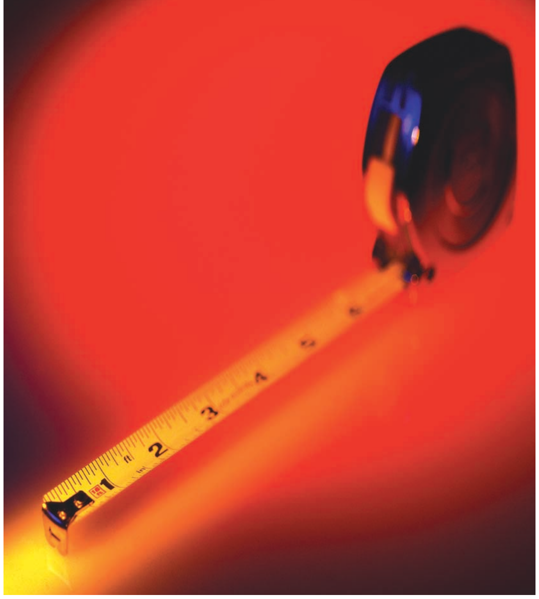


Does your Web site function smoothly enough to deliver government services? Combining evaluation techniques gives you a multidimensional answer.

Fred B. Wood, Elliot R. Siegel, Eve-Marie LaCroix, Becky J. Lyon, Dennis A. Benson, Victor Cid, and Susan Fariss

A Practical Approach to E-Government Web Evaluation

The World Wide Web revolution has now swept through much of government, as well as the private sector. Over the last five years, federal, state, and local governments in the United States have progressed from little or no use of the Web for delivering government services to, in 2003, using the Web as a major, and increasingly as the primary, means of service delivery. At the federal level, the 1996 Clinger-Cohen Act and the 1993 Government Performance and Results Act (GPRA) mandate not only efficient delivery of services to the public but the use of performance measurements to verify success. Thus, usability and performance have become integral components of Web site deployment as a key platform for e-government. IT professionals in and out of government need a practical framework with which to design and implement Web evaluations.

Just as it was important to evaluate earlier modes of service delivery—using face-to-face meetings, telephone, and paper mail—it is equally important to apply appropriate methodologies to ensure that Web-based service delivery meets customer and citizen needs. Relying on any single evaluation strategy is likely to yield incomplete, misleading, or erroneous results.

MULTIDIMENSIONAL APPROACH

Using a robust, multidimensional Web evaluation strategy is key to successfully evaluating Web-based e-government (see Figure 1). Web evaluation methods fall into four major classes:

- *Usability testing.* This category includes various techniques for obtaining feedback from a limited number of experts or users, the latter typically in a controlled laboratory environment.
- *User feedback*. Many methods exist for getting direct, usually qualitative feedback from actual Web site users.
- *Usage data*. In this category are various approaches for collecting quantitative data about Web usage levels, primarily from Web log analysis.
- *Web and Internet performance data*. These methods involve measuring the Web site's technical performance, using metrics such as latency, availability, and data transfer rate.

Specific methods are appropriate for obtaining different types of information at various stages of the Web site's life cycle. Table 1 lists various evaluation methods and their relevance at each stage.

USABILITY TESTING

Usability testing techniques involve obtaining feedback on Web site design and functionality

Published by the IEEE Computer Society

1520-9202/03/$17.00 © 2003 IEEE

either from experts or from users in a controlled laboratory environment.

Heuristic or expert review

In this type of usability testing, a Web usability expert reviews your Web site, compares it against generally accepted Web design and functionality principles and standards (a practice known as heuristic analysis), and suggests design improvements. These may include site layout and structure, navigation tools, search function, fonts and colors, and so forth.

Strengths. Expert review brings an independent, outside perspective to your Web site development. It provides a larger context for Web developers and applies the cumulative learning and expertise about what works best in Web design.

Limitations. Outside experts may not understand the intended audience or the government policies and budget constraints limiting your Web design options. Additionally, no matter how objective they are, experts have their own biases and opinions.

Typical cost. $5,000 to $10,000.

