

MauroNewMedia

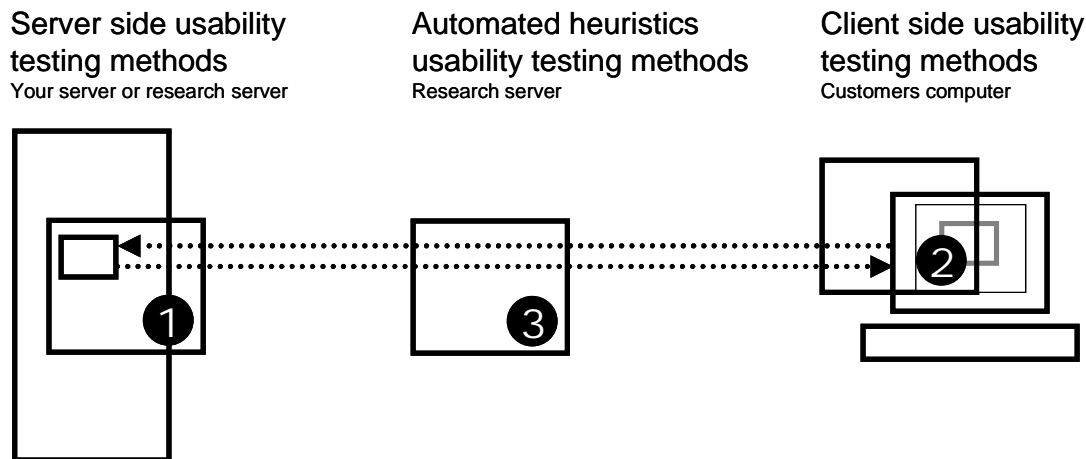
White Paper

Version 2.0

Publication date: 12/2/2002

Length 57 pages

Professional usability testing and return on investment as it applies
to user interface design for web-based products and services
(a review of online v lab-based approaches)



Author
Charles L. Mauro
President
MauroNewMedia, Inc.
524 Broadway
New York, NY 10012
212-343-2878
Email Cmauro@MauroNewMedia.com

MauroNewMedia¹

Professional Usability Testing and Return on Investment as it Applies to User Interface Design for Web-Based Products and Services (*A review of online v lab-based approaches*)

Author
Charles L. Mauro²
President
MauroNewMedia, Inc.
524 Broadway
New York, NY 10012
212-343-2878
Email Cmauro@MauroNewMedia.com

¹ For a detailed summary of the services and expertise profiles of MauroNewMedia, please refer to <http://www.mauronewmedia.com> or review “About MauroNewMedia” in Part 12 of this White Paper. For additional information on User-Centered Design, visit our public interest web site at <http://www.taskz.com>.

² For biographical information on the author, please refer to “About the Author” in Part 11 of this White Paper.

Abstract

Professional Usability Testing and Return on Investment as it Applies to User Interface Design for Web-Based Products and Services *(A review of online v lab-based approaches)*

Professional usability engineering and testing is a well-established development discipline that has been used extensively to create some of our most successful military and commercial systems. With the maturation of the web as a delivery model for information and E-Commerce, the formal science of usability will become increasingly important. This paper discusses the return on investment (ROI) implications of integrating formal usability testing methods into web development projects. Online and traditional lab-based approaches are discussed and compared for their respective strengths and weaknesses. The white paper provides detailed technical descriptions of current online usability testing methods and draws conclusions about the future of this important new customer response testing methodology. It includes a comprehensive trade-off matrix useful in making decisions about important technological approaches and research benefits. This white paper covers material delivered in a series of Executive Briefing Sessions presented by Charles L. Mauro, President MauroNewMedia. The sessions were held in New York City, Stamford, Connecticut and Chicago in late 2002.

12/2/2002

Charles L. Mauro
President
MauroNewMedia
Cmauro@MauroNewMedia.Com

Table of contents

Part 1: Science and History of Usability	5
Part 2: Business Rationale Behind Professional Usability Testing.....	9
Part 3: The Science Behind Professional Usability Testing.....	17
Part 4: Online vs. Lab-Based Professional Usability Testing	24
Part 5: The State of the Art in Online Testing Tools	32
Part 6: Methodology trade-off matrix	41
Part 7: Selecting the Right Approach.....	47
Part 8: Comprehensive Approach / <i>MetricPlus</i> [®]	51
Part: 9 About <i>MetricPlus</i> [®]	52
Part 10: About the Author	54
Part 11: About MauroNewMedia (MNM)	55
Part 12: Informal peer review and acknowledgements	56
Part 13: Recommended Reading and additional information	57

Part 1: Science and History of Usability

New, yet well proven

Even though “usability testing” has just recently become a priority for web development teams, this important science is a well proven and highly effective development tool that has been used by leading software and hardware engineering teams for decades. For more than 60 years, professional usability engineering and testing has played a critical role in the design and development of important products and systems. From the outset it should be clear that professional usability engineering and, by association, professional usability testing is a specialized field of expertise requiring formal experience and training in the cognitive sciences and related fields.

The following is a formal definition of **usability testing** discussed in this paper. This definition is meant to set the framework for what will soon be a migration toward the use of scientifically based usability engineering and testing in mission-critical web development programs.

***Definition:** Professional usability testing is defined as a formal research methodology that adheres to the processes and rules of scientific investigation as developed and taught in formal graduate level programs in the cognitive sciences. Practitioners of this type of research hold advanced degrees in human factors engineering, ergonomics, or other relevant cognitive science fields. This field of expertise is also known under the professional terms of human factors engineering, usability engineering, human-computer interaction and cognitive ergonomics. It is important to note that, within the overall field of formal usability science, “Human-Computer Interaction” (HCI) has become a well-recognized sub-specialty. The primary focus of this paper is on the application of professional usability engineering and testing methods to screen-based products and services. A large and active body of work is taking place outside the HCI field.³*

Why Important Systems Work

The operational effectiveness of many of our most important military, aerospace, and commercial systems is based, in part, on the application of professional usability science and related testing. Formal usability testing has been applied to all manner of products and services, ranging from consumer products to the large networks that combine human participants with computer-based systems. As a formal discipline, the methods can be used to determine the relative and absolute effectiveness of the connection between

³ This definition adheres tightly to established expertise and experience profiles provided by the Human Factors and Ergonomics Society and other professional organizations that deal with professional usability research as a formal development discipline. For an interactive definition of User-Centered Design, an important aspect of formal usability science, visit: <http://www.taskz.com/definitions>.

people and machine or to define and optimize your customers' satisfaction with your web-based system or service. History has shown that professional usability testing is a powerful and effective method for determining objectively the speed, reliability, and satisfaction users experience as they interact with your screen-based product or service.

Lab-based usability testing

From its inception in the 1940s until the late 1990s, formal usability testing was executed in a laboratory-based setting. This process has included design and execution of controlled experiments based on observations of humans interacting with machines. In addition to this more "applied" form of testing, a considerable amount of basic research was conducted in areas known to have a large impact on the usability of technology-based products. These studies focused essentially on skill acquisition, reaction to different types of stimuli, and issues related to human information processing. The rapid advance of the general science of human factors engineering (recently known as usability engineering) is based on the proper use of lab-based testing methods. It is important to point out that lab-based testing is even more applicable today than in the past. A key insight of this paper is the point that lab-based testing can be combined with new online customer response and behavior tracking tools to create an even more robust understanding of how customer/users interact with your screen-based products and services.

Migration to Screen-Based Delivery Models

During the past 10 years, it has become even more important to apply professional usability science to the creation of effective screen-based delivery systems such as web sites. There is an overriding reason for this trend. Clearly, many of the features and functions that were previously performed by traditional products and services are migrating, at an accelerating rate, to screen-based delivery. We can see this migration clearly by looking at how personal banking is migrating from a human-mediated interaction (Your Friendly Bank Teller) to a machine-mediated experience that includes ATMs, your home computer, and even your cell phone. This shift has lead many in the field of professional usability engineering and testing to the conclusion that research and development of screen-based products and services is now a fully formed sub-specialty of usability science called Human-Computer Interaction or HCI for short.⁴

Expanding Area of Research

There is general agreement that the design and testing of screen-based systems requires special expertise and focus as compared to design and testing of traditional three-dimensional products and human-mediated services. This new focus on the primary and secondary issues surrounding the science of Human-Computer Interaction has lead to the development of significant academic research at several leading universities in the United States and abroad. Research dollars are beginning to flow into corporate research labs for

⁴ There are now numerous professional interest groups with-in well established professional societies such as ACM, HFES, SIGCHI and SIGGRAPH that specifically address this new area of specialization.

the specific purpose of developing more rigorous methods for solving a wide range of usability issues related to human-computer interaction methods including speech recognition and other interaction modes. These efforts focus on a wide range of applications, from fixed workstation designs to hand-held devices such as cell phones, PDAs, ATM machines, and PCs to complex process control centers containing hundreds of terminals, large-scale projection displays, and 3-D mapping systems

Definitions Are Important

Thomas Kuhn in his text *The Structure of Scientific Revolutions* acknowledges that the establishment of formal definitions is a sign of a maturing science. For the purposes of this paper we use the terms “user interface,” “interface,” “customer experience,” and “human-computer interface” interchangeably. All of these terms describe that part of a screen-based delivery system that we see, touch, and interact with as a means of achieving a predetermined set of goals and objectives.

Generation One of the Internet and Subjective Decision Making

Until recently applying rigorous, professionally executed usability research in the design of web-based products and services has not been a part of most web development efforts. Generation One of the internet used an unstructured software development methodology that did not take advantage of professional usability testing. Most critical user-interface design decisions were left to the intuition of the development team. We now know that in many cases this approach yields error-prone and complex screen-based systems.⁵ It is essential that high-level development managers understand the benefits and limitations of professionally executed online and lab-based usability testing. Both of these methods offer tremendous advantages to those teams that know how and when to use such problem-solving tools in the development of world-class web-based software design and engineering solutions.

Guru Usability and Unmet Promises

Along with the dramatic increase in funding for internet startups during the late 1990s came the rise of guru usability experts offering services at very high rates. Along with the rise in guru usability came the concept that usability science could be force fed to development teams based on a few days of expensive consulting. This practice led to the mistaken impression that formal usability engineering was a quick fix, even for complex usability problems. This unfortunate trend led many in the software engineering community to view this important new science as transitory and having little effect. In reality, professional usability engineering and testing is a fully bona fide profession with a long and clear history for improving the design and acceptability of screen-based systems. Most of these gurus have migrated back to seminars where their views make

⁵ For an interesting discussion of these issues see the article “Why E-Com firms are in Flat Line Mode” by Charles L. Mauro at http://www.taskz.com/ucd_Gone_in_a_flash_indepth.php

sense and can be taken in context of the seminar setting. Guru usability does not have a place in complex product development settings.⁶

⁶ For a detailed discussion of the issues surrounding guru Usability see the article by Charles L. Mauro “Is a High Priced Usability guru a good investment?” at http://www.taskz.com/ucd_high_priced_usability_guru_indepth.php

Part 2: Business Rationale Behind Professional Usability Testing

Five Usability Statistics that Are REAL

Over the past 8 years the internet press has proffered many statistics documenting the importance of usability. Much of the reported research was anecdotal. Here are five critical statistics that have been shown to be true based on research undertaken by MauroNewMedia (MNM) in a wide range of projects and vertical applications.

1. For every dollar spent acquiring a customer you will spend \$100 dollars re-acquiring them after they leave because of poor usability or bad customer service. In several large studies conducted by MauroNewMedia this statistic has been verified. In fact, in some settings such as online banking the cost of re-acquiring customers may be so high as to make such efforts not worth action. Professional usability testing, if properly structured, can be used to address the problems of customer rejection rates directly by subjecting systems to analysis using critical incident techniques and other methods aimed at identifying critical customer experience design flaws that lead to loss of customer confidence. These methods are well proven and powerful. Online testing systems can be useful in addressing this problem as well.

2. More than 95% of your customers will use less than 5% of the features and functions of your site. Customers will NEVER use about 75% of the functions on your site. This is an essentially correct finding. The reasons for this are many and complex. During early stages of development, however, professional usability testing clearly show which features and functions are relevant to your overall business objectives. By using testing methods that focus on feature-function tradeoff analysis, it is possible to reduce dramatically the complexity of the user-interface design itself and more importantly the cost and complexity of the entire software- and hardware-based system. This reduction can have a dramatic effect on project lead times and costs.

3. The single largest predictor of call center volume is your web site's usability. Calls cost an average \$22-\$30 per call. The interesting aspect of this statistic, which is true, is that call center volume can be fully predicted by executing professional usability testing research during development. Even more important is the fact that if customers do resort to call center support they will be far more likely to become repeat users of phone-based problem resolution. This means that, in real dollar terms, the actual cost per call is probably far greater than the standard rate quoted above.

4. For every \$10 spent defining and solving critical usability problems early in development using professional usability research, you will save about \$100 in development costs. By using professional usability engineering and testing early in development, we have seen clients dramatically reduce the complexity of their software through elimination of unnecessary features thus reducing the cost of coding and more importantly the cost of testing and fixing bugs in the system. In one large commercial client, the time to execute a quality assurance test fully was reduced by 85% because of a decrease in features brought about through the application of formal usability engineering and testing during the early phases of development. This amounted to a savings of approximately \$15 million and a reduction in schedule by 18 months over the prior software development iteration.

5. For every dollar you spend improving the visual design or style of your site, you will receive virtually no improvement in sales. The same dollar spent on improving core behavioral interactions with your site's critical way-finding and form-filling functions will, however, return \$50-100 if executed in a professional and rigorous manner. In several large studies conducted by MauroNewMedia during the past 5 years, it was clear that spending large sums on web site design (re-design) efforts produced almost no benefit in terms of improving the business performance of large E-Com offerings. In one large client's case, serial re-design efforts by several large web development firms used approximately \$100 million in development fees. Yet the number of new customers declined, those retained remained level, and almost no customers were migrated to other services or were involved in cross purchasing of products or services.

