Survey Measures of Web-Oriented Digital Literacy[*]

Eszter Hargittai[**]
Northwestern University

Abstract

This paper presents survey measures of Web-oriented digital literacy to serve as proxies for observed skill measures, which are much more expensive and difficult to collect for large samples. Findings are based on a study that examined users' digital literacy through both observations and survey questions making it possible to check the validity of survey proxy measures. These analyses yield a set of recommendations for what measures work well as survey proxies of people's observed Web-use skills. Some of these survey measures were administered on the General Social Survey 2000 and 2002 Internet modules making the findings relevant for the use of existing large-scale national data sets. Results suggest that some composite variables of survey knowledge items are better predictors of people's actual digital literacy based on performance tests than measures of users' self-perceived abilities, a proxy traditionally used in the literature on the topic.

Keywords: methods, surveys, measures, skill, Web use, digital literacy

[*] I thank Paul DiMaggio, Scott Lynch and Peter Miller for helpful discussions. I am also indebted to Ron Anderson and the anonymous referees for their valuable suggestions for improving the manuscript. Generous support from the Markle Foundation and NSF grant #IIS0086143 is kindly acknowledged. The project has also been supported in part by a grant from the Russell Sage Foundation. I am also grateful to the Dan David Foundation for its support.

[**] Contact the author at sscore05@webuse.org.

Introduction

An increasing body of literature exists on how people are incorporating the Internet into their everyday lives (Fallows 2004; Howard and Jones 2003; Wellman and Haythornthwaite 2002) and in particular, how online behavior differs across different segments of the population (DiMaggio, Hargittai, Celeste, and Shafer 2004). As diffusion has spread across the population, a growing number of scholars have been looking at differences in online behavior among Web users in addition to simply exploring differences in access statistics (Mossberger, Tolbert, and Stansbury 2003). Much of such research relies on survey data gathered about people's online activities. Survey measures are helpful because they allow for the collection of data from relatively large sample sizes making possible various quantitative analyses and the potential generalizability of the findings for a larger population group. However, some questions are hard to assess through survey questions. One such area concerns information about people's digital literacy.

An existing line of research has focused on people's computer skills (e.g. Dutton and Anderson 1989; Shashaani 1994) with some emphasis on Internet skills in recent years (Hargittai 2003; Hargittai 2002b; Mossberger, Tolbert, and Stansbury 2003). However, most of the existing literature is based on people's *perception* of their computer skills – often referred to as "self-efficacy" (Bandura 1977) – instead of measuring actual abilities through observations or survey items that measure users' actual knowledge of computer and Internet-related terms and functions.

This paper contributes to the literature on refined measures of Web use and digital literacy studies in particular by presenting possible survey measures of people's online skills *derived from* measures about the actual online skills of users

assessed through performance tests. Findings are based on the online abilities of one hundred randomly selected Internet users in a New Jersey county. Using data on both people's actual Web-use skills and survey questions measuring their knowledge of Internet-related items, we can determine whether identical survey questions administered on the General Social Survey Internet modules in 2000 and 2002 can be used as a proxy for people's online skills. On the basis of these analyses, I recommend the creation of index variables for both GSS 2000 and GSS 2002 as proxies for digital literacy measures. I also make recommendations about which survey questions should be administered on future surveys for optimal composite measures of Web-oriented digital literacy.

As Internet use diffuses to an increasing portion of the population, we need measures beyond simple access statistics for a refined understanding of potential inequalities stemming from differentiated Internet use. A focus on variation in digital literacy allows us to see what segments of the population may be best poised to benefit from the medium. As research has shown, merely having access to an Internet-connected machine does not result in informed users (Hargittai 2003). If some people are unable to find information online while an increasing number of services relevant to daily life become easiest to access on the Web (e.g. financial services, product information, government forms) then the segment of the population with low digital literacy levels will become increasingly disadvantaged in our digital world.