Usability lab testing

For this type of evaluation, your organization invites a small number of users to participate in structured testing of your Web site. The users perform a series of tasks using the Web site, and test facilitators monitor and record user behavior and cognitive processes to better understand exactly how users

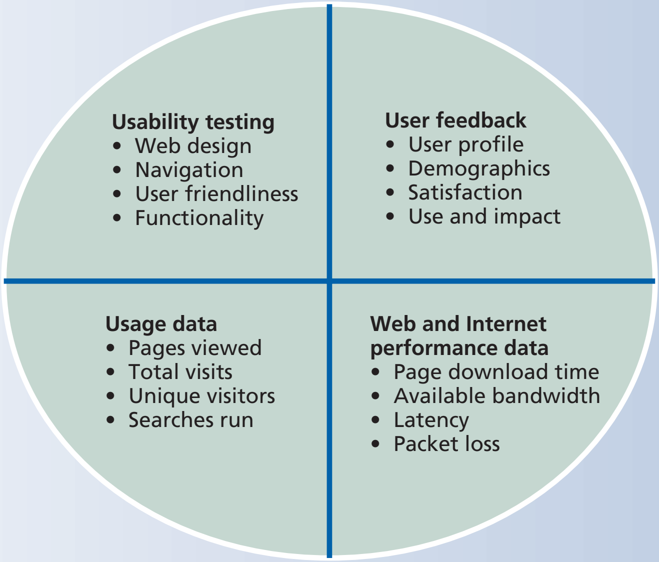


Figure 1. Web evaluation: A multidimensional approach.

Table 1. Criteria for selecting Web evaluation methods.

	Web site life cycle stage		
Evaluation method	**Development**	**Operations**	**Improvement**
Usability testing			
Heuristic or expert review	✓✓		✓✓
Usability lab testing	✓✓✓		✓✓
Informal usability feedback	✓		✓
User feedback			
Online internal user survey		✓✓✓	✓✓✓
Online external user survey		✓	✓
Focus group	✓✓	✓✓	✓✓✓
Nationwide syndicated survey	✓	✓✓	✓✓
Unsolicited user feedback		✓✓	✓✓
Usage data			
Web log data analysis		✓✓✓	✓✓✓
Internet audience measurement	✓	✓✓	✓✓
Web and Internet performance	✓	✓✓	✓✓

✓✓✓ = Very important; ✓✓ = Moderately important; ✓ = Less important;
No check mark = Generally not applicable or not important.

navigate and attempt to find specific information. Facilitators also solicit users' opinions about the Web site design, functionality, and user friendliness, and they ask for users' suggestions for improvement.

Formal testing typically takes place in a computer lab with a workstation that can record keystrokes and spontaneous comments. The workstation also typically videotapes the user's hand movements. Usability testing usually includes between six and 12 users, tested individually. This type of testing is an iterative activity that an organization should conduct periodically if possible. In the early design stages, this type of testing can use paper mockups and prototypes of Web pages.

Strengths. Usability lab testing provides very detailed and specific feedback on site design and functionality. The resulting written and video record of user interactions with the Web site facilitates a rigorous, objective usability analysis from the user's perspective. Frequently, usability lab testing identifies unanticipated problems in site design and navigation.

Limitations. In part because of cost, organizations can usually do lab testing with only a small number of users; generalizing from the results to a large user community can be difficult. However, some experts believe that usability testing with six to 10 individuals will normally capture the vast majority of key design or navigation issues—assuming a relatively homogenous user community. If your organization has several distinct and different user segments, it might need to conduct several usability lab tests with users from each segment.

Typical cost. $15,000 to $30,000 for testing with a dozen individual users. This is the all-inclusive cost, covering recruitment, protocol development, testing lab, test implementation, analysis, and reporting of the test results. Providing some of these items in house can reduce costs.

Informal testing's most effective use is in testing specific design issues.

Informal usability testing

Here, your organization conducts usability testing informally, with individuals or small groups of users working at individual workstations or in a standard computer lab setting, instead of a specially instrumented usability lab. Typically, the informal settings don't let you record keystrokes or hand movements, so this approach relies instead on the facilitators' observations. Facilitators ask participants to perform a series of tasks and answer questions.

Strengths. Informal testing complements formal testing. Its most effective use is in testing specific design issues—for example, where to place a button, or the choice of formats and colors. You can repeat informal testing on various issues intermittently during a Web design or redesign process.