This finding clearly shows the complexity of creating a steady increase in business performance through re-design of large complex sites. If these development teams had used rigorous usability testing methods before re-design, they would have seen immediately that the visual style of their site was not the driving force for improving the customer acquisition, retention, and migration. In a recent study conducted by OpinionLab,⁷ it was found that of 12 major site re-designs only about 20% of the sites achieved their prior level of subjective ratings by customers. The remaining sites were actually judged worse by customers. Obviously, we must ask the critical question "Where is all the development money going for site design and re-design"?

What are the Hard Benefits?

Professional usability testing methodologies have been proven to deliver significant return on investment.⁸ Two decades ago the U.S. military discovered that many complex problems involving interactions between human participants and advanced technology

⁷ For a copy of this report visit, <http://www.opinionlab.com/>

⁸ For a comprehensive list of case studies on the application of professional usability engineering and testing visit <http://www.mauronewmedia.com/casestudies.html> also see "Cost-Justifying Usability" by Bias and Mayhew at <http://www.mauronewmedia.com/reading.htm#uid>

could not be solved without application of a structured, professional usability testing approach. This formal research methodology makes it possible to reduce user-induced errors and training time and, most importantly, to cut software development lead times and costs. Professional usability science delivers profit to the bottom line by increasing the rates of customer acquisition, retention, and migration. This new science is fundamentally a management tool for making objective, mission-critical design decisions about screen-based and other interactive computing products and services. In the end, formal usability testing dramatically improves the users subjective and objective experience with the system. This leads to improved brand attribute conveyance, increased user satisfaction, and improved productivity as determined by reductions in both user task time and critical errors.

Where Does the Money go in E-com Development?

If we look in detail at how costs are allocated in a typical large-scale E-Com development project, we see that optimization of the customer experience or user interface design is critically important and very costly. For example, in most projects and especially those that are aimed at a broad consumer profile, it is common for the design of the user interface to use 50-70% of total system development costs.⁹ In fact, user-interface design costs can be as high as 85% of the total development costs. This is likely if a project undergoes serial iterations involving major changes in E-Com strategy or customer experience design. In one large financial services E-Com effort, more than \$500 million was spent without a single screen being delivered to the customer. During this effort none of the seven firms retained in the design of the new site conducted professional usability research. At the end of the costly development cycle, the final site was still complex and did not result in significant increases in customer acquisition, retention, or migration. For a detailed discussion of the impact poor usability engineering has had on the general category of Financial Services see the article cited below (Mauro 2001).¹⁰

Improving Your Percentages

In another research project conducted by MauroNewMedia where professional usability research and testing were used in the early phases of development, costs associated with the design of the user interface constituted between 15% and 25% of the total development costs. This dramatic reduction in fees was correlated with early identification of a core feature set and an ability to prioritize user-interface design solutions based on usability testing in a rapid prototyping development environment. If professional usability engineering methods including usability testing are used early and frequently, development costs are significantly reduced.

⁹ In a series of studies undertaken by MauroNewMedia it was discovered that funds expended on user interface design averaged between 50%-70% of the total system development costs. This finding was independent of the size of the effort and the category of E-Com service under development.

¹⁰ For a copy of the paper "Usability and Online Financial Services, Big Losses" by Charles L. Mauro at http://www.taskz.com/ucd_Usability_financial_indepth.php

Feature bloat

A primary benefit of professional usability testing is its ability to address the critical issue of feature bloat brought about by poorly structured business objectives and constantly changing design specifications. As previously mentioned, in traditional software, and even more so in large E-Com web development efforts, about 5% of features available to the customer are used 95% of the time.¹¹ A more staggering statistic is the fact that some 70% of user-interface design features are never or rarely used. This data holds true for many web-based products and services and even for standard software packages and can be verified using even the simplest log file analysis. Fewer features lead to simpler software and infrastructure and greater user satisfaction. Even the best professional usability research will fail in the face of poorly defined business objectives and unclear strategy.

User Interface Impacts Infrastructure As experienced E-Com development executives know, changes to the user interface often have the single largest impact on infrastructure and development schedules. Nothing pushes a project over budget and behind schedule faster than changes to a functioning user interface. On the other hand, no single aspect of the E-Com system is more subject to opinions and executive directions for changes and updates. In the real world, many of these changes are without foundation in either reliable customer testing or professional user-interface design research. Recently, it has been shown that professional usability testing methods can be an invaluable resource in determining objectively what aspects of a web-based delivery system needs to be changed and, more importantly, why enhancements are required. By linking user-interface design changes to objective customer feedback, professional usability research removes user-interface design from the subjective opinion of the developer. This approach places design changes in a more manageable and less politically charged position within the overall context of site re-design and upgrade.

Call Center Costs

Professional usability research can play a critical role in reducing call center volume and call duration by objective identification of critical user interface interaction events (critical incidents) that lead to call center intervention. Data from rigorous usability research can be used to structure call center problem resolution databases and interfaces and to plan call center volume levels. At the end of the day, professional usability research is a cost effective and powerful way to reduce the cost of call center support by identifying and resolving critical user problems created by interface design configurations and procedures that lead to poor usability. The interesting aspect of this benefit is that usability research can be used in the testing and design of user interfaces that have as a specific business objective the reduction of call center costs. When this is a major

¹¹ Statistic based on extensive log file analysis of large sites in the financial services, consumer products and automotive industries. Analysis undertaken by MauroNewMedia between 2000 and 2002. This finding is supported in other proprietary studies undertaken by MauroNewMedia for traditional software.

business objective of a development effort, methods can be put in place that address this problem before any hard coding of the system is begun.

Return on Investment for Professional Usability Testing

Robert Pressman in his book *Software Engineering: A Practitioners Approach* says that it is important to solve problems early in the software development cycle. Even Pressman's model and related costs, however, pale against the real costs of not addressing problems related to usability. Poor usability has an impact not only on system reliability (Pressman's main point) but also on customer acquisition, retention, and migration. Pressman and others have shown that for each phase of development that proceeds without formal usability testing the cost of fixing usability problems increases by a factor of 10. We can see immediately that costs rapidly expand to very high levels. This impact is especially important for interfaces that have a high transactional component and that offer customers goods and services that are fee based. In those cases, customers tend to be loyal up to a point, and when they flee because of poor usability or customer experience design they are exceedingly costly to recapture as discussed in Part 1 of this paper.

Early error detection and Return on Investment

Solving one serious usability problem early in the development cycle may require minimal costs in terms of actual usability testing fees. Leaving that same problem until after launch, however, will cost at least 100 times as much to fix. In studies by MauroNewMedia, the actual cost of solving complex usability problems after beta was closer to 1000 times original costs. It is important to note that the impact of poor usability is rarely known within the context of larger web development Return on Investment (ROI) modeling because the impact sphere often covers cost centers that are not part of the normal web development ROI model. Such costs often include employee training, fulfillment, facility maintenance, returned goods, and lost cross-sell opportunities. These are only a few of the factors that poor usability impacts.

Currently, those web development teams that do use professional usability testing tend to do so too late in the development cycle. As a result, they are not achieving a significant ROI either on usability testing or on the resulting design changes. But more important the benefit of usability testing is being lost in the larger context since the cost of making changes based on usability research grows dramatically with each release cycle. As we can see from Figure 1, the greatest ROI for professional usability research is achieved during the basic concept development phase of a large project. This is the time when usability engineering and testing return the most benefit. Yet, many teams tend to perform usability testing much later in the development cycle.

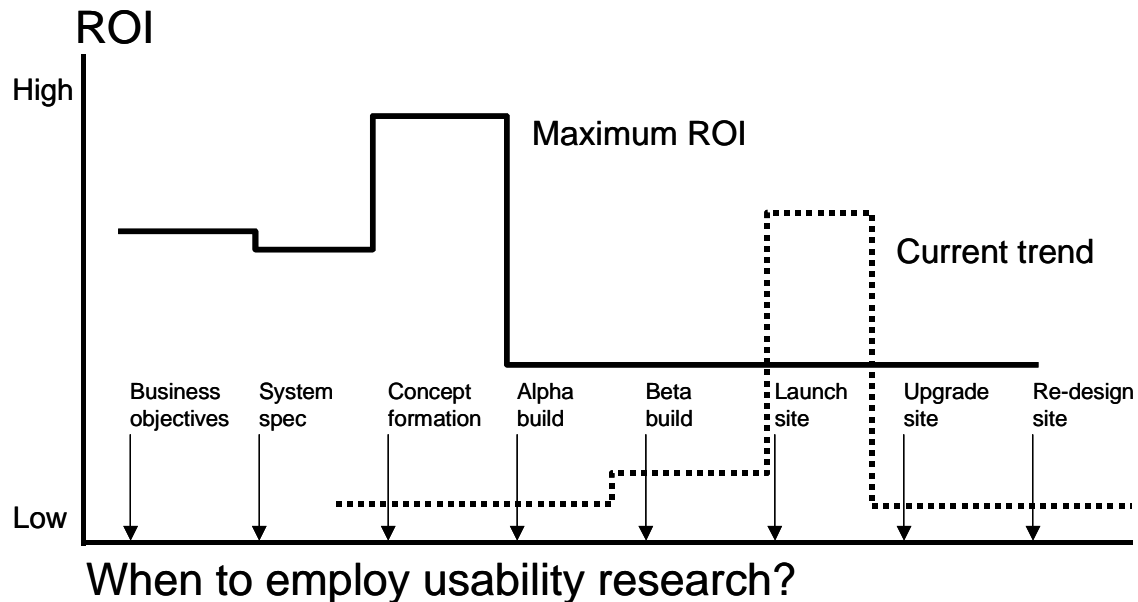


Figure 1: ROI and professional usability testing by phase of development

This timing, however, is not the important point in terms of ROI and usability testing. After launch or even well into the beta release cycle, some complex usability problems simply cannot be fixed without a major re-design of the infrastructure. This fact is especially true for interfaces that require the use of databases and search functions for the location and purchase of items from a large inventory. These systems routinely fail to meet baseline usability requirements and cannot be improved without major investments in new infrastructure and programming. In one large study undertaken by MauroNewMedia involving an E-Com site selling consumer products, the site's search engine returned a wrong or incomplete list of search queries 57% of the time. On average 46% of the site's customers left without locating the items they wished to purchase even though such items existed and were available on the site. By applying professional usability testing in the design of the search query system, the E-Com site could have been improved by a full order of magnitude. Waiting until the database and related search functionality was complete, however, meant spending more than \$1 million in re-design and programming. The cost of a professional usability testing study early in development would have been about \$25,000.

Pay For It But Do Not Use It?

It has always been a curious fact that some large E-Com clients will use rigorous usability testing and advanced human factors engineering methods at various phases of development, then ignore the results. On more occasions than we would expect, executives charged with managing large web development teams ignore the findings of such research and rely on subjective instincts when making mission-critical decisions. In any field other than web development, ignoring such decision-making models would be

career threatening. In the first generation of web development, however, this was the model of preference for development executives. This scenario is rapidly changing as we begin to understand that web development like all other complex product delivery problems must be based on the best science we have. Professional usability engineering is one component of this science.

Not This Time Around

Another common mistake web development teams make is thinking of usability testing as strictly an evaluative tool and not a development discipline. By taking this view, development teams often wait until far too late in the development cycle before using professional usability testing. This inevitably leads to cancellation of usability testing efforts or postponing such research until the next iteration. This approach is common even among the largest web development teams. It is a major reason why many web-based systems are poor examples of usability. Often this delayed response to the need for formal usability testing becomes part of the de facto development model for even the largest web design efforts. Clients routinely cancel and delay usability research year after year as their sites degrade into exceedingly poor levels of usability and acceptability. In reality, this delayed response is always the result of a real need by the development team to just get the site up and then conduct usability testing. Until web development teams start projects with appropriate and well-conceived usability engineering, untold billions of dollars will continue to be wasted.

The time to use professional usability testing is **before** a site design or re-design not after. As we mentioned, the ROI for using professional usability testing after launch is exceedingly low. If you are about to undertake a major web design effort, **now** is the time to speak to a professional usability-engineering group about how to proceed and when to use the methods and practices of formal usability science.

Customer Acquisition Costs

If we factor into the cost-benefit model the cost of customer acquisition for web-based services (for example, acquiring one online banking customer costs about \$1,500), the real costs of poor usability begin to emerge. Not only do we need significant funds to fix the problem that is sending customers away, we also need major investment in marketing fees just to balance the number of customers who are leaving the service due to poor usability. This is the big picture on ROI and usability testing. We can, however, easily expand the model to include other costs such as call center support costs and the impact of negative peer recommendation caused by poor usability. It soon becomes clear that using professional usability testing and related sciences to solve usability problems fast and early can have a significant ROI.

How Much Is the Right Amount to Spend on Usability Problems?

Consider that \$250,000 dollars spent on professional usability engineering and testing during early phases of development or site re-design can save millions of dollars later in the life of an online service. This same \$250,000 spent after launch of a poorly designed

system will rarely solve even one significant usability problem in a meaningful manner. In their landmark book, *Cost-Justifying Usability*,¹² Bias and Mayhew present compelling examples of ROI-based benefits flowing from the application of professional usability engineering and testing. But the most important issue related to ROI modeling and E-Com development is the fact that it is virtually impossible to define the attributes of an effective cost-benefit model without objective quantification of the usability of the screen-based delivery system through rigorous and formal user testing. Yet, in several studies undertaken by MauroNewMedia in the design of complex, mission-critical interfaces, corporations routinely program ROI models without estimating user-induced errors and task efficiency. ROI models that lack objective user performance variables are a waste of management resources no matter how compelling the Customer Relationship Model (CRM).