In-depth measures of online skill

I draw on data from a project I conducted in 2001-2002 on people's Web-use skills. I defined skill as a user's ability to locate content on the Web effectively and

efficiently.[1] I gathered data on one hundred randomly selected Web users' online skills using in-person observations and in-depth interviews. Participants performed online tasks in a research setting. All of their online actions were recorded and later analyzed to see whether they could locate various types of content online and how long they took to do so. Hargittai (2002a) describes the methodology in more detail. The in-person observations of people's online browsing behavior resulted in two measures of online skill:

1. Percentage of eight tasks completed successfully *(effectiveness)*
2. Amount of time spent on the eight tasks *(efficiency)*

Subjects looked for information on a) job or career opportunities; b) a site that compares different presidential candidates' views on abortion; c) a used car for purchasing; d) tax forms; e) information about local cultural events (movie time listings, theatre shows); f) music to listen to online; g) children's art; and h) a museum's or gallery's Web site. See Hargittai (2003) for copies of all study instruments.

Survey measures of digital literacy

In addition to looking for various types of content online, participants were also presented with survey questions to measure selected aspects of their Internet-related knowledge. In sum, four different types of measures were collected about digital literacy levels.

1. Four yes/no self-report questions about digital literacy (DL)

[1] Undoubtedly, there are numerous online actions one can consider when measuring Web-use skill. Here, I focused on the aspect of information retrieval instead of person-to-person communication, because many forms of group discussion are also contingent upon the ability to find relevant groups with which to discuss topics.

2. Thirty-eight 5-point (self-reported) ratings of degree of understanding of DL-related items
3. Thirty-seven multiple choice (MC) tests of DL (sub-sample only)
4. An overall (self-report) rating of "Internet" skill

Here I present these measures in detail including the exact wording of the survey questions.

1. Yes/no self-reports of digital literacy (4)

DOWNLOAD - Do you know how to download a file from the World Wide Web to your computer? (17% NO)[2]

UPLOAD - Do you know how to send a file that is on your computer's hard drive to someone using another computer? (26% NO)

OPENATT - Do you know how to open an attachment someone sent you via email? (4% NO)

SRCHENGN1 - Do you know the name of any search engines? (13% NO)

2. Five-point self-reported ratings of DL items (38)

Exact wording of the question for the items below:

How familiar are you with the following Internet-related items? Please choose a number between 1 and 5 where 1 represents having "no understanding" and 5 represents having "a full understanding" of the item. (none, little, some, good, full)

Modem, Browser, Server, ISP, HTML, "bcc" option in email, Flaming, Spam, Spider, Boolean expression, MP3, JPG, XML, Meta-search engine, Natural language, Proximity operators, .gov ("dot gov"), Banner ad, Weblog, Usenet, Message thread, Filtering software, Cookie, DNS parking, Mirror site, P3P, Click-through, Image map, E-zine, Meta-tag, Frames, Shareware, Preference setting, Remote login, Refresh/Reload, Newsgroup, PDF

3. Multiple-choice tests of the same DL items (37) as in #2 administered on random 36 percent of participants[3]

Exact wording of the question for the items listed in #2 above:

Please choose the correct response to all of the following multiple choice questions.

What is …?

[2] Descriptive statistics are presented in parentheses.

[3] One of the 41 items ("advanced search") was accidentally omitted from the multiple-choice section of the survey so it is not possible to run a validity check on that item.

What does … stand for? (used for acronyms)

4. Self-reported rating of Internet skill

Users were asked to answer the following question measured on a five-point scale (not at all skilled, not very skilled, fairly skilled, very skilled, and expert): "In terms of your Internet skills, do you consider yourself to be..." On a scale of 1-5, the mean self-perceived skill level is 2.88 in the sample (st.d.: .73).

Digital literacy measures on the General Social Survey

Some of the self-report questions from this study were replicated on the General Social Survey 2000 and GSS 2002 Internet modules (the items included in GSS 2002 were based on preliminary results from this project). Here I list these items (including some descriptive statistics in parentheses to give an idea of the variance in responses based on the national GSS samples).