Limitations. This type of testing usually does not permit rigorous analysis of user feedback. It provides a more qualitative than quantitative perspective, and is not a substitute for usability lab testing.

Typical cost. This type of testing incurs no direct cost if the client organization has or has access to a suitable computer lab and has a reasonably skilled facilitator on staff.

USER FEEDBACK

User feedback methods involve collecting mostly qualitative feedback directly from Web site users. The "users" could be the organization's own users, members of external panels of users, or a subset of the general population.

Randomized online user surveys

By this method, a random selection of Web site users have the opportunity to respond to a pop-up survey when they visit the site. If users elect to respond, they either take the survey on the main Web site or are redirected to another site. In either case, after they complete the survey, the server returns users to the page where they received the pop-up survey request. The number of questions typically ranges from five to 20, and the average completion time is typically five minutes or less.

Strengths. A randomized online user survey yields results with higher statistical validity than a self-selected, "bounceback" survey, which lets anyone respond. The online survey capability lets you efficiently survey large numbers of users in a relatively short period of time. Heavily trafficked Web sites can usually obtain 1,000 to 3,000 respondents over a two-week period, a sufficient number to provide high levels of statistical validity. The online electronic format facilitates quick pretesting and easy modification of the survey questions as needed. It also expedites the data collection, analysis, and reporting process, compared to paper-based mail surveys of the pre-Web era. In addition, the organization can benchmark the results against other surveys.

Limitations. Average response rates to online user surveys are in the 5 to 10 percent range. The "nonresponse bias" is still an issue: Some types of users, such as first-time or low-frequency users, might be less likely to respond to the survey at all. Thus, the results can be skewed toward more frequent users.

Typical cost. $20,000 to $30,000 for a survey with 1,000 to 3,000 respondents. This figure includes survey instrument design, technical setup, pretesting, and data collection, analysis, and reporting.

Online "external" user panel surveys

Here, rather than surveying users when they visit the Web site on their own initiative, the organization offers the survey to members of an external user panel who are directed to the Web site. Several private companies maintain user panels, ranging in size up to several tens of thousands of people. The company asks various subsets of these panels—for example, people from specified geographic areas or racial, ethnic, or age groups—to participate in an online survey. This approach usually generates higher response rates than randomized user surveys, and you can use it to obtain comparative user feedback on several Web sites. However, this is no substitute for surveying your own users directly when they visit your Web site. Costs are comparable to randomized surveys.

Focus groups—in person or online

In this type of evaluation, a small group of users (typically six to 10) provides feedback about a Web site. A moderator follows a prepared script with a series of queries about the site and, sometimes, provides exercises for users to try on the site (for example, search for information on a specific topic). A focus group session typically takes about 45 minutes to an hour. Organizations traditionally conduct focus groups face-to-face, but "virtual" focus groups—in an online environment similar to a moderated chat room—are possible.

Strengths. Focus groups provide deeper insights into user feedback about a Web site, because they allow interaction between users and the moderator. In-person focus groups permit more flexibility, consideration of nonverbal cues and responses, and generally deeper discussion. Online focus groups are very efficient; they minimize or eliminate travel (for participants and client staff) and level the playing field by assuring more balanced participation. (Domination of the discussion by the more talkative people is less likely online than face-to-face.) In addition, online focus groups permit automatic recording and transcription, facilitate geographic diversity among participants with the desired demographic profile, and let the client make back-channel suggestions to the moderator as the session progresses.

Limitations. Focus group results do not generalize to the entire user community. Compared to online focus groups, in-person focus groups can be difficult to schedule and arrange. However, by definition, online focus groups exclude people not using the Web, and they tend to self-select the more Web-savvy users. Focus group results depend strongly on the facilitator's skills.

Typical cost. $10,000 for each in-person focus group; about $5,000 online. This figure includes logistics, facilitator, transcription, reporting, and participant recruitment and compensation.

