the other. Qualitative ratings for the three interfaces mostly conformed to the research hypothesis that user perceptions of usability, learnability and satisfaction would increase as element layout proximity decreased. Subjects rated Form C, the form with interleaved data to be entered and data entry fields, as being the most usable and satisfying to use. Form A, the form which maximized the distance between data to be entered and data entry fields, was rated as being the least usable and satisfying to use. Subject ratings of learnability did not conform to expectations; subjects rated Form B, the intermediate-distance form, as being the most learnable.

It was expected that subjects would rate Form A as involving the most discomfort and exertion to use, using Form B would be rated more comfortable and less effort-intensive to use than Form A, and Form C would be rated with a perception of the least amount of discomfort and exertion. Subjects did rate Form A as inducing the most discomfort, but also as involving the least amount of physical exertion. Form B was rated as involving the least amount of mental exertion.

Time-on-task data, i.e. "stopwatch" data, are a staple of conventional usability evaluation methods [2][9][13]. The information captured regarding time-on-task provides a somewhat better indicator of which interfaces exhibit usability issues. Time-on-task measures, it should be noted, tend to exhibit a decreasing slope as subject familiarity with identical or similar task scenarios increases [11]. Thus a null hypothesis for comparing the time-on-task results for Groups I and II is that time-on-task for each group will decrease at a uniform rate.

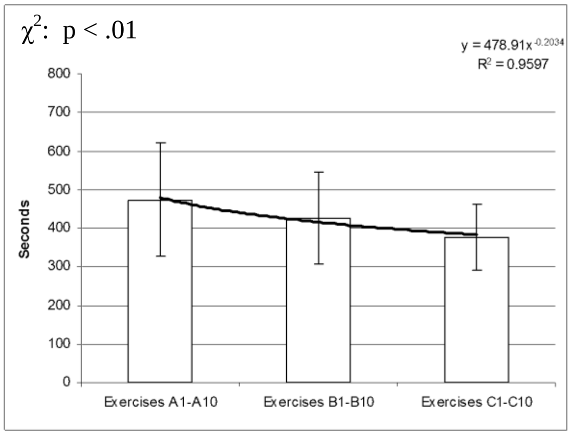


**Figure 4. Mean Aggregate Time-On-Task, Group I**

The null hypothesis in this case did not hold. The times-on-task for Group I (Figure 4) decreased at a sharper rate than those in Group II (not shown). This is as expected given that the three interfaces which Group I interacted with were presented in decreasing order of element layout proximity, whereas the three interfaces which Group II interacted with were presented in increasing order of element layout proximity. The time-on-task data imply that task efficiency increases as interface element proximity decreases.

The segregated keyboard logging data indicated that task effectiveness, as measured by keystrokes necessary to accomplish a task, also tends to increase as interface element proximity decreases. As with time-on-task, when interfaces were presented in decreasing order of element closeness, task completion keystrokes decreased as expected. When interfaces were presented in increasing order of element closeness, the same flattening of the curve was observed as was seen with the time-on-task charts. The curve-flattening indicates that an increase in effectiveness due to learning over time is in effect colliding with the ineffectiveness burden imposed by the wide spacing between the interface's elements. Similar trends are seen in data for the total number of correction-keystrokes, i.e. the number of keypresses necessary to undo a mistake. Correction-keystrokes is defined as 2 × "Backspace" keypresses + 2 × "Delete" keypresses + any arrow-key presses (note that the experiment disabled cut-and-paste and highlight/delete input features). Figure 5 shows these data for Group I.