GSS 2000

Yes/no self-reports of digital literacy

DOWNLOAD – Do you know how to download a file from the World Wide Web to your computer? (20% NO)

UPLOAD – Do you know how to send a file that is on your computer's hard drive to someone using another computer? (31% NO)

GSS 2002

Three-point self-reported ratings of digital literacy items

Exact wording of the question for the items below:

Are you very familiar, somewhat familiar or not familiar with the following Internet terms:

ADVSRCH – Advanced Search (23% NOT FAMILIAR)

MP3 – MP3 (47% NOT FAMILIAR)

EZINES – E-zines (81% NOT FAMILIAR)

PREFSETS – Preference Settings (27% NOT FAMILIAR)

NEWSGRPS – Newsgroups (40% NOT FAMILIAR)

The descriptive statistics suggest that respondents in the project on which this paper draws tended to be somewhat (a few percentage points) more knowledgeable about Internet-related terms than those in the General Social Survey (in both 2000 and 2002). However, the overall ranking of items is similar in the two studies. Terms that most people knew well in the in-depth study correspond to the items that most people also knew in the GSS and the least familiar items were the same in both samples. These similarities suggest that findings about the survey measures based on the in-depth study sample are generalizable to use of the GSS Internet modules.

The validity of self-reported ratings of DL items

To test the validity of self-reported scores on digital literacy items, a subset of respondents answered multiple-choice questions about thirty-seven of the terms. They were presented with five options out of which one was the correct response. Appendix 1 presents the Pearson's and polychoric correlation coefficients for the 37 self-reported ratings and the multiple-choice question results. I use both coefficients because the Pearson's correlation coefficient tends to underestimate the relationship of variables when used for ordinal-level data (Lynch 1999).[4] The coefficients in the table indicate that there are statistically significant correlations between the majority of the measures. Three variables did not show any variance on the multiple-choice measures making it impossible to calculate meaningful correlations between those measures and self-reported levels of understanding. Nonetheless, for the majority of

[4] I thank Scott Lynch for giving me access to his program to estimate the polychoric correlations.

the variables, the self-reported knowledge measure is a good indication of people's actual knowledge of the terms.

The relationship between behavioral and survey measures of digital literacy

I calculated the Pearson's correlation coefficients between the self-reported ratings and the two items measuring actual ability: a) percentage of tasks successfully completed (*effectiveness)*; and b) amount of time spent on the eight tasks *(efficiency)*. The third and fourth columns in Appendix 1 present the results of these analyses for the entire sample. Items that were replicated on the General Social Survey Internet modules are highlighted in bold.

The signs of the coefficients are in the expected direction. For percentage of tasks successfully completed, the correlation coefficients are positive suggesting that understanding the various computer and Internet-related terms is positively correlated with users' ability to find content online. The negative coefficients for time spent on tasks shows that those with better understanding of computer and Internet terms took less time to look for information online. *In the majority of cases the coefficients are statistically significant for both outcome skill measures.* This suggests that the self-reported ratings of digital literacy items may be used as a proxy for actual skill measures. The next step is calculating the optimal index for measuring digital literacy using survey questions.

Composite measures of digital literacy

Based on the above findings about the relationship of survey and actual measures of online skill, I recommend creating a composite variable for measuring Web-oriented digital literacy using survey questions. Using information from the

coefficients presented in Appendix 1, I created a composite index excluding those knowledge items, which exhibited low correlations with the outcome variables. This new index variable yields a correlation coefficient of 0.573 (p=.000) and –0.540 (p=.000) for successful completion of all tasks and for total time searching, respectively. This new variable is the sum of the self-reported ratings of the following *seven items*: MP3, PREFERENCE SETTING, REFRESH/RELOAD, NEWSGROUP, PDF, ADVANCED SEARCH and DOWNLOAD. This is the best possible index based on the findings from the study. The index has a Cronbach's alpha of .89.

Using the General Social Survey, we are limited to questions asked on the surveys in 2000 and 2002 so I discuss those separately. The composite variables I present here in the case of both of these surveys is based on the self-rated items that exhibited the highest and most statistically significant correlations with actual measures of skill.