Resources

Web sites

- **"Evaluating Our Web Presence: Challenges, Metrics, Results," presentation abstracts from a symposium cosponsored by the National Library of Medicine/National Institutes of Health and CENDI, a US Government interagency science information group, 2001; http://www.dtic.mil/cendi/activities/04_17_01_eval_pres_abstracts.html.**
- **Internet2 End-to-End Performance Initiative, a resource for measuring, evaluating, and improving performance of the Internet and Web infrastructure; University Corporation for Advanced Internet Development; http://e2epi.internet2.edu.**
- **Usability.gov, a resource for designing usable, useful, and accessible Web sites and user interfaces; Communications Technologies Branch, National Cancer Institute, National Institutes of Health; http://www.usability.gov.**

Publications

- ***Designing Web Usability,* Jakob Nielsen, New Riders Publishing, 2000.**
- ***Measuring the Difference: Guide to Planning and Evaluating Health Information Outreach,* Pacific Northwest Regional Medical Library and National Library of Medicine/National Institutes of Health, 2000; http://nnlm.gov/evaluation/guide.**
- ***User-Centered Web Design,* John Cato, Pearson Education/Addison-Wesley, 2001.**
- ***Web Metrics: Proven Methods for Measuring Web Site Success,* Jim Sterne, John Wiley & Sons, 2002.**
- ***Web Performance Tuning,* second edition, Patrick Killelea, O'Reilly & Associates, 2002.**

Nationwide syndicated survey

By this method, the organization buys access to the results of third-party surveys of a randomized sample of online users of specified information or other Web-based services—these surveys are usually conducted by telephone, but sometimes by mail or e-mail. The typical sample size is 1,000 to 2,000. Sometimes the survey company also conducts a comparison survey of offline users of the same type of services, with a sample size of 500 to 1,000. Private-sector companies most commonly conduct these types of surveys on a syndicated, multiple-client basis to reduce the per-client cost to an affordable level.

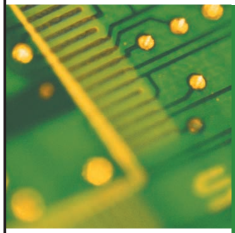
Strengths. Such surveys provide a nationwide look at user views, preferences, and behaviors in a defined market segment that would otherwise be unaffordable for most organizations. On the national level, these surveys typically provide results with strong statistical validity—this is less true for regional or local breakouts.

Limitations. Results from this type of evaluation can sometimes be too general. The syndicated survey is not a substitute for surveying actual users; some of the questions can be off point because the survey company is accommodating the interests of a wide range of clients in a single survey instrument.

Typical cost. $25,000 to $40,000 is the typical annual subscription cost for a survey with 2,000 respondents. This figure includes full reporting, data tabulations, and special analyses.

Unsolicited user feedback

Most Web sites provide one or more opportunities for unsolicited user feedback—for example, via an e-mail box, a link on the homepage, or a help desk phone number. Informal feedback from individual users can provide valuable insights into customer needs, problems, and preferences. Organizations can use this feedback for troubleshooting, identifying possible new or modified features, and developing questions or exercises for more formal surveys or usability testing. The cost consists primarily of the staff time required to monitor, analyze, and respond to the feedback.



USAGE DATA

Usage data is quantitative data about the Web site's usage levels. This kind of evaluation usually involves Web log data analysis or the collection of similar data by companies that measure Internet usage.

Web log data analysis

To make this analysis, your organization most likely uses Web log software installed on the Web site server to collect usage data. Several commercial off-the-shelf (COTS) and open-source software products are available for Web log analysis. The Web logs typically collect data on pages downloaded (or page views), total visits, and unique visitors—the three most commonly used and widely accepted Web metrics. The analysis software reports various drill-down data, such as frequency of visits, referring URL (where the user came from), frequency of page use and search terms, and place and country of origin (when discoverable). To protect user privacy, the software collects no personally identifiable information.

Strengths. This kind of evaluation provides a wide range of quantitative data on usage at relatively low cost—COTS software will usually suffice. It lets your organization track overall usage trends over time, and you can compare log data with other sources of usage information (such as online surveys and external Internet audience measurements).