¹² For a specific reference on where to purchase “Cost-Justifying Usability” by Bias and Mayhew visit http://www.taskz.com/reading_indepth.php

Part 3: The Science Behind Professional Usability Testing

Important Concepts of Professional Usability Testing Methods

At the heart of both lab-based and online professional usability testing is a fundamental commitment to the use of proven methods of scientific investigation in measuring the connection between users and web-based products and services. This adherence to a formal research methodology is based on the premise that such research can lead to optimized systems that effectively combine human skills and limitations with leading technology to produce robust and financially successful E-Com delivery models. The formal roots of professional usability testing are based in the cognitive sciences. Lab-based testing (the only form of testing until the emergence of online methods) developed from a need to optimize the design of military offensive weapons systems during the Second World War. It was discovered that by applying principals of observational research and cognitive modeling,¹³ it was possible to determine in relative and absolute terms the performance of the connection between users and all manner of technology, not just screen-based interfaces. In the larger context of systems development, the realization that using formal research methods could optimize the connection between people and machine opened the door to development of high-technology weapons systems. Before this development, weapons performance began to overtake the abilities of the human participant. System reliability dropped so low that weapons development was becoming unpredictable. With the discovery and integration of professional usability science into weapons development programs, however, our role shifted to the center of the development process and weapons development began a near-vertical trajectory in terms of accuracy, reliability, and ability to deliver mission objectives.

The critical question for web development executives is how to execute this process and how to retain and manage project resources. The type of methodology used, be it traditional lab-based testing or online tools, can only be determined after careful and professional study design. In all successful professional usability testing programs, the following four key aspects of the formal research methodology must be present, properly linked, and managed.

- **Component 1: Clearly articulated hypothesis (business objectives).** Although this component may seem self-evident it is important to emphasize how critical it is to have a well-defined statement of what is to be tested and why. In the majority of usability testing projects for Generation 1 web designs, the development team had no clear idea of what aspects of the web-based customer experience were to be tested and why. In most cases what might be thought of as

¹³ In this context “cognitive modeling” is the ability to objectively define and document the human information processing system with a specific focus on how decision making impacts machine design (interface design) and other issues such as training, recruitment and job and system design.



hypothesis formation is actually definition of a formal set of business objectives for the site. At the most fundamental level, usability testing in any form is wasted without a well-defined set of business objectives that are then used as the basis for design and execution of a study that maps user performance to actual business profit and loss (P&L). When business objectives are clearly articulated, user performance can be measured in terms that will have a real impact on the P&L of the delivery system. Time and errors and customer acquisition, retention, and migration rates have meaning that can be translated into hard interface design attributes and testing sequences. This component of formal usability testing is the most important yet the most frequently overlooked. Surprisingly, sometimes even usability testing professionals do not understand the need for the development of formal business objectives. In studies conducted by MauroNewMedia, it was found that special tools and data-gathering methods are critical in the creation and approval of reliable business objectives.

- **Component 2: Proper experimental design.** For professional usability testing to be effective, the team must have an appropriate and well-defined experimental design. This means that a formal study must be designed that makes use of appropriate testing methods and protocols. In many so-called usability studies, users are polled in a traditional focus group format or even in an informal setting using one-on-one observations and verbal protocol methods. These methods may not be appropriate for making mission-critical decisions. Such methods can be generally useful in defining simple problems and usability issues. In the end ,however, an appropriate experimental design is an essential component of all professional usability testing projects. As a special note, many online testing vendors offer professional consulting services in experimental design. Use such expertise with caution. The design of an effective study is a complex task requiring special unbiased expertise. Professional usability testing firms are often the proper resource to turn to for study design and testing methods selection.
- **Component 3: Reliable data.** There has been a popular notion that conducting usability research with a small group of users can yield answers to complex usability questions. This approach popularized under the general description Discount Usability Testing¹⁴ is **not** an acceptable method for producing reliable data in an experimental setting. Certainly, we would never make mission-critical decisions using small subject samples with poorly articulated user profiles. A professional usability testing study is only as good as the data collected. It is necessary to have appropriate sample sizes, to recruit respondents who reflect your customer profile, and to never allow a study to be identified with the sponsor. Studies conducted on the site of the client are basically meaningless in

¹⁴ For a detailed discussion of the problems associated with Discount Usability Testing see the paper by Joe Dumas "How Many Participants In A Usability Study Are Enough?" Published as a technical paper in the "Essays on Usability" Edited by Russell J. Branaghan for The Usability Professionals Association at UPAssoc.Org or call 312-596-5298 for more information.

terms of experimental reliability, yet many studies are conducted in such a manner. Respondents do not deliver reliable or objective responses on the site of the sponsoring organization. If you employ professional usability experts in the design of the study and properly fund the actual study so an appropriate number of respondents are tested you will have reliable data. This is one area where online usability testing methods offer significant advantages over traditional lab-based methods. Large samples sizes ranging from a few hundred to thousands can be polled in online studies. Such studies, however, have other limitations that will be discussed later in this paper. Debra Mayhew, a leading usability expert, has written extensively on the topic of short-cutting professional usability testing methods. For an interesting exposition on this topic see her articles in Taskz.com.¹⁵ Other interesting papers on ROI and usability testing can be found at the University of California Berkeley Computer Science web site¹⁶ where there is a homepage set up that focuses entirely on these issues.

- **Component 4: Data must be properly interpreted.** The data that flows from rigorous usability testing is complex and multidimensional. This data is difficult to analyze and summarize. It is critical that experts with a background and knowledge in formal usability science review data from usability studies. What seems obvious at first view often has latent meaning. Even the very best data can be misleading if it is not subjected to proper statistical analysis. In fact, data from users in professional usability studies can appear technically confounded. This state is best expressed in what is known as the four paradoxes of usability testing.¹⁷ It is important to note that world-class usability testing is of little use if it is not supported by a dedicated development team prepared to implement recommendations and changes. At the end of the day, formal usability science is focused on creating solutions to complex human-computer interaction problems. Solutions to these problems require a **team** effort extending beyond the role of the usability professional.

As an expert with more than 25 years experience in formal usability testing, it is clear to me that more than 90% of the usability testing undertaken during Generation 1 of the internet did not meet baseline requirements for a professional usability testing protocol. Generation 1 of the web suffered from bad usability science and guru usability science. Both left many development teams with concern over the benefits and methods of this critical development tool. This opinion will change as professional usability testing methods increasingly find their way into large-scale web development efforts. There is no other process for optimization of the human-computer interface that is cost effective and

¹⁵ See *Usability Testing: You get what you pay for* at By Debra J. Mayhew 2002 at http://www.taskz.com/ucd_usability_testing_indepth.php. For another interesting article on this important topic see the publication by Carol Righi Ph.D at http://www.taskz.com/ucd_righi2_summary.php

¹⁶ The UC Berkeley ROI site can be viewed at: <http://www.sims.berkeley.edu/~sinha/UsabilityROI.html>

¹⁷ The 4 Paradoxes of usability testing are instances where data appears to be contradictory or confounded. They are discussed later in this White Paper.

reliable. It is not a matter of **if** such methods will be applied it is only a matter of **when** and **how**. A new generation of E-Com development executives is increasingly finding formal usability science the central framework for making mission-critical decisions.

Improving Usability vs. Improving Visual Style

There has been a running debate among large web development teams, often times extending into the office of the corporate CFO and CTO, about how large budgets should be allocated when undertaking major web design or re-design efforts. Should you spend \$250,000 dollars with a web development agency for an updating of the visual branding and information architecture of the site? Should you spend that money on professional usability research and enhancement of the procedural interactions of the customer as they interact with your system? There is a reliable and well-developed framework for making these types of decisions. Web development teams in critical decision-making scenarios do not, however, routinely apply these methods. As a result, allocation of technical and consulting services often turns into a hotly contested political battle that does little for team morale or development of a cohesive vision. By drawing on fundamental research from the cognitive and management sciences, MauroNewMedia has developed an effective framework for addressing these complex questions.

Return on Investment Modeling and customer behavior

The best way to determine an effective means for allocating development funds is to frame the issue in terms of return on investment (ROI) for the funds to be expended. Unfortunately, over the past 2-3 years ROI modeling, much like usability testing, has become a popular buzzword in web development teams. As any seasoned development manager well knows, all ROI modeling comes unglued unless the basic model is sound and the data itself is reasonably accurate. In our work with leading corporations on these issues, we have seen ROI models for allocation of web development resources range from complex mathematical simulations to rule of thumb decision-making by the CFO and CTO over lunch. In fact, both approaches can be valid if those involved in the process have a fundamental understanding of the complex balance that must be obtained between visual design and usability engineering of the customer experience. To the surprise of most development executives, there is a well-reasoned research method for addressing these issues. If we look objectively at the role of visual design and usability engineering in the creation of a powerful customer experience design, surprising issues emerge that can guide our decision-making.

Hygiene Factors: Why Visual Design Goes Only So Far

Research principals from the behavioral sciences deal with the basic concept of how much benefit users of technology will find from increasing levels of visual or graphic design improvements in the user interface. Drawing on this research, it is clear that visual design enhancements beyond a certain acceptable level will add little by way of improved customer satisfaction or, far more important, have little to no impact on customer acquisition, retention, or migration. This is not what most web development agencies

communicate to corporate development executives when selling re-design efforts. This fact is, however, clear. If you continue to improve the visual branding or graphic design of the user interface beyond an acceptable level, you will do nothing to improve critical business drivers. Your ROI will be essentially negative or flat. But this is not the whole story. This research also convincingly shows that poor levels of look and feel will have a negative impact on customer response. The critical question is, What constitutes an acceptable level of visual look and feel? This question is best answered by a form of market research known as “visual and interactive brand attribute testing.”

These tests executed using online testing tools tell you what level of visual design is appropriate and if you need to spend time and money on re-designing and upgrading the visual brand of the site. The important point is that such decisions are based on objective feedback from your customer base. Customers ultimately determine what is appropriate and acceptable. This form of research must be conducted before funds are allocated for design and development. A study of this type routinely offers significant ROI. It is important to note that hygiene factors and the relative weight of such factors vary depending upon the industry group, product, and target audience. The only way to determine the relative importance of the visual style of your interface objectively is to map the design against industry-specific references and to test the site against a formal set of visual brand attributes.

Behavioral Factors: Where the Money Must Ultimately Go

In addition to hygiene factors, there is another set of factors that must be addressed if you are to optimize the ROI for development funds during design or re-design. This second set of factors is known as behavioral factors. They are the point-by-point interactive procedural factors that allow the customer to execute transactions on your site. The important aspect of behavioral factors is that they are determined by mapping customer online behavior against actual business objectives. These factors have a directly measurable impact on critical customer interaction variables such as

1. Customer acquisition costs
2. Customer retention costs
3. Customer migration costs
4. Customer support costs
5. Customer training and skill support costs
6. Process improvement costs
7. Software design and development costs
8. Software quality assurance and testing costs
9. The cost of user-induced transaction errors
10. The cost of increased task time and task complexity

All of these issues form the basic business performance model for E-Com initiatives. In the design of a new site or planned upgrade of an existing site, these factors map directly

to the observed behavior of the customer as they enter, navigate, and perform transactions on your site. These factors are not significantly improved by increasing the visual style of the site but are directly and profoundly impacted by the actual interactive nature of the site in terms of the user's decision-making processes and sequences. Although this point has been made before in this paper, it is well worth repeating. These processes can only be understood and optimized through application of methods and practices of the cognitive sciences. It is from these sciences that professional usability engineering and testing draws its basis. In other words, how the site is organized from the very first eye-scan of the users until they migrate to your highest paying user profile the site must be modeled and optimized against rigorous and proven customer decision-making models. Figure 2 is a schematic representation of the basic concepts at the center of ROI and fund allocation decision in a large web development effort.

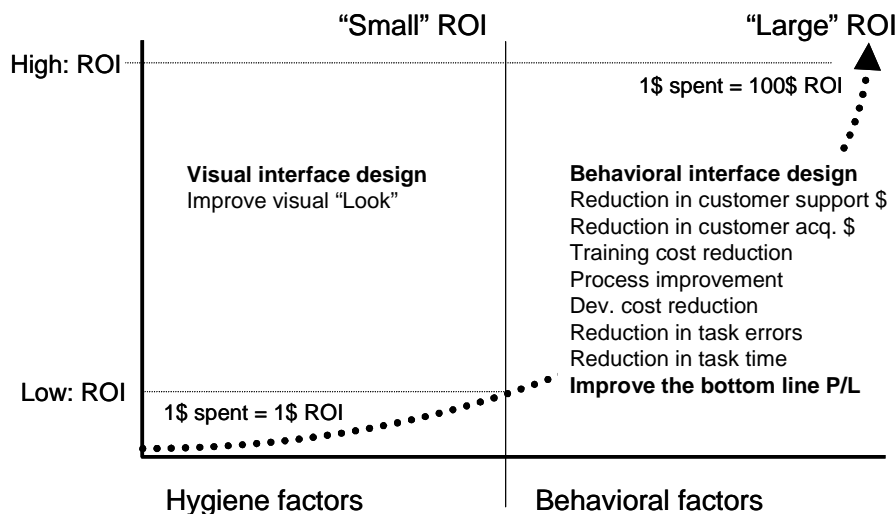


Figure 2: ROI decision variables

A comprehensive discussion of these issues is beyond the scope of this paper and certainly there are contrasting views on the appropriate model for ROI decision making. For a discussion of these factors visit <http://www.taskz.com> under the heading of Executive Primer.¹⁸

Unfortunately, many web development executives have not had the benefits of exposure to rigorous professional usability engineering and testing methods. This leaves them without sufficient background for critically evaluating project proposals and team expertise profiles when allocating development funds. Even the best MBA programs and IT graduate schools offer a superficial overview of software development methods using

¹⁸ For a detailed discussion of this issue visit <http://www.taskz.com> homepage and read through all 7 sections of the Executive Primer. The Executive Primer is a series of papers organized by development category. These papers discuss the specific functions of each critical development discipline and provide links to other information.



professional usability engineering and testing case studies or course materials.¹⁹ This problem is widespread and clearly complicates the success of many large-scale web development efforts. There is also growing awareness on the part of web development executives that current methods using poorly defined expertise profiles based primarily on the visual design bias has been ineffective in delivering highly engaging and appropriate screen-based systems.