For GSS 2000 a composite of the DOWNLOAD *and* UPLOAD *variables* may be used as a proxy for skill. In the user study discussed in this paper, the correlation of this index variable with actual measures of skill is higher than any individual correlation coefficient at 0.5511 (p=.000) for successful completion of all tasks and -0.4475 (p=.000) for total time spent on the eight tasks. The composite variable's Cronbach's alpha for the study sample is .77, for the GSS sample it is .65.

For GSS 2002 an index variable constructed of ADVANCED SEARCH, PREFERENCE SETTING, *and* NEWSGROUP *variables* yields the best proxy for

actual skill.[5] Based on data from the study reported here, the correlation of this constructed index variable with actual measures of skill is higher than any individual correlation coefficient at 0.5579 (p=.000) for successful completion of all tasks and -0.5061 (p=.000) for total time spent on the eight tasks. The Cronbach's alpha for the variables included in this index is .83 for the New Jersey study sample and .76 for the national GSS sample.

Survey measures of digital literacy as predictors of actual online skill

Most existing research on people's computer use skill – a focus more common in the literature than Internet-use skills – draws on information about people's self-perceived skills instead of measures of actual abilities. For comparison purposes, respondents in this study were also asked to rate their self-perceived Internet skill level. As mentioned earlier, the mean of this measure is 2.88 in the study sample (st.d.: .73). Another possible proxy for skill may be the amount of time people spend online. Those who spend more time online have more opportunity to refine their online abilities. A similar argument may be made for number of years one has been an Internet user. Over time people may well pick up skills and improve their digital literacy. I consider the predictive power of these variables on actual skill as well.

To test the power of the proposed composite measures based on people's self-rated understanding of digital literacy items, I compare the predictive power of the more traditional self-perceived skill measure and the Internet time use measures to the predictive power of the new constructs on actual skill. In Table 1, I present

[5] Although the variable EZINE also exhibited statistically significant correlation, the actual level of correlation was considerably lower than the figures for the other three variables making the value of the construct weaker than that of individual variables and thus EZINE is not included in the index.

the results of the various survey measures' predictive power with respect to actual skill. The first row shows the adjusted R^2 for amount of time spent on the Web as a predictor of actual skill, while the second row shows the result for number of years a respondent has been a user. The third row displays the result of the self-perceived skill measure regressed on the actual skill measure. The fourth and fifth rows show the result of the indexes created from the variables available on the GSS 2000 and 2002 Internet modules respectively regressed on the actual skill measure.[6] Finally, the last row shows the predictive power of the seven-item composite variable based on the most highly correlated survey measures of skill. This is the best predictor of skill and thus the recommendation from this study is that the seven items that make up this scale should be included in future surveys as a measure of people's Web-oriented digital literacy.

Table 1. The Predictive Power of the Various Survey Measures of Actual Skill

Survey Measures	**Actual Skill Measure Adjusted R^2**
Time spent on Web weekly	**.048**
Years using the Internet	**.114**
Self-perceived skill	**.239**
GSS 2000 index	**.297**
GSS 2002 index	**.304**
Seven item best index	**.321**

[6] I created a composite variable including all of the digital literacy items available on the GSS 2002 Internet module. The Cronbach's alpha for these variables is .79, which his slightly higher than the alpha for the three variables in the construct. However, when checking the predictive power of this larger index variable, the results suggest that it is not as a good a predictor of the measures resulting from the performance tests as the smaller construct. The adjusted R^2 for the 5-item scale is .28 suggesting that the index variable that only contains the three most highly correlated variables is a better proxy for actual skill than a sum of all available variables.