Limitations. The error factor in Web log data analysis is significant because the software measures usage by tracking the IP addresses of the computers being used, not of the users themselves. This tends to undercount users in institutional settings such as libraries—where many people use a computer with a single fixed IP address—and users who are redirected through proxy servers used by Internet service providers. On the other hand, it tends to overcount individual users that have computers with dynamic IP addresses, where a new IP address is assigned each time the user logs in; these could be incorrectly counted as a new, different user at each login. In addition, special metrics—for example, "number of database searches conducted"—might require custom software.

Typical cost. $500 to $2,000 for a COTS software site license, plus staff cost for data analysis, presentation, and software maintenance and upgrades.

Internet audience measurement

By this technique, private companies collect usage data from large panels of Web users who agree to have their Web surfing monitored constantly. Each participant's computer collects data on all his or her Web use; the company server then aggregates the data. Panel sizes range from about 50,000 to 1.5 million participants, covering usage in US homes plus—in some cases—office, school, and international usage. The companies use demographics and census data to extrapolate the usage data to US or global estimates.

Strengths. This technique offers one of the few ways to get usage data for an industry or market sector and for overall US and global Web use. It provides comparative data on the usage of your Web site versus competitive Web sites in a defined market. Using the common metrics of pages viewed and unique visitors, this technique is useful for developing time series Web usage trends.
Limitations. Data collection and extrapolation methods vary by company. The differences in panel composition and methodologies mean that usage data from different companies are not strictly comparable. Your organization should consider these results as estimates, not precise measures.
Typical cost. $35,000 to $40,000 per year, for a subscription service that includes online Web access to monthly measurement data.

Web Evaluation Basics

1. **Plan your Web evaluation, then evaluate your plan. An evaluation plan is an important part of a Web development strategy. Update it periodically.**
2. **Evaluation is a necessary part of Web infrastructure development. Good evaluative information can help optimize your investment in Web technology and content. It can also provide a solid basis for making future improvements.**
3. **Web evaluation is an iterative process, not a one-shot deal. Web technology, content, and users change over time; periodic evaluation helps keep your Web site aligned with its customers and your organization's mission.**
4. **Some Web evaluation is better than no evaluation. The marginal utility of initial Web evaluation activities, if properly implemented, is usually high. Important insights typically result, even if time and money do not permit the most comprehensive program.**
5. **Quality counts in Web evaluation. Evaluative data may be difficult to analyze and subject to varying interpretations. Time taken to assure quality control in data collection and analysis will be time well spent.**

WEB AND INTERNET PERFORMANCE DATA

The methods in this cluster focus on the Web site's technical performance and its Internet connectivity. A key question is, how fast can users download Web pages via the Internet? The download speed is a significant determinant of user satisfaction. Download speed depends on several factors—the Web site's design and content, the Web server software and configuration, the local area network (LAN) between the Web server and the Internet connection, the type and speed of the connection to the Internet backbone, the backbone itself, and—at the user's end—the Internet and LAN connectivity and computer platform.

You can use internal and external means to collect performance data. At the application level (HTTP), common metrics include the time in seconds to download the Web site front page and a breakdown of download time by each page element (text, graphics, and so on). At the transport level (TCP), common metrics include bulk transfer capacity (effective available bandwidth), latency or round trip time (for packets to transit from sender to receiver and back), packet loss, and packet routing stability.

IT staff collect technical performance data using common testing software such as Iperf, Ping, and Traceroute. In addition, commercial vendors can use a proxy test network that emulates users downloading Web pages to collect the data. You then monitor the data for operational anomalies and analyze it for trends over time. Organizations can perform this kind of testing on an ad hoc or continuous basis—or for defined periods of time.
Strengths. This approach helps Web managers understand whether technical issues associated with the Web site or its Internet connection are degrading performance and, by extension, user satisfaction. It provides a basis for tracking performance and comparing with benchmarks and other Web sites. Finally, it is useful for troubleshooting technical issues.
Limitations. Because Web and Internet performance vary over time, a full understanding of this type of data requires a longer-term monitoring program and a commitment of management and resources. Performance data may not be fully comparable, unless methods and testing protocols are clearly specified and understood. In addition, monitoring pathways to multiple geographical sites can be expensive.
Typical cost. For in-house testing, most test software is freely available, but the cost of staff time for data collection, analysis, and reporting can be significant. External testing typically costs $10,000 to $15,000 per URL per year. This includes online access to a Web-based data repository with some analytical and drill-down capabilities.