¹⁹ In lectures and presentations at leading MBA programs the author routinely encounters curriculum that does not cover rudimentary principals of formal usability science or User-Centered Design.

Part 4: Online vs. Lab-Based Professional Usability Testing

Study background

In a study conducted by MauroNewMedia²⁰, more than 100 online usability testing products and systems were evaluated for their benefits and limitations. The study focused on identifying underlying technical approaches and limitations in the current online usability testing environment. If good experimental procedures are followed and the proper tools are used, robust and meaningful results flow from the use of online testing methods. There remains, however, a gap between the technical efficiency of new online tools and the fidelity of traditional lab-based research methods. Clearly, this paper addresses the current state of online tools, but it is important to note that a gap exists at the junction of online and traditional usability testing methods. The next generation of professional usability testing systems will likely be a hybrid of these two approaches. Such systems are currently under development and are showing promise.²¹ When combined with lab-based testing the combination offers new levels of insight into customer behavior and opinions related to usability and satisfaction. Overall, online usability testing must be viewed as a powerful and important addition to the web-development decision-making process. The critical question development executives are faced with today is how and when to use these new tools in the process of migrating their web-based business models to more productive levels.

The Basic Concept Behind Online Testing

Fundamentally, online usability testing infers usability by looking at events that have taken place as customers interact with your web site. It is important to understand that in most tools you cannot actually observe the behavior of the user except to the extent that a web-based tool can track mouse clicks and page delivery. Missing from the observational data in any online tool is a large body of relevant observational data. This data will include what the user is actually looking at, their facial expression, body posture, and most important their detailed interactive behavior linking the screen with their eyes hands and thinking patterns This is the exact opposite of lab-based testing, which gathers information from real-time observation of the user interacting with your web site or software while undertaking assigned tasks in a laboratory setting. **Online testing is a classic pattern recognition task, and lab-based testing is a real-time event-recording task.** This is an important distinction because it defines at the most basic level the

²⁰ MauroNewMedia conducted a detailed audit of over 100 online usability testing tools during the summer of 2002. The study involved examination of several leading products including both server side and client side applications. Combined approaches were also examined. The primary criteria for inclusion in the study was written claims in the vendors marketing materials making specific reference to the provision of “usability ratings” as an offering of the research tool.

²¹ One such system METRIC PLUS is currently under development by MauroNewMedia and is described in summary form in part 10 of this document. Other vendors are also developing tools and methods that bridge this important gap.

strengths and weaknesses of online usability testing. From the outset you cannot witness in real-time what your customers are actually doing as they navigate your site and make decisions along the way. All you know objectively is that an event has taken place, such as clicking on a link to call a new page from your server, remaining on a page and looking at information, requesting pages in a specific order, that can be used to infer some level of user behavior. You do not, however, know why they have undertaken this task or if what they have chosen to do can be classified as an error or a correct operation. In a lab-based testing this can also be true unless you have given the respondent the task in advance. The advantage of lab-based data is that you can ask the respondent to discuss what they are thinking at the time they undertake tasks. Advanced forms of online testing currently under development will allow acquisition of this type of data. To illustrate these issues an analogy is useful.

What Hansel and Gretel can tell us about online testing

A powerful and interesting analogy is the story of Hansel and Gretel. As the story goes the father takes the children through the woods on a circuitous walk that leaves them at a clearing assuming that they will not be able to find their way home. Unbeknownst to the father, during the journey the children leave a trail of pebbles. Much to the father's surprise the children find their way home by following the trail of pebbles. In the analogy we see that a click on a link by your customer sets in motion a series of actions that are very much like the pebbles left by Hansel and Gretel. When a mouse action takes place it marks an event in the browser that sends a request to your customer's ISP that sends a message to other servers on the internet until finally the request finds its way to your server where it is logged and dealt with by your software. At the start and end points of this journey interesting things happen. When the user clicks on a link, we say that it is a client-side event. When something takes place on your server, it is known as a server-side event. Fundamentally, online customer behavior research deals with sensing and documenting events either on the customer's machine (client-side) or on your server (server-side). But in reality almost all online vendors attempt to infer customer behavior by looking at events on both your customers' machines and on your server. The question is, What can we really determine by peering intently at such small pebbles? The answer is, like an examination of the pebbles of Hansel and Gretel, we can learn a great deal.

In the case of Hansel and Gretel we can tell by how far apart the pebbles are how fast they were walking. If the pebbles are tightly grouped at certain points along the trail we can assume that they slowed to look at something of interest. If we find pebbles scattered about we might assume that they were frightened by a bear and simply dropped the pebbles and ran. If the pebbles were on the right side of their footprints Hansel was dropping stones (he was right handed). If they were on the opposite side, Gretel was in charge of pebble dropping. We might find long stretches of no pebbles at all in which case we might infer that they simply forgot to drop pebbles as they attempted to figure out where they were as they walked. We can see that the information in the data of pebble dispersion can be interesting and to a certain extent we can infer several levels of

behavior on the part of Hansel and Gretel. This analogy should not be carried too far except to say that there is a lot of information even in pebbles left on the ground.

TCP/IP and Other Internet Concepts of Stone Laying and Way Finding

Imagine if you will what we can learn from a technology that is based on accurate recording of time and place. The internet is essentially like a massive recording machine where everything we request and receive is, by design, location and time dependent. For example, we can tell with a high degree of certainty what the customer requested (web page URL), where they came from (their URL), how long they were on a page before requesting a new page, the order of pages requested, and a multitude of other variables. But what we cannot tell for certain is what they were actually doing at the time these events took place. We cannot know this because we were not there to observe their behavior nor were we there to probe for insights by asking them questions about why they did something, the very essence of lab-based testing. But in some type of usability research, that is, when what we want to do is recognize patterns in what they are doing, this is not a problem. If we want to know how they feel when they undertake these events, we can send them a questionnaire and ask them how they felt about trying to find the latest money market yield on your web site. By correlating their behavioral interactions (page tracking) with their subjective opinions (from a survey) we can learn a great deal about their behavior and their satisfaction. With a proper understanding of how to look at the behavior research and with a well-designed survey we can tell a great deal about how they felt and what they did. But we cannot tell for certain unless we follow Hansel and Gretel for a while to confirm that what we infer from the pebbles is what is actually happening. Herein enters the important connection between online and lab-based usability testing. One without the other is bound to leave us lost in the woods.

Objectively, what can we tell from the pebbles and by association the internet communication events taking place as your customer's request traverses the internet and finds its place in your server and then returns the requested page to the customer? Like in the example of Hansel and Gretel, we can infer a great deal but we cannot be certain about anything except a few discreet timed events. The current thrust in online testing systems is toward finding new ways to expand the reliability of internet-event recording on both the client- and server-sides of the data flow. To achieve this, vendors must treat the customers' sites with special code, treat your server with special code, or do something in between that helps them record and organize the small pebbles of information being spread over the forest floor of the internet. This is not a fairy tale; it is potentially a very big business.

How Online Usability Testing Systems Are Currently Positioned in the Marketplace

It has become an unfortunate but clear trend for most online usability testing vendors to market their systems by pointing out the weaknesses in traditional lab-based usability testing approaches. This approach has done little for the credibility of some vendors who feel that their tools can and should completely replace traditional usability methods. This is a shortsighted approach. A brief discussion of the comparison of claims made by

online testing and traditional lab-based testing is worth noting here. Figure 3 shows a tradeoff matrix for online vs. lab-based testing methods. This is a typical chart used by some online testing vendors.

Online usability testing attributes	Traditional lab-based usability testing
Low cost (high cost/benefit)	High cost (low cost/benefit)
Fast project turnaround time	Slow project turnaround time
Large sample size	Small respondent sample size
Immediately scalable	Difficult to scale to larger samples
High quality standardized data	Short application period for data
More realistic test conditions	Unrealistic test conditions (lab-bias)
Makes possible longitudinal studies	Cannot easily do longitudinal studies

Figure 3

A complete analysis of the claims of online testing vendors compared to traditional lab-based testing is beyond the scope of this paper, but it is worth noting that nothing that involves testing of humans as they interact with complex technology such as a web site is as simple as the comparison matrix makes it seem. In fact, lab-based testing evolved specifically because it was learned from direct observation that users often do not report their opinions accurately or persist in the execution of complex tasks without professional probing and careful recording of actual real-time behaviors. On the one hand, online testing tools do not offer the same level of observational fidelity made possible by lab-based testing. On the other hand, online testing can offer advantages that lab-based testing cannot achieve. There are clear and well-informed reasons to use both methods.

Basic Theory Behind Web Site Usability Testing

When we look at the primary and secondary business objectives of large successful E-Com sites it is often best to describe usability or customer interaction performance in terms of scenarios. These scenarios can be thought of as compilations of observable behaviors that involve two basic types of tasks: (1) way-finding and (2) form-filling. These two types of goal-seeking behaviors are schematically represented in Figure 4.

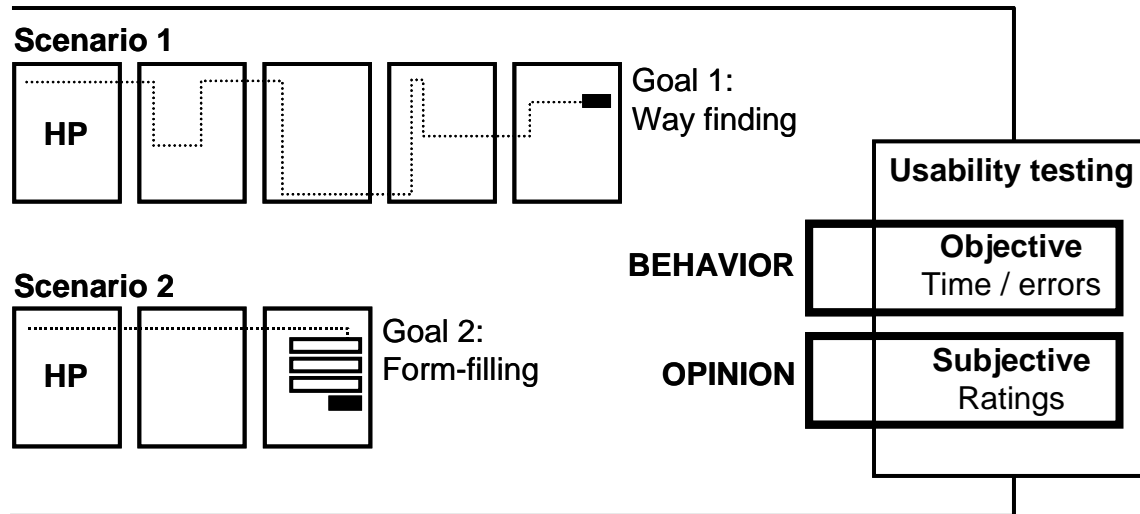


Figure 4

It is possible to categorize the majority of web-based customer interactions as either of these two types of tasks or as combinations of both. By adopting a scenario-based model, it is possible to prescribe a formal method for analysis of the benefits and limitations of online customer response testing methods and tools.

All That Is Old Is New Again

The approach of investigating a user's interaction with a screen-based system by characterizing the experience as a series of linked events is not new. In fact formal task analysis²² has been used successfully for more than 50 years as a means of defining the overall and detailed user interaction sequence and related performance in both complex and simple interactions with technology-based systems. It turns out that web sites are based on their technical underpinnings, excellent candidates for the application of formal task analysis or scenario-testing methodologies.

Best Way

By adopting a task-based scenario approach in the analysis of usability for web site performance, it is possible to identify an optimum pathway or task profile. These profiles can be useful in determining benchmark way-finding and form-filling sequences. Reliably determining the usability of a web site using this approach is, however, more complex than we might imagine at first glance. As has been mentioned previously, benchmarking and testing the usability of a series of tasks must include the gathering of two forms of data: (1) user's subjective or opinion data and (2) user's actual objective behavior data. This requirement is a hard and fast rule imposed by rigorous research methods required by professional usability engineering and testing procedures. This

²² For an excellent technical description of formal task analysis see the publication titled "Task Analysis For Instructional Design" at http://www.taskz.com/reading_indepth.php.

approach is nonnegotiable if we want to understand how the design of our sites impacts on our customer's ability to find information and execute tasks.

Typical objective behavioral variables include time and errors as measured against an optimized task profile. Subjective measures includes rankings, scales, and open-ended comments covering the customers' responses to the overall and detailed customer experience taken after they have executed an assigned task. The task-based scenario is an especially powerful method for evaluating conceptual approaches during prototype development by using simulations that allow the usability research team to learn objectively from variations in site way-finding and form-filling designs. This is a powerful yet rarely applied research method that can be implemented at low cost using online and lab-based testing methods.

Home page usability requires special expertise and methods

Whereas task-based scenario analysis is the preferred model for investigating the usability of your site's critical business performance, testing your home page for usability and customer experience design is a level more complex. In addition to being the entry point for both way-finding and form-filling tasks, the home page also serves as a high-level organizational framework for the customers first time and repeated visits to the site. Home page analysis is part and parcel to the analysis of overall usability testing but expands to include issues of visual and interactive brand attribute conveyance and transfer of learning issues. Testing of these issues requires special tools and methods.

Four paradoxes of usability research

After testing hundreds of products and systems for usability and customer response during the past 25 years, it is clear that some aspects of this type of research frequently defy common sense. In fact these contradictions are so common that they are described as the Four Paradoxes of Usability Research.