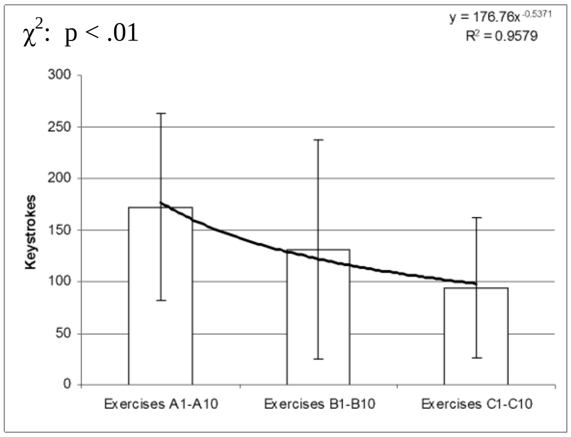


**Figure 5. Mean Aggregate Correction-Keystrokes, Group I**

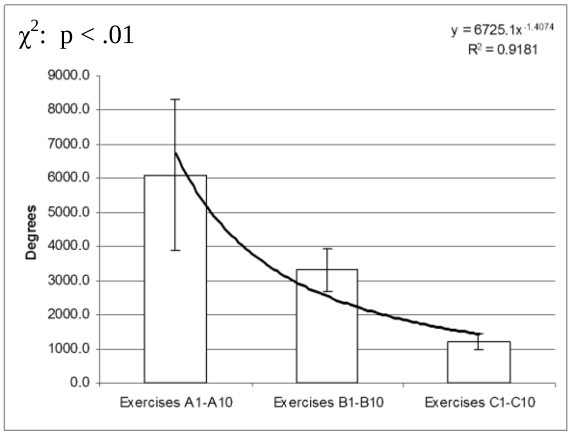


**Figure 6. Mean Aggregate Gaze-Path Traversal, Group I**

The keystroke and time-on-task data indicate quite definitively that there is some sort of underlying usability issue with Form A which is inhibiting user effectiveness and efficiency. The eye-tracker data confirm this finding and furthermore show the underlying cause of the usability issues. There is a marked difference in the eye-movement distances required for Forms A, B and C. An increase in related element proximity strongly correlates with shorter gaze-path traversal as well as shorter jumps between points-of-interest within the interface. In the case of eye movements, order of presentation did not induce any "learning effect" i.e. decrease in required eye movement as time progresses. Figure 6 shows eye-movement effort for Group I; the curves for Group II (not shown) were very similar, indicating that eye-movement trends do not vary whether forms are presented in decreasing or increasing order of proximity.

## 4. Ongoing and future research

Layout is but one of several interface design concerns. Guidelines have been formulated for several other areas of design, including widget characteristics, element interaction, functional sequencing, dialog phrasing, online or inline documentation, colors, fonts, frame sizing and placement, and several additional items [3][9]. It would be valuable to conduct total-effort metric verifications of best practices for each of these areas.

*Fitts' Law* is a simple predictive formula which specifies that the time required for a user to acquire a stationary target by manipulating a moving object will vary depending upon the distance to the target and the size of the target [10]. An experiment is currently in progress to verify Fitts' Law using a TEM framework. In this experiment, subjects are asked to click on targets placed at various radii from a center-point and then click back onto the center-point. Mouse-clicks, mouse-pointer path traversal, and eye-movement metrics are recorded for each subject.

## 5. Conclusion

The experiment described above verifies that a total-effort metric approach provides a greater breadth and depth of insight into usability issues than more conventional evaluation methods. Designers who knew nothing in advance about the design of Forms A, B and C could use the eye-tracker data in combination with the keystroke and time-on-task data to make a reasonable conclusion about the underlying nature of each form's usability. Using the TEM methodology, designers would not only be able to conclude which form factor was problematic, but they would also gain insight into the specific nature of the underlying problem.

Myers observes that the typical software engineer reacts to the subjective nature of usability testing as it is currently practiced with frustration and skepticism [8]. Total-effort metrics provides data that is quantitative, objective, and presumably more palatable to software designers. The authors believe that a TEM approach to usability testing is more in line with traditional software testing practices and has the potential to leverage greater overall acceptance of usability testing within the software engineering community.

## 6. References

[1] N. Bevan, "International standards for HCI and usability", *International Journal of Human-Computer Studies* 55(4), Academic Press, Oxford, UK, 2001, pp. 533-552. Emphasis added.

[2] Dumas, J. S., and J. C. Redish, *A Practical Guide to Usability Testing*, Ablex Publishing Corp., Norwood, NJ, 1993.

[3] Fowler, S., *GUI Design Handbook*, McGraw-Hill, New York, NY, 1998.

[4] International Standards Organization, "Software engineering — product quality — part 1: Quality model", *ISO/IEC 9126-1:2001(E)*, ISO, Geneva, Switzerland, 2001. Emphasis added.

[5] Jones, C., *Software Quality – Analysis and Guidelines for Success*, International Thomson Computer Press, Boston, MA, 1997. Emphasis added.

[6] O. V. Komogortsev, C. J. Mueller, D. Tamir, and L. Feldman, "An effort-based model of software usability", *Proceedings of the International Conference on Software Engineering Theory and Practice* (Orlando, Florida, July 13 – 16, 2009), SETP-09, 2009.

[7] C. J. Mueller, D. Tamir, O. V. Komogortsev, and L. Feldman, "An economical approach to usability testing", *Proceedings of the 33rd Annual IEE International Computer Software and Applications Conference* (Seattle, Washington, July 20 – 24, 2009), COMPSAC '09, 2009.

[8] Myers, G., *The Art of Software Testing*, John Wiley & Sons, Inc., Hoboken, NJ, 2004.

[9] Nielsen, J., *Usability Engineering*, Academic Press, Boston, 1993.

[10] Pressman, R. S., *Software Engineering: A Practitioner's Approach*, McGraw-Hill, Boston, 2005.

[11] F. E. Ritter, and L. J. Schooler, "The learning curve", *International Encyclopedia of the Social and Behavioral Sciences*, Pergamon, Amsterdam, 2002, pp. 8602-8605.

[12] D. Tamir, O. V. Komogortsev, and C. J. Mueller, "An effort and time based measure of usability", *Proceedings of the 6th Workshop on Software Quality* (Leipzig, Germany, May 10-18, 2008), ICSE '08, 2008.

[13] Tullis, T. A., and W. Albert, *Measuring the User Experience: Collecting, Analyzing and Presenting Usability Metrics*, Elsevier/Morgan Kaufmann, Amsterdam, 2008.

[14] N. Vukelja, L. Müller, and K. Opwis, "Are engineers condemned to design? A survey on software engineering and UI design in Switzerland", *Lecture Notes in Computer Science* 4663, Springer Berlin, Heidelberg, Germany, 2007.