CRAFTING A WEB EVALUATION PLAN

Lesson one is to plan your Web evaluation strategy up front, as part of your overall Web site development plan. An ad hoc approach will likely yield less-than-optimal results. (See the "Web Evaluation Basics" sidebar for this and other lessons learned.)

The planning approach we suggest starts with a checklist of the range of possible Web evaluation methods. Table 1 indicates each method's relative importance for initial Web site development, ongoing Web operations, and further

Some Practical Benefits of Web Evaluation

- **Improving layout, navigation, and search functions (user-friendliness).**
- **Better matching the Web site to its intended audience.**
- **Gauging the relative importance of features and functions.**
- **Assessing the need for new or modified features and functions.**
- **Diagnosing and improving technical performance and download times.**
- **Identifying underserved or underrepresented user groups.**
- **Evaluating the impact of Web site usage on user knowledge and behavior.**
- **Evaluating the impact of outreach and promotional activities on Web site usage.**
- **Tracking Web usage trends in the context of overall trends in the relevant market space.**
- **Comparing Web site usage and impact against agency mission and performance goals.**

Web site improvement. Not all methods apply to all evaluation needs or stages in the Web site life cycle. However, the multidimensional approach helps assure that evaluation results include some redundancy and crosschecking, so that one method's relative strengths compensate for another's limitations. By looking for patterns in the evaluation results, you can have much greater confidence that the results are painting a reasonably accurate picture of how customers rate the Web site and how well the Web site is performing.

Naturally, the number of methods you select will depend in part on the size, complexity, and usage level of the program and Web site you are evaluating, and on your evaluation budget. Table 1 can help you select Web evaluation methods at each stage in order of relative importance, from highest to lowest, as funds permit.

The transition to e-government offers many opportunities but also major challenges. Well-designed and smoothly functioning Web sites can be a strong platform for delivering a wide range of government services electronically. But to ensure this outcome, a robust Web evaluation strategy is a must.

A multidimensional Web evaluation strategy—combining several of the approaches we've described in this article—lets you triangulate multiple perspectives to provide a more complete and accurate overall Web site evaluation. This gives you the best prospect of better understanding your users, developing and improving your Web sites, and gauging the impact of Web site usage. The "Some Practical Benefits of Web Evaluation" sidebar lists other likely positive effects of a coherent evaluation strategy.

Finally, as important as Web evaluation is, organizations must ultimately evaluate Web-based platforms for delivering information and other government services in the context of overall program goals and objectives. In this sense, Web evaluation is an integral part of program evaluation in the age of e-government. ■

Fred B. Wood *is a computer scientist at the National Library of Medicine. Contact him at fred_wood@nlm.nih.gov.*

Elliot R. Siegel *is associate director for health information programs development at the National Library of Medicine. Contact him at siegel@nlm.nih.gov.*

Eve-Marie LaCroix *is chief, Public Services Division, at the National Library of Medicine. Contact her at eve-marie_lacroix@nlm.nih.gov.*

Becky J. Lyon *is deputy associate director, Library Operations Division, at the National Library of Medicine. Contact her at becky_lyon@nlm.nih.gov.*

Dennis A. Benson *is chief, Information Resources Branch, National Center for Biotechnology Information, at the National Library of Medicine. Contact him at dab@ncbi.nlm.nih.gov.*

Victor Cid *is a computer scientist in the Office of Computer and Communications Systems at the National Library of Medicine. Contact him at vcid@nlm.nih.gov.*

Susan Fariss *is a systems librarian in the Public Services Division at the National Library of Medicine. Contact her at susan_fariss@nlm.nih.gov.*

We gratefully acknowledge management leadership on Web evaluation from Donald A.B. Lindberg, director, National Library of Medicine (NLM); technical support from Simon Liu, director, NLM Office of Computer and Communications Systems; and the assistance of Joyce Backus, Gale Dutcher, Larry Kingsland, Cindy Love, Terry Luedtke, Barbara Rapp, and Bryan Zhou of the NLM staff and Tara Piazza of Fulcrum Analytics. The NIH Evaluation Set-Aside Program provided partial support for this work.