Paradox 1: Objective and subject data can be negatively correlated. It often comes as a surprise to those just starting out in this relatively new science that when comparing two systems for usability, strange things surface. For example, when executing a comparative study of two site designs, it is common for users to report high levels of subjective satisfaction for the interface on which they made far more errors and took more time on task execution. Conversely, we also see customers making fewer errors and taking less time but reporting subjectively that they like the system less than one on which they objectively make far more errors and take more time on task completion. We can see at once the problem with collecting one type of data. In both cases, we will be making decisions on the wrong conclusions. It is critical that objective and subjective performance be optimized. Simply to adopt an error-prone and time-consuming design that customers seem to like initially is short sighted because it is time and errors that ultimately affect large business issues such as data entry errors, server load system response time, and customer call center volume. Truly powerful and effective E-Com systems are optimizing for both objective and subjective customer performance. This

finding has been shown valid in many other applications including aviation, process control, aerospace, and consumer products.

Paradox 2: What looks easy on first impression can be difficult to use.

MauroNewMedia calls Paradox 2 the First Impression Paradox, and it is alive and well in web site design and development today. Web development teams, experienced mostly in the graphic arts, applied their skills to interactive web-based services to make a site appear initially simple to use. Some sites, however, end up being hideously complex. In fact so complex that both way-finding and form-filling are simply lost to the sites commitment to visual design. This is a problem that can be difficult to identify early in development but one that is critical to the successful allocation to development funds and resource retention. In a surprising way, sites can also impart the reverse reaction to customers. In other words, a site can appear complex but be simple to use. It is clear that no single form of research such as group focus methods can be used to get at customer response to issues that are imbedded in detailed interactions with your site. The best way to deal with the first impression paradox is by conducting longitudinal testing so that customers can interact with a system in a repeated form. This will almost always allow the customer to become aware of task time and errors and to determine if such performance problems are critical to their overall acceptance of the system. Ultimately, the measure of any E-Com system is in hard numbers relating to customer acquisition, retention, and migration. Subjective and objective data taken in a longitudinal formula is an excellent way to get at these issues.

Paradox 3: Subjective ratings can be skewed positive

Recently, market research professionals, most of the academic stripe, have been looking into the reliability of subjective rankings and scales as a means of gauging customer response to products and services. This new research suggests that customers, when asked to evaluate a site or system, have a significant tendency to judge the site more positively than they might under other circumstances. This positive skewing of subjective data complicates the reliability of the research data. For example, in a recent analysis of a research study MauroNewMedia discovered that simply by asking respondents to evaluate a new site redesign, the overall rating of the site was skewed a full point in the positive direction. This was compared to the same question that asked respondents to evaluate a web site. Clearly there is a substantial halo effect stemming from the very idea that the development team has spent time and effort on a site redesign. This is only one small example of biased question development. This bias is especially true when conducting survey type data collection using an online tool and is a disadvantage of online testing methods. In a lab-based setting respondents can be carefully questioned for skewing during debriefing

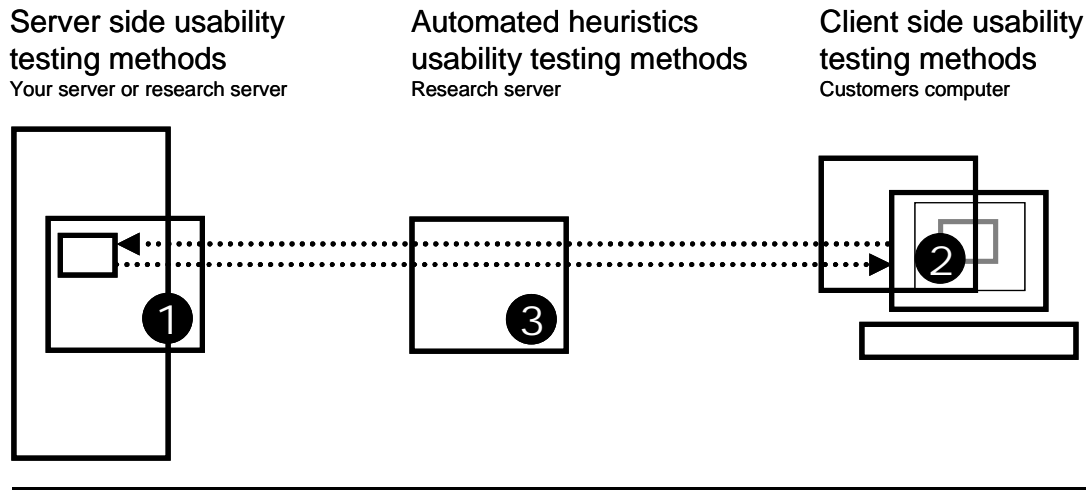
Paradox 4: Behavior is more important than look and feel

It is difficult for web development teams to understand that how successful customers are at executing tasks is more important than whether the sites they create win the latest web design awards. The reason that this is true is rooted in the fundamental idea that a

successful web presence is ultimately a means for driving revenue to the bottom line. A simplified but useful way of thinking about a web-based delivery system is viewing it as a tool that has certain performance specifications that must be met in its design. In the development of a traditional product-user interaction, performance metrics are developed before the product is turned over to the industrial design or styling team. For some inexplicable reason web development from the beginning has this process backward.

When you start a web design effort by optimizing visual style, driving profit to the bottom line is very difficult. In reality, however, we know that creating an effective means of supporting productive user behavior is the most important development variable. This goal can have wide meaning. For example, large not-for-profit institutions need to build confidence on the part of those who contribute currently and in the future. A consumer electronics site must ensure an error-free and rapid selling sequence and must cross-sell to the greatest extent possible. An educational institution must support potential applicants, incoming students, ongoing students, alumni, faculty, and administration in a manner that reflects the values and strengths of the school. All of these E-Com initiatives rely on web-based experiences to build their interactive brand and support growth. In all cases, the manner in which your site supports and encourages productive and satisfying behavior on the part of the user will determine the success of your business model. Design awards do not win customers. Engaging, error-resistant, and productive user interfaces do win customers. In addition, they garner that all important form of web marketing - peer recommendations. Historically, professional usability engineering and testing is the only proven method for mapping behavior to design and pushing value to the bottom line.

Part 5: The State of the Art in Online Testing Tools



Survey Tools and Little Else

A review of the current online usability testing offerings shows that about 95% of the current systems are objectively little more than web-based survey systems. There are well over 100 of these systems offered by a wide range of market research firms and internet start-ups. It is important to note that a web-delivered survey is likely to report the same findings as a traditional-based survey if the online version is properly optimized for screen-based delivery. All of these tools allow you to create customized surveys of your own design. Screen-based survey design is not, however, the same as creating a traditional survey. If you directly translate a traditional paper-based survey form to screen-based version you will **not** get the same results in both formats. Typical problems with online vs. traditional surveys can include greater percentage of uncompleted questions, and higher respondent confusion. Also, as was normally good practice in traditional survey design, but is now mostly ignored, you must pilot test your online survey. Things that can be easily dealt with by a survey administration person turn into major issues with online surveys. Remember, no one is there to help the respondent. Many of these systems do not properly track partial completes or reasons for termination, which means that data from abandoned surveys is often lost. In a poorly designed online survey it is sometimes difficult to separate usability problems with the survey design from usability problems with the product or service being tested. Several online systems offer simple, well designed question and survey formatting features that reduce the probability that you will end up testing the usability of the vendors survey forms as opposed to the usability of your site.

Effortless Survey Form Creation

Several of the tools examined make study design and layout of online forms easy and direct. In fact, it is so easy and direct it is almost fun. The point here is **important**: A poorly designed study thrown together quickly without careful research design is worse than no research. Just because these tools make design and layout of the survey forms easy, do not think for a moment that they do the same thing for the actual design of the study. Nonbiased questions, improperly developed scales, and poorly defined attribute maps are created everyday by well meaning development teams who lack professional training in study design. Online testing tools do nothing to change this issue. There are vendors who encourage access to their archive of questions developed for other studies. Referring to these samples often causes more confusion than help. In the end the same adage applies to research design that applies to software development - garbage in, garbage out. We suggest you use the money you save on recruitment fees to increase the quality of your study design. It will pay off several times over.

Too Much Technology Experience

No matter what online survey tool you use, gathering survey-type research studies online can skew your data far in the direction of respondents who are comfortable with browser interfaces and the use of the internet. In terms of general profile, internet users tend to be younger than the general population, better educated, more positive toward technology-based products, and wealthier. For some markets, this eliminates online testing altogether. If you are polling younger populations especially teenagers or college students, however, online surveys produce a wealth of data compared to traditional formats. Numerous methods are available that you can use to normalize the effects of skewed populations. The best use of online survey research is for those populations that find it engaging and easy to use.

As has been stated before, if you take the time to design your online survey properly, the results can be robust and compatible with traditional surveys. If you are inclined toward complex study designs such as conjoint analysis, some online survey vendors make conditional branching available. This cannot be done easily with traditional survey methods.

But the most important point is that using survey data to determine usability is not reliable. This type of testing involves sending users to sites and asking them to undertake a series of tasks. After each task and at the end of the total task sequence these systems deliver a series of screen-based questions to the user. Under this model, site usability is determined only by subjective opinion of the user. This is **not** a reliable means of gathering data required for making mission-critical design decisions. Even though such tools claim usability testing as a benefit, they do not gather objective data on user behavior with sufficient detail to ensure a reliable profile of customer behavior. The online survey approach does, however, have a significant benefit. You can send customers to competitive sites and have them undertake a task. Yes, even your

competitor's site can be explored and rated by respondents. This research is possible because all of the code that monitors and gathers the customer's responses is on your server or the server of the online research firm. Virtually nothing touches the competitor's site, and there is no way for the site owner to distinguish your test respondents from those who visit the site in the normal course of events.

Tracking Behavior at the Macro Level

Approximately 5% of currently available online tools contain simplified behavior-tracking capabilities. The range of events that can be objectively tracked by current online systems varies depending on what technological approach is taken by the vendor. The implications that such an approach has on behavior tracking are significant. In almost all cases, however, tracking behavior appears to be essentially an afterthought. Most vendors deal with behavior by adding a simplified log file analysis. At best, this produces macro event tracking. Within the context of these specific limitations, however, initial research comparing online and lab-based testing methods is indicating a strong correlation between some important usability variables. For example in a paper presented at the 2002 UPA annual meeting Thomas Tullis²³ reported high correlations between baseline variables including task time and task completion rates. The research showed poor correlations for subjective ratings. It is important to note that micro level behaviors that deal with critical incidents are not discussed in this important paper and remain a primary strength of lab-based testing. Make no mistake, this is a serious limitation and one that online testing vendors work hard to minimize in their sales and marketing efforts.

It is important to understand from the beginning that mapping critical user behavior to user-interface design is complex even in the best of cases. As any professional usability expert will tell you, it is hard work logging hours of video tape data so that time and errors can be properly analyzed. Many of the current online systems do not allow micro behavior analysis normally part of traditional lab-based testing, where a professionally trained usability researcher can probe for critical insights and record micro level events that determine real user performance. If you are concerned with true time and event analysis, including error type analysis, lab-based testing must be combined with online testing methods.

Linking Behavior and Opinions Is No Small Problem

Upon detailed analysis it is clear that many online testing tools are struggling with a way to create a scientifically valid link between user behavior and user-defined subjective data. In the end, online usability testing is most productive when objective and subjective data can be gathered in a valid and integrated manner. Currently even the best online behavior tracking tools have limitations that often limit their widespread use by the professional usability testing community and by web development teams. It is, however, still important to understand how behavior tracking is being developed and what the

²³ An Empirical Comparison of Lab and Remote Usability Testing of Web Sites By Tom Tullis, Stan Fleischman, Michelle McNulty, Carrie Cianchette, and Marguerite Bergel, From Fidelity Investments Contact: tom.tullis@fidelity.com

benefits and limitations are for each approach. Eventually, behavior tracking will become a more robust component of successful online usability testing tools.

Technical Approaches to Behavior Tracking: Limitations and Benefits

All online systems that offer behavior-tracking functions do so by capitalizing on the structured nature of basic internet protocols including TCP/IP. As information bounces and careens around the internet from your customer's desktop to server and back, events take place that can tell us a great deal about the behavior of customers. Some of the information is useful; some is interesting; but most of it is useless. The question then becomes, What is good data and where do we look for meaningful insights? This question turns out to be complex and interesting.

Volley of Data

Events taking place as a result of a customer's mouse click on your web page viewed inside their browser sets in place a series of events that eventually leads back to your server and back to the customer in a type of round robin data exchange. The interesting point is that at various times in the exchange, what is requested, where it is sent, and timed events are captured and stored in the system itself. Most of the detailed information is stored on your server, but some is also stored on the customer's machine. To reiterate: The customer clicks a link on your web page, which sends a request over the internet to your server; your server sends the requested page (data) back to the user, and so on. In theory this is how the web works. The important question becomes, What do these series of events tell us about your customer's behavior? The answer depends entirely on where we look as the exchange of information takes place.

Looking Only at Server Activity - Server-side Analysis

Some online tools look only at the incoming and outgoing requests taking place on your web server. We call this server-side analysis. Log file analysis is a typical method used to view customer behavior using this technique. From this simple data stream we can derive a great deal of valuable data. All of this data is, however, macro level. For example, we can determine the frequency of pages requested, order of pages requested, IP addresses requested, exit pages, entry pages, and hundreds of other data points. Gathering accurate timed events is illusive, however, because of unpredictable internet latency issues. The actual time it takes for the roundtrip data transfer to take place can vary widely based on where your customer is on the internet, the connection speed, the traffic on your server, the customers ISP, and other variables. As has been stated before, at best you can only infer behavioral patterns from this data stream. Still, properly analyzed this data can be very useful as a means of supporting other forms of usability research. The final point on server-side analysis is that we can only infer customer behavior based on where pages, when pages, and how many pages are being requested and sent to customers from your web server. Some vendors imply that a great deal of actual customer behavior can be derived from this data stream. Used in isolation, server-side behavioral analysis is highly limited.

Treating Your Site with Special Research Code

Several technological approaches are used by vendors to add more data to the basic server-side approach. For example several vendors produce a small string of code that must be inserted into the HTML code of every page or just the critical pages of your site as it sits on your server. This approach adds a significant level of detail to your server-side analytics. It also makes it possible for customers to request a survey-type form on any page at any time. These user-requested popup survey forms are only called into view by customers performing a mouse rollover of a small icon located in a common position on all important pages of the site.²⁴ The down side of this approach is the always present question of how conspicuous to make the icon that users must interact with to call forth the popup form. If the icon is too subtle, the stimulus is too low and customers are not even aware of the system. If the icon is too bold, it can interfere with the visual look and feel of the site. This question turns out to be a non-issue if you are willing to explore the conspicuousness of the icon to achieve the proper balance. This type of treated site technology can produce staggering amounts of data that in itself requires special tools and analytical methods.

Transparent Treatment of Your Server with Tracking Code

You can use various online vendors to treat your site with code that is virtually transparent to the users visiting the site. This small string of code allows you to gather significant additional data about your visitors, such as the IP address for each visitor, referral sites, time on page, time on site, and most frequent visitors. This code is normally managed by specialized software from vendors such as WebTrends and dozens of other software firms. This approach does not allow you to poll customers in any way. No surveys and no behavior tracking per se is possible. This type of data can, however, be combined with other research methods to confirm and support critical findings on customer response and usability. Frequently, log file analysis is used as part of professional usability testing projects. We recommend that all clients use some form of treated server analytical tools. Treating your site with code can also be used to enhance the quality of the data derived from each page requested by the users. When we treat our sites with customer-tracking code, significant detail about your customers' behavior can be obtained. For example, we can obtain accurate page order tracking and data on what, where, and when customers are interacting with our sites. Again, from this data we must infer customer behavior because we will know virtually nothing about the customers' opinions of the site or their level of satisfaction. We will not be able to determine in a meaningful manner errors and other problems customers may be encountering as they navigate the sites.

This approach cannot be used to conduct competitive analysis of sites since we cannot treat competitor's sites with the necessary code. On the other hand, this method can be a powerful means of tracking in virtual real-time the implications of changes to a site. By strategic use of this approach, we can selectively monitor critical pages for user behavior.

²⁴ For an example of this technology visit our corporate web site at <http://www.mauronewmedia.com> An OpinionLab Icon is located in the lower right of all screens on the site.

For example, if we make changes to a registration page and see time on page values increase and exit page rates for that page also increase we can infer that changes have had a negative impact on customer usability.

The most critical negative aspect for this approach is the always-present complaint by the IT team that the required code will compromise server performance or security. Some IT teams refuse to put foreign code of any type into the pages of their sites. This refusal is unfortunate because the data flowing from a simply treated site normally outweighs the downside. The second problem with this type of behavior tracking is that it produces staggering amounts of data that must be critically summarized for appropriate insights. Many corporations use server-side behavior tracking but do virtually nothing meaningful with the data that is generated. It is important to have a usability professional aid in the analysis of server-side data if meaningful insights are to be obtained.

Page Caching and Other Wonders of the Internet

A far more important point, and one that online usability testing vendors do not tell us, is that a major innovation of basic internet technology **page caching** causes huge problems. This clever idea that saves a page you already visited on your hard drive dramatically reduces the total traffic on the internet. Without page caching we might not have an internet. If a page is cached and a customer goes back to that page during a session, your server does not know a thing about it - zip - nothing. Therefore, if you are looking at errors on your site by examining navigation into and out of specific page sequences, you have no data. If the user backs up using their browser Back button, and more than 90% of users do this when lost as a primary error correction strategy, your server will never see it and therefore neither will the web development team making enhancements to the interface. Your server sees a perfectly acceptable sequence while the user may be backing into and out of forms with important frequency and related frustration. It is important to note that this problem can be dealt with by experimental design and or use of alternate technological approaches. Usability problems that do not surface in click stream data may be reflected in subjective response to questions by the user. This approach, however, is not always the case. Yes, there is code that you can send quietly to your customer's machine that turns page caching off. But no corporation would even think about using such an approach. So in the end, server-side usability analysis, no matter what its form, is limited and data is confounded and can be highly misleading if you use it improperly. If you combine this type of data with professional lab-based testing it can, however, be very powerful

Limitations Are Important

There are limitations with all server-side implementations for online usability testing systems. For example, we cannot tell what the user is actually doing on a page: Are they scrolling up, down, left, right? What exactly are they looking at? What information are they reading? How are they exploring the data presentation scheme? Are they switching to other browser windows? Did they go to lunch or did the dog unplug the computer? These behaviors are **not** available if the online testing system uses treated server-side

technology. Any local event that takes place on the user's machine that does not result in a mouse click (or other event that causes a message to be delivered to your server) cannot be tracked. The implications of these variables are significant: If you cannot detect behavior, you cannot infer behavior. Recently, various vendors have developed and are testing tools that allow micro-level tracking of user behaviors. These tools will aid in this critical problem and show significant promise for the future.

Looking at Events on Our Customers' Machines

An approach used by some online testing vendors allows us to record browser-level events on the customer's machine. In other words, all events taking place on the user's machine including scrolling, Back button use, and viewing cached pages can be recorded and sent to the research server. The question that springs immediately to mind is why doesn't everyone use this approach? The answer is that to record actual user behavior at the browser level successfully, online usability testing vendors require that the customer download software ranging from an applet to a new specialized browser. With this software installed on your customers' machines it is possible to record more of their task sequences as they take place within the browser functionality. Note that this does not allow us to record anything else that the customer is doing with other applications.

There are two basic limitations to this approach. First, downloading and setting up a browser requires expertise that applies a bias to the respondent population. Vendors who use this technology steadfastly deny that this is a problem. But it is. Users who are comfortable with executing such downloads are at the upper end of internet technical expertise. When testing this user group, MauroNewMedia routinely finds them more positive about all types of interfaces, that they make fewer errors, and that they take much less time performing tasks than customers representing a broader internet profile. This confounds data on several levels; most specifically, it puts a significant positive bias on behavioral and subjective data. Usability problems tend to be minimized, which is not a good thing.

Panels and Other Research Methods

This bias problem has led vendors to attempt to reduce bias by prescreening and developing panels of respondents who have been helped through the download and setup process by phone support. By carefully screening respondents into these testing pools, some measure of bias is reduced. How much is another matter. Once vendors spend the time and effort to set up users with specialized browsers, they tend to reuse respondents for more than one study. This practice is not experimentally appropriate and may further complicate respondent reliability. Finally, vetting of respondents is at best a hit and miss affair as respondents who do not begin to meet the stated profile sometimes show up in these studies. The benefit of this approach is that it is fast. If we work with a vendor such as Vividence (the current best in class for this approach), we will be able to turn studies around almost overnight. This approach is a more reliable means of getting at macro behavior because it tracks gross events that take place within the customer's browser. For the most part, we will know when users back up, how much time they spend on which

page, their page navigation sequences, and their form-filling sequences. We will not know what they are thinking or looking at as they go about their tasks.

Like the first approach, this method has a significant benefit in testing competitive sites. You can send customers to any site and have them undertake take a study without the site owner being aware of the study. This is possible, again, because all the code that monitors and gathers the customers' responses is on your server or the server of the online research firm. Virtually no specialized research code touches the competitor's site, and there is no way for the site owner to distinguish your test respondents from those who visit the site in the normal course of events. It is important to note that a strict statement that only three approaches does not hold. In fact upon detailed technical analysis it can be shown that some vendors use hybrid versions of server-side and client-side methods. Clearly, this is a rapidly changing issue that will require constant updating as vendors produce and upgrade products to deal with known problems and to deliver more robust data to clients.

Third Approach: Semi-Automated Heuristics

A third approach involves a new online technology that models actual user behavior based on software-based "bots". This approach is generally known as **semi-automated heuristics**. The basic theory behind this approach was developed over the past 15 years and involves the idea that by looking at procedural steps undertaken by users in a theoretically optimized path you can infer usability. Numerous attempts have been made to develop systems that can examine the HTML code for each page of a site and infer from the code what the level of usability the site should be for an idealized user. This approach is a form of cognitive modeling and can be useful as a design tool. The primary conceptual model for these systems is based on a generalized mathematical model *n* (developed by the vendor) that is based on the concept that time on task and other variables are strongly correlated with general usability.²⁵ This is generally not found to be true. These tools may, however, show promise for some forms of baseline usability modeling. It is probably too early to make a reasonable estimate as to the final usefulness of semi-automated heuristics.

Unfortunately, in web applications these tools can be wildly misleading because they look essentially at the hard physical elements imbedded in the HTML file, such as the number of buttons, amount of text, number of graphic elements, and number of clicks to reach an identified target field in the site. What these systems cannot do is evaluate the syntax and context of text on the page. This means that a site can be written entirely in Russian and still be judged excellent in terms of usability even though the customers understand virtually no Russian. These systems are interesting, but they are only as good as their underlying models.²⁶ WebCriteria has previously used this type of approach in their online product testing system know as MAX.

²⁵ Thanks to James Landay Ph.D. from NetRaker for his review of this issue and related products.

²⁶ The primary conceptual model for these systems is based on the GOMS system developed by Card, Moran and Newell more than 15 years ago. Other leading researchers including John Anderson at Carnegie Mellow have expanded these conceptual models.

The latest research on Semi-automated Heuristics

Interesting work is currently being undertaken by researchers at UC Berkeley and other major universities on automated heuristics.²⁷ Through the application of factor analysis, web pages judged as outstanding by a jury of web design experts have been analyzed to identify underlying physical attributes that are common to high ratings. Factors include the amount of text on a page, the amount of white space, and the number of different fonts. This approach represents a promising and potentially powerful design tool.

Understanding the trade-offs

The number of products and vendors entering the online market is increasing every month. At best the entire field of online usability testing is a moving target. As a means of assisting in the understanding of the primary functional issues including technical approaches and related usability implications, we prepared the following trade-off matrix. By scanning the rows and columns of the matrix you can obtain a good overview of the primary and secondary issues at work in this important new field of research.

²⁷ See the paper by Melody Ivory and Marti Hearst at (<http://webtango.berkeley.edu/papers/ue-survey/p470-ivory.pdf>) for an excellent discussion of automated heuristics and related research methods.

Part 6: Methodology trade-off matrix

Methodology	Technology implications	Overall usability Testing implications	Impact on opinion tracking	Impact on behavior tracking	Top tier vendor
Online testing					
1 Treated online method					
Client side treatment (adding specialized browser or applet to your customer's machine)	Requires download of special applet or browser on to user's machine. Set up is dependent on users' browser version and configuration. Requires commitment to vendor and data may not reside on your server but server of vendor. Customer may complain about placing code on their site.	Produces significant bias in user profiles and requires use of specialized panels of respondents.	Can deliver wide range of survey data collection formats. Data may be biased based on use of panels.	Allows for behavior tracking of browser level events.	Vividence
Server side treatment (adding specialized tracking code to your server)	Requires imbedding specialized tracking software into the HTML code of your site. Requires commitment to vendor and data may not reside on your server but server of vendor. IT dept may complain about placing code on your site.	Can produce large volume of data sometimes of questionable value. Properly structured data can be a powerful tool in longitudinal studies.	Some applications do not allow opinion tracking; other applications can be used to create innovative, on demand pop-up surveys	Allows macro level behavior tracking based on recording of events flowing to your server from customer's browser.	WebTrends for log file data OpinionLab for online pop-up survey data
Combined online approach requiring client side and server side treatment (adding specialized code to both your customer's computer and your server)	Requires imbedding specialized tracking code into the HTML code of your site and download to customer's machine. Can be costly and complex to manage. You are tightly bound to one or more research vendors, which limits your access to new tools. Tools do not share data easily.	When combined with survey tool can produce detailed subjective and objective data. However produces significant respondent profile bias based on download and set up requirements	Can deliver wide range of survey data collection formats. Data may be biased if respondents are based on panels.	Can produce rich data set if proper tools are combined and analysis is properly executed. Can be complex to set up and use but may produce powerful insights.	Requires combining tools from various vendors. NetRaker is largest vendor with multiple offerings that can be combined under the methodology
2 Untreated online method	Technology implications	Overall usability Testing implications	Impact on opinion tracking	Impact on behavior tracking	Top tier vendor
Online survey tools	No download to either your server or user's computer required	Good for delivering surveys and collecting subjective opinions of usability. Not possible to track user behavior. Survey must be carefully designed for maximum effectiveness	Can be used to test users opinion of any site without treating site in any way. Excellent for simplified competitive analysis. Properly designed surveys can provide robust data	Only delivers opinion data to the research team. This provides subjective view of usability issues, which can be highly misleading.	Many vendors in this space. Over 100 survey tools currently offered.
Online behavior tracking tools	No download to either your server or user's	If properly designed and understood tool	Excellent for competitive	Does not accurately report	Several vendors in this space with



	computer required	can produce relevant data	analysis. Properly designed studies can provide interesting and useful data if combined with survey tool	user behavior after pages are cached on customer's machine. There are methods for addressing this problem.	significant development underway to address this important research need.
Combined survey and behavior tracking methods	No download to either your server or user's computer required. Usually executed in ASP approach so that your system is totally unaffected by study execution. Data normally belongs to you.	Can produce good basic view of high level opinions of usability issues. Study must be properly designed but results can be powerful if combined with lab-based testing.	Excellent for competitive analysis. Properly designed studies can provide interesting and useful data with no requirement for customer to download special software.	May not accurately report user behavior after pages are cached on customer's machine unless system is designed to account for this problem.	NetRaker is a system in this category that address behavior and opinion tracking.
Automated heuristics (<i>this method uses a software-based BOT to analyze the content of your HTML code and then infers usability based on a mathematical model</i>)	Requires access to your site by automated BOT that examines page level code. Process is transparent to site being evaluated and no special down loads required on server or client side.	Interesting approach that in the long run may become a powerful tool for predicting basic levels of usability for systems under development.	System generates artificial opinion of site usability. Can be HIGHLY misleading based on the type and complexity of the site being evaluated and model applied.	System generates artificial description of user behavior on your site and then infers usability. Can be misleading based on the type and complexity of the site being evaluated and model applied	WebCriteria MAX technology Also various tools developed by Melody Ivory and team at UC Berkeley
Lab-based testing					
Heuristics analysis	Requires use of video recording system to log and record experts' interactions with software.	Good for very high level opinions of usability if the expert has clear understanding of business objectives and usability science.	Fast and inexpensive. Can produce very good results on limited set of issues. Good for applying standardized guidelines and procedures.	Data is only as good as the individual executing study and their ability to predict user behavior.	MNM and many other PS firms.
Small group lab-based testing	Requires use of video recording system to log and record user interactions with software. Study must be conducted in formal lab setting. Can be more costly and time consuming than simple online study. Data has short shelf life if changes are made to site.	If executed by usability professional provides highest quality objective and subjective data sets but sample size restricts significance of data. Allows for use of well proven research methods for solving complex usability problems not possible using online tools.	Produces the most reliable and robust view of user subjective opinions and interaction with your interface. Can be used on any profile without bias effects. Allows for complex debriefing and in-depth interviews critical for understanding user behavior and motivation.	The most reliable and proven method for critically documenting and understanding user behavior including detailed definition of errors and time on task both critical in interface optimization and ROI modeling	MNM and other professional service firms listed in the HFES listings as full members.
Large group lab-based testing	Requires use of video recording system to log and record user interactions with software. Study must be conducted in formal lab setting. Can be more costly and time consuming than simple online study. Data has short shelf life if	If executed by usability professional provides highest quality objective and subjective data sets Larger sample size adds significance to data. Allows for use of well proven research methods for solving complex	Produces the most reliable and robust view of user subjective opinions and interaction with your interface. Can be used on any profile without bias effects. Allows for complex debriefing and in-depth	The most reliable and proven method for critically documenting and understanding user behavior including detailed definition of errors and time on task both critical in interface	MNM and other professional service firms listed in the HFES listings as full members



	changes are made to site.	usability problems not possible using online tools.	interviews critical for understanding user behavior and motivation.	optimization and ROI modeling	
Hybrid methods					
4 Combined online and Lab-based approaches	Technology implications	Overall usability Testing implications	Impact on opinion tracking	Impact on behavior tracking	Top tier vendor
Combined untreated online and small group lab-based testing. <i>(this hybrid approach combines un-treated online behavior and opinion tools with small group lab-based testing)</i>	Minimal impact on your site and users machine	If properly designed a combined study produces very robust data set by combining insights gained from online study to refine small group lab-based testing sessions. Allows for use of predictive statistics	Cost effective and produces reliable and robust view of users' subjective opinions and interaction with your interface. Can be used on any profile without bias effects. Allows for complex debriefing and in-depth interviews critical for understanding user behavior and motivation.	Very cost effective, reliable and proven method for critically documenting and understanding user behavior including detailed definition of errors and time on task both critical for interface optimization and ROI modeling	MetricPlus® MNM in combination with online testing tools from various vendors
Combined server side treated online and lab-based approaches <i>(This hybrid approach combines specific treated online and opinion tools with small group lab-based testing)</i>	Requires imbedding specialized tracking software into the HTML code of your site.	If properly designed can produce robust data set by combining insights gained from online study to refine small group lab-based testing sessions. Excellent for longitudinal studies	Low cost and high data rates. By using some methods customers' opinions can be captured continuously and used to improve effectiveness of small group lab testing. Can track changes to site and their impact on opinion in real-time.	More expensive than using only online tool but data is far more rich and meaningful. Delivers significant actionable insights in terms of behavior. Can track changes to site in real-time.	MetricPlus® MNM in combination with OpinionLab
Other methods					
Group focus testing	Standard video recording required	Minimal to negative impact on understanding usability issues.	Opinion on usability biased by group interactions. Data of little value.	Cannot track behavior	Hundreds of market research firms

Better Way To Measure Learning Effects

Although there are significant downside issues with current online usability testing tools, there are also significant upside advantages that are important. The most significant factor is the ability to run longitudinal studies of web site customer responses. This ability has always been a shortcoming of traditional lab-based approach. The importance of this capability cannot be taken lightly. For example, in large lab-based studies where MauroNewMedia has been able to return to a site after users have been active customers for several weeks or even months, we always see significant changes in both behavior and attitude. This fact is explained by well-proven concepts from the field of usability engineering, most specifically research related to skill acquisition. But the important point is that some of these variations in short-term vs. longer-term response to an interface design have been dramatic, which means that initial feedback was either overly positive or overly negative. Some of the tools discussed in this paper address this problem by their very nature, specifically those methods that are continuously available

in user-defined popup tools such as those offered by various vendors including OpinionLab. This is a major reason to consider such tools.

Sample Size Can be a Benefit

It is well known that statistical reliability is increased with increased sample size. Traditional lab-based testing, however, often has a small sample size. There are ways that traditional lab-based testing gets around this problem, such as targeted samples and distribution sampling. There is, however, no substitute for large sample sizes of several hundred respondents. A large sample dramatically improves significance in most but not all studies. The question of sample size is complex and should be dealt with by consulting with usability testing experts well versed in study design. It is important to note that you will receive more data and have better statistical reliability if you do a lab-based study using 60-80 respondents than you will have with an online survey using thousands. This fact is because of the fundamental limitations in the fidelity of data that you can gather even with the best online tools. There are instances where a large sample size is the only way to go. Such cases involve questions of naming, visual branding, and other highly subjective variables. In these cases larger sample sizes have no substitute.

Data Capture

The time usability professionals spend on data capture in traditional lab-based testing is about 20% of the total project time. With most online tools, this phase of research is automatic. This does not mean that analysis of the data is automatic unless you are only looking for top line results. In that case most online tools cannot be beat. In the best products mentioned in this paper, data capture and display occur in real time. As respondents complete their surveys, we see the results. This has both positive and negative issues. Early results can be misleading, and not everyone knows how to interpret the data as it comes in. A word of caution: Do not allow anyone but the research director to have access to the data until most of the respondents have finished the study. Above all, do not give access to the real-time displays to top management. Your e-mail file will look like *War and Peace*. It is also worth noting that all good online tools allow you to download your data from the study into Excel formats. This feature eliminates data entry errors when coding survey forms and observation sheets. On average, data entry errors will range between 2 and 4% for manual entry of survey data. Cross-checking is the only way to deal with this problem in traditional studies.

Data Analysis

Most currently available online tools provide basic statistical summaries that deliver top-level results such as mean, range, and standard deviations. Few of the current tools provide more sophisticated tests of significance or correlation analysis, both critical in any important study. Therefore, you must download your data and run these types of analyses in Excel, SPSS, Statview, or other comparable program. This task is not simple because of the need to align and coordinate import/export functions. In any important study this depth of analysis should be left to a usability testing or market research professionals. Although it may seem a bit harsh to state this, but drawing conclusions

from online research data is usually not the strength of online data-gathering processes and should rarely be trusted to the professional services staff of the online tool vendor.

Data Is Not Information

In concluding this discussion, it is important to point out a benefit of online testing that can quickly turn into a major problem. When you use online testing methods and gather data from a large sample size, the data will be overwhelming if any portion of the study allows open-ended comments or verbatim data. The online testing vendor, and more importantly, your staff will have no reliable way of dealing with these responses. Yet this data always contains highly relevant insights. To help resolve this problem, MauroNewMedia developed and uses a software-based analytical system called *Comment Tool* to examine these files. We routinely work with studies that return 5000 to 150,000 open-ended comments. Extracting meaningful data from these files is no simple matter. When properly executed, however, the data is both rich and robust. *Comment Tool* can be used on any file of comments from any online testing vendor.

A special note of caution on survey design. Many online testing vendors offer to throw in study design as part of the total cost of the project. You may save money, but your data can be seriously compromised. It is better to save money by cutting back on the number of respondents instead of accepting a free study design offer.

Incentive Payments and Respondent Management

Finally, online testing offers benefits in acquiring respondents and providing them with appropriate incentives. As any market research professional knows, respondent recruitment and screening is complex and difficult. People do not show up, they show up at the wrong time, they show up late, they show up sick, they leave in the middle of studies, they are cranky about the muffins. There is no end to the problems of respondent acquisition and management. Online testing reduces these problems by allowing respondents to take the study at home or at work. This solution can save time and be less costly. In major metropolitan areas, the standard recruiting fee for lab-based studies ranges between \$50 and \$150 per respondent. This fee is normally paid to a recruiting service or testing lab. Online recruiting fees can be dramatically less depending on your profile requirements. In the end, what you need is a respondent's first name and e-mail address. It is the responsibility of the recruiting service to verify profile details and to ensure compliance. There are several agencies that do a good job of online respondent recruiting. A word of caution: No matter how much you pay for recruitment, respondents show up who are wrong or not qualified. In the hundreds of usability studies undertaken by MauroNewMedia during the past 25 years, we frequently have had respondents recruited and screened who simply were not qualified. Agencies make mistakes or fill in with wrong people at the last moment hoping you will not notice. People lie about their background to get into the study. Lab-based studies always do a three-question cross check when the respondents show up. There is no way to do this with online testing. It is safe to say that in studies that require a highly vetted respondent pool, online testing is not the way to go.

The ability to restrict who takes an online test and who sees a new site is **not** protected or controlled under online testing conditions. For example, MauroNewMedia recently received three new web site concepts from a colleague who took an online test for a competitor. If you do not want the competition to see your site do not test it online. If you do, it is safe to assume that anyone will be able to see your online test.

Incentives

One aspect of formal usability testing that online testing does well is respondent incentive fulfillment or payment to respondents for successfully completing the testing session. Some vendors are much better at this than others. There is a huge range of options available for incentive fulfillment, from sending respondents a check within two working days to providing Amazon gift certificates. The important point is that incentive fulfillment is handled by all best in class vendors in a reliable and transparent manner. An important point to remember when planning an online study is the cost of incentive fulfillment itself. If you have a study of 500 respondents (great for statistical reliability) and each receives a \$25 gift certificate you have just spent \$15,000 on incentives plus a dollar or so for incentive handling. This is a real cost that you will not normally see discussed by most online testing vendors.

Part 7: Selecting the Right Approach

In the end, it is not a question of online vs. lab-based usability testing. The most important point is that mission-critical decision-making must be based on a detailed understanding of human skills and limitations. These types of decisions can only be made with a full and direct understanding of the behavioral and attitudinal attributes of customers as they interact with our screen-based delivery system. All forms of professional usability testing including lab-based and online methods help address this critical research need. If you take away only one insight from this paper let it be that there is a new science that deals with the connections between computer and user and that field will play a central role in the future of all successful E-Com systems. That science is professional usability engineering and testing.

Selecting the Right Research Approach and Tools

How do we decide on an approach that will be cost effective yet yield the best results? Are we better off with a long-term relationship with a large online testing vendor or with a lab-based approach? An effective model for determining which approach to take is to ask yourself the following question: What is the impact of making the wrong decision when determining what to change or optimize in the design of my E-Com initiative?

This question is graphically represented in Figure 5. When we look objectively at the complex question of determining if a mission-critical decision is right or wrong, it is important to take multiple views of the problem. The figure combines several issues that interact with each other to produce insights into this complex question.

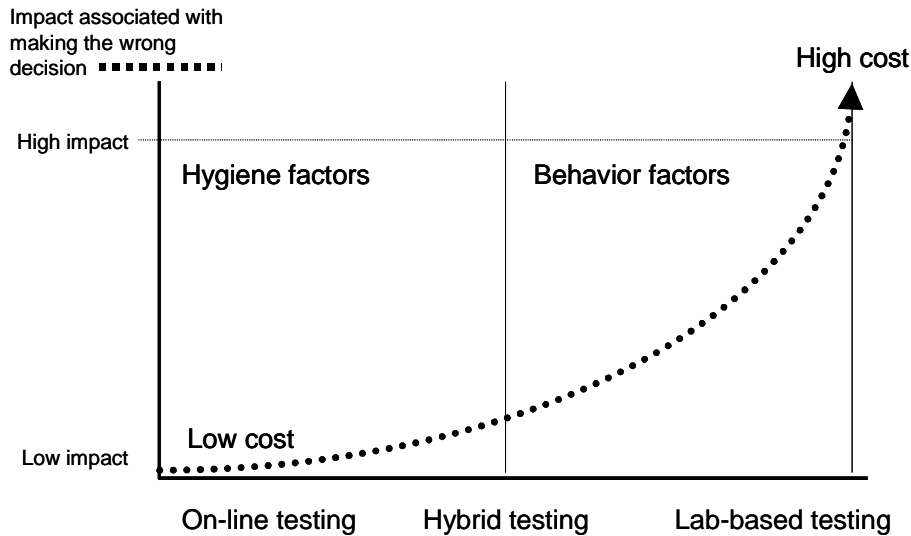


Figure 5

Low Impact Issues

If we are faced with questions related to hygiene factors, such as visual design acceptability or visual brand transfer, then lower cost online testing can be very useful. In this case, we can expect reasonable confidence when making decisions on overall graphic impressions and visual branding. In fact, these issues are often better addressed with online testing methods where larger sample sizes are helpful and there are no possible error states resulting from page caching or population bias resulting from installation of specialized browsers. It is important to note that the reliability of decision-making with respect to visual design concepts and graphic visual branding can be dramatically improved by the use of online testing tools. It is also true that making a mistake in such factors (assuming you meet baseline criteria) is small. Please note that how such studies are actually designed is far more important than the online tool selected.

MauroNewMedia strongly recommends that you never compromise on study design and related costs. A poorly designed online study is still a bad investment no matter how automated and fast the response. At MauroNewMedia we use a visual design testing methodology refined and updated over a 15-year period. Testing for visual design factors is an excellent application for online testing methods. If you are faced with more complex issues such as customer acquisition, retention, and migration, the cost of making an error in design execution can be career threatening.

Acquiring Customers Is a Complex Issue

When faced with customer acquisition questions, online testing tools begin to lose their cost benefit. Overall, the concept of online customer acquisition design is complex and tightly interwoven with issues of learning transfer, stimulus response compatibility, and concept formation. These are issues that require serious experimental design and cognitive modeling. They do not lend themselves well to online testing methods because

it is critical that the customer be carefully and professionally interviewed for key concepts. We have found that writing standardized questions to address such complex problems is nearly impossible. More traditional lab-based testing, therefore, begins to assume a larger role in increasing the reliability of our decision-making process. If we are being asked to address critical questions related to increasing customer acquisition rates, we cannot conduct research solely with online tools. The fidelity and richness of the data will simply not be enough to dramatically reduce the probability of making the wrong decision.

Retaining and Migrating Customers Is also Complex

At the heart of customer retention and migration is the subjective and objective behavioral interactions and experiences of the customer with your web-based delivery system. Have no illusions about these issues. They are totally determined by the interaction between people and machine or web site and customer. This is **behavior** in its most complex and relevant form. Therefore, if we are addressing issues of low retention and migration rates we must seriously consider high-quality, professionally executed lab-based research. When faced with consulting assignments that deal with these issues, we routinely use combinations of several tools. The cost of making the wrong decision when faced with these questions means the difference between success and failure. Many large web development groups, however, have strongly resisted this recommendation and either ignore this development tool or allow market research department to handle this critical area of development. This is not the way of the future. The usability must go in before the graphics go on. Entire industries have adopted professional usability testing and engineering. It is about time web development teams did the same.

Most Important Problem of All

Whereas there is a great deal of discussion in the web development field about what methods and processes to use in the creation of powerful and effective web-based products and services, there is a much bigger problem. Today the single largest problem facing web development teams is an almost complete lack of professionally defined business objectives. This issue holds true for even the largest corporate clients. Few E-Com development teams possess clearly articulated and well-defined business objectives that can be mapped to underlying consumer behavior in a web-based environment. In simple terms, this lack makes design and development of successful E-Com initiatives a hit and miss affair. More importantly, poorly defined business objectives make it virtually impossible to bring a professional level of design decision-making to such efforts. Therefore, it is clear that no amount of clever online research tools, good experimental design, or complex statistical analysis will cover up poor business objectives development.



Before we use any of these methods, it is critical that we put in place a well-reasoned and fully vetted set of business objectives known as User-Centered Design (UCD). UCD can be helpful in framing business objectives that can be mapped to customer experience design. In the end web-based delivery systems can only be optimized against a set of performance standards that are strategically and financially driven. To do otherwise is to waste your corporation's profit for certain loss.

Thank you

Charles L Mauro
President
MauroNewMedia
cmauro@mauronewmedia.com



Part 8: Comprehensive Approach / *MetricPlus*[®]

As part of an ongoing effort to create value for large and small developers of effective web-based products and services, MauroNewMedia has developed tools and systems that integrate research and methods from the usability sciences into a comprehensive product offering. *MetricPlus*[®], our combined online and lab-based testing system, is one such tool. For more information on *MetricPlus*[®], please refer to *About MetricPlus*[®] in this paper.

Charles L. Mauro
President
MauroNewMedia
10/14/02

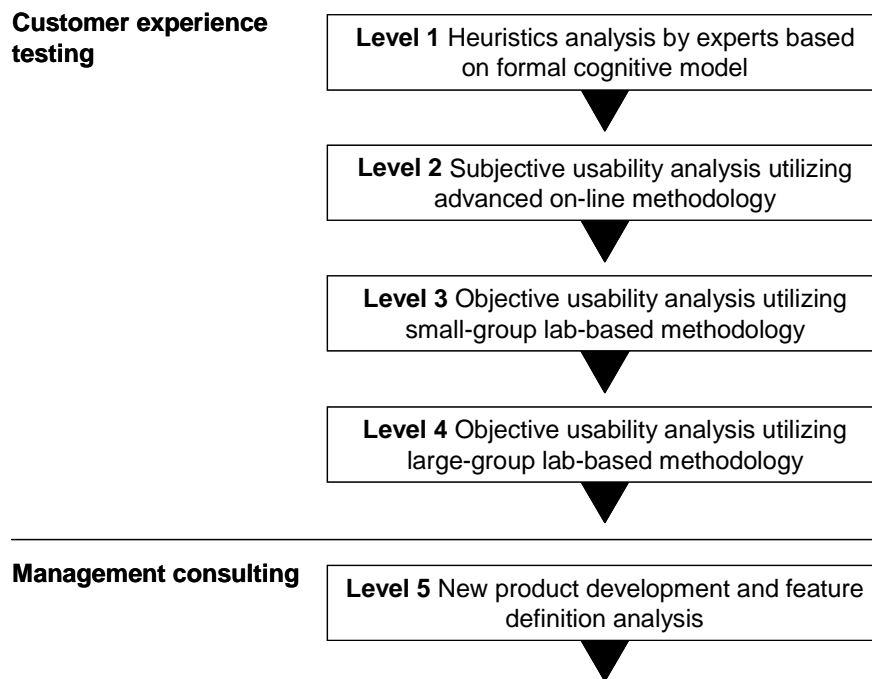
Version 10

Part: 9 About *MetricPlus*®

Effective Customer Experience Optimization System Offered By MauroNewMedia

It is becoming increasingly clear that the creation of a powerful, engaging, and profitable screen-based delivery system is far more complex and demanding than was imagined during the first generation of internet-based E-Commerce. It is clear that there is a pressing and important need for a comprehensive approach to this difficult problem. As a means of addressing the demand for an integrated approach to customer experience optimization, we developed *MetricPlus*®. This usability testing and user-interface design development methodology provides a means for critically determining the proper balance between online and traditional testing methods. More importantly, *MetricPlus*® is structured to apply formal expert resources in a consulting environment to a wide range of usability and customer experience optimization issues including business objectives development, team conflict resolution research, and formal user-interface design and concept testing. *MetricPlus*® combines more than 25 years experience in formal usability research and user-interface design with the latest online testing methods and management decision theory. It has been created to address the complex usability testing and user-interface design issues faced by development teams as they undertake mission-critical updates and baseline re-designs of their important web-based properties.

With *MetricPlus*® you can choose from five levels of technical consulting services tailored to meet the specific business objectives of the sponsoring organization. Projects range from short-term usability testing assignments to comprehensive site optimization research.



In many of the projects using *MetricPlus*[®] a combination of online and traditional lab-based testing methods has been used. For example, it is often possible to follow online testing sessions with focused lab-based testing based on data derived from the online research. This approach adds significant depth and focus to traditional usability testing projects. In projects where visual branding and visual look and feel are being defined, we find that online testing methods, if specifically designed for these issues, are without peer. Online testing is also effective in administering product attribute trade-off analysis (commonly known as conjoint analysis). This type of study can be effective in determining what combinations of attributes are most preferred by customers of either online or even traditional products and services. A word of note: These studies are complex and costly to design and conduct online but the data can be robust and reliable.

When faced with questions of increasing customer acquisition, retention, and migration rates we need to use combinations of tools and approaches. These questions always involve integration with market research, strategic planning teams and are complex and difficult projects. By using the proper combination of expertise, insight, and research methods, professional usability engineering and testing can dramatically reduce the probability of making the wrong decisions. For a detailed description of *MetricPlus*[®] please send an email request to Cmauro@MauroNewMedia.com.²⁸

²⁸ MetricPlus is a copyrighted software development methodology which has been used in the development of some the most successful screen-based products and services currently in operation today. For more information on MetricPlus and our partners please send an email request to Cmauro@MauroNewMedia.com.

Part 10: About the Author

Charles L. Mauro
President and Founder
MauroNewMedia

Charles Mauro holds an MS in Human Factors Engineering from New York University. At NYU Rusk Institute of Rehabilitation Medicine he was a NIOSH research fellow. He received grants and fellowships from the Ford Foundation, National Institute Occupational Safety and Health, and The National Endowment for the Arts.

Since 1975, Mr. Mauro has been President of MauroNewMedia, a leading provider of professional usability engineering and high-performance user-interface design consulting services. Before forming MauroNewMedia, Mr. Mauro worked directly with product design pioneers Henry Dreyfus and Raymond Loewy. During Détente he managed the first product development program undertaken by a U.S. firm for the Soviet Ministry of Science and Technology. He was responsible for research and development of mission-critical user interfaces, including designing primary trading systems used on the floor of the New York Stock Exchange.

Mr. Mauro's experience spans more than 25 years and includes consumer, commercial, military, and aerospace applications. He consults on a regular basis with Fortune 500 clients and leading start-ups.

Mr. Mauro has received numerous citations and awards, including the Alexander C. Williams Award from the Human Factors and Ergonomics Society and citations from NASA and Association of Computing Machines (ACM). He served on national and international panels and chaired two ANSI standards committees. Mr. Mauro also served on the Presidential Design Awards Program for the NEA and was a founding member of the Human Factors Society Special Interest Group on Consumer Products.

Mr. Mauro consults on a regular basis with leading corporate CEOs on strategic issues of screen-based customer experience design, usability, interactive brand development, and fixed-to-virtual migration of products and services. He has been accepted in federal court as an expert witness for design patents, trade dress, and other intellectual property issues related to user-interface design. He holds many U.S. and international patents. Mr. Mauro is widely published in the professional and popular literature and has been quoted in *Fortune*, *Business Week*, *The Wall Street Journal*, *Science*, and other leading business publications. He routinely speaks at national and international conferences on design, usability, and web development. Mr. Mauro lectures at leading graduate programs



including MIT Sloan School and Stanford. His first major book *Usability: the Bottom Line* will be published in 2003.

Part 11: About MauroNewMedia (MNM)

Founded in 1975, MauroNewMedia (MNM) is a leading provider of professional usability engineering and user interface design solutions for mission-critical software applications. MNM expertise is focused on the creation of online customer experiences that provide outstanding performance in terms of user acceptance and overall ease of use. These claims are objectively determined through professional usability testing.

MNM is not a vertically integrated software development firm. Our focus and expertise profile is solely user interface design based on the application of professional usability engineering and testing. MNM often joins with leading strategic planning, engineering, and software development firms to solve problems that meet the business objectives of the sponsoring organization. The primary expertise profile of MNM covers professional human factors engineering and user interface design. Principals and primary consultants hold advanced degrees in either human factors engineering or user interface design.

The solutions of MNM are currently running at the heart of the world economy. These solutions contribute on a daily basis to the productivity and profitability of several of the world's most successful institutions and government agencies. Such clients include The New York Stock Exchange, Goldman Sachs and Co., and NASA.

MNM has extensive user interface design experience in several service categories. This experience includes a demonstrated history of creating effective solutions for user profiles ranging from retail consumers to high performance institutional traders. MNM expertise, when combined with that of state of the art online research tools, provides an experienced professional consulting resource for addressing mission-critical E-Com development problems.

For more information on MNM please visit our corporate web site at <http://www.MauroNewMedia.com> or our public interest site on User-Centered Design at <http://www.taskz.com>

Part 12: Informal peer review and acknowledgements

The author would like to thank the following usability professionals, vendor representatives, colleagues, and documentation experts for their review and recommendations related to this document. There were others whose time and effort are acknowledged anonymously...thanks to all.

It is also important to note that the entire field of on-line usability testing is rapidly undergoing new and exciting changes. Therefore it is impossible to cover advances that may have occurred following production of this document. Finally, it is important to note that the limitations and benefits of on-line vs lab-based testing are the opinions of the author and that vendors and other professional usability may not be in complete agreement with the finding stated in this paper. However, it has been our goal to produce an objective document without the force or impact of individual vendor points of view. We hope to have achieved this important goal. Please send your comments to Charles L. Mauro at Cmauro@MauroNewMedia.com.

- **Dr. Corrado Ronchi Ph.D.:** Thanks for special assistance in reviewing products and testing systems, for critiquing drafts, providing recommendations, and for reviewing technical descriptions and definitions.
- **Tim Dowling Ph.D. Associate Professor University Connecticut:** Thanks for reviewing drafts and providing recommendations related to online survey tools and behavior tracking methods
- **Dr. Peter Mitchell Ph.D. President Ergo Research Inc.** (peter@ergo9.com) Thanks for reviewing the final draft and providing recommendations and comments on content and methods.
- **Sarah Hiner** Thanks for reviewing the initial and final drafts and providing recommendations and comments helpful in clarification of survey design and overall technological approaches
- **James Landay Ph.D. Chief Scientist, NetRaker Corporation:** Special thanks for reviewing the final draft and providing recommendations and comments on the overall field of usability science and clarification of online testing methods and practices.
- **Harriet Serenkin:** Special thanks for recommendations and comments on content and style and for her assistance in editing. (serenkin@abacus96.com)

Part 13: Recommended Reading and additional information

For more information on user-centered design, usability engineering, and other aspects optimizing the usability of products and services visit <http://www.taskz.com>.

Relationship with vendors mentioned in this publication

MauroNewMedia supports the field of online usability testing. As part of our on-going research effort for clients, we use the tools and methods of several vendors mentioned in this paper. We do not, however, expressly support nor recommend the tools and methods of any one vendor.

Charles L. Mauro
President
MauroNewMedia