Conclusion

As the Web evolves, more and more information is available on the network to users. Search and classification services continue to develop and evolve to help users deal with the demands of the increasingly vast amounts of available information and help users find material of interest to them. While these services have certainly made online content more accessible to some, their mere existence does not guarantee that people will be able to navigate efficiently the literally billions of pages that make up the Web (Hargittai 2004; Rieh 2004; Spink, Wolfram, and Jansen 2001). Users differ with respect to their awareness of various search engines and the optimal ways to use them (Hargittai 2003). Today's search engines are still not well-equipped to deal with simple queries that contain no more than one word, yet the majority of queries on search engines do not include more refined information (Spink, Jansen, Wolfram, and Saracevic 2002; Spink, Wolfram, and Jansen 2001). This limits their utility for numerous users and limits the ways in which these users may benefit from the medium.

As the Internet spreads to an increasing portion of the population and as online services start permeating more and more parts of people's daily lives, nuanced measures of Internet use will gain importance for research on the social implications of information technologies. The validity of survey measures is an important challenge for social scientists. In this paper, I contribute to the literature on Internet use and methodology by proposing a survey measure of Web-oriented digital literacy that is based on verifying the validity of the measures derived from their relationship with actual skill measures.

Given that some of these measures were administered on the General Social Survey Internet modules in 2000 and 2002, researchers using those publicly available data sets will be able to incorporate these nuanced measures of Internet use into the analyses of large-scale national data bases.[7] Since the items identified here as important predictors of actual skill are measured on a five-point scale, their inclusion on future surveys should be possible with relatively little effort as compared to refined user studies. Yet they will yield more reliable estimates of people's actual skills than the currently dominant survey measure of self-perceived user skill allows.

[7] Undoubtedly, some knowledge about Internet-related terms will change over time across the population as particular features and services become increasingly well-known by users. Nonetheless, because the study reported in this paper was administered within the same timeframe as the GSS Internet modules, findings from this study can be generalized to use of the GSS.

Appendix 1. Pearson's and polychoric correlation coefficients of self-reported ratings and multiple-choice

Pearson's correlation coefficients and polychoric correlation coefficients for five-point self-reported ratings and multiple-choice measures of digital literacy items; Pearson's correlation coefficients for relationship between survey items and a) successful completion of tasks; and b) total time spent on eight tasks. Items replicated on the GSS are highlighted in bold. Items included in the proposed "best index" are shaded.
* <.05; **<.01; ***.005

Digital Literacy Items	Pearson's correlation coefficient	Polychoric correlation coefficient	Correlation with Successful Completion of Tasks	Correlation with Total Time Spent on 8 Tasks
Download	N/A*	N/A	**0.5272*** **	**-0.4392*** **
Advanced search	N/A	N/A	**0.5110*** **	**-0.4261*** **
Preference setting	**0.5052*** **	**0.982514**	**0.4730*** **	**-0.4215*** **
Newsgroup	**0.5640*** **	**0.871793**	**0.4710*** **	**-0.4680*** **
PDF	**0.6866*** **	**0.855970**	**0.4647*** **	**-0.4186*** **
Refresh/Reload	**0.6912*** **	**0.762428**	**0.4509*** **	**-0.4739*** **
MP3	**0.5582*** **	**0.717611**	**0.4112*** **	**-0.4265*** **
Upload	N/A	N/A	**0.4762*** **	**-0.3771*** **
E-zine	**0.6660*** **	**0.865540**	**0.3323*** **	**-0.3729*** **
Banner ad	**0.7289*** **	**0.895455**	**0.4332*** **	**-0.3342*** **
.gov ("dot gov")	**0.4915*** **	**0.887191**	**0.4310*** **	**-0.3309*** **
HTML	**0.5231*** **	**0.601987**	**0.4227*** **	**-0.4334*** **
Search engine	N/A	N/A	**0.4186*** **	**-0.2392***
JPG	**0.7006*** **	**0.853337**	**0.4105*** **	**-0.4059*** **

* No measure is available because no variance was observed in the responses to the multiple-choice question.

Shareware	**0.6726*** **	**0.868788**	**0.4099*** **	**-0.3053*** **
Browser*	N/A	N/A	**0.4050*** **	**-0.2965*** **
Frames	**0.6661*** **	**0.875697**	**0.4014*** **	**-0.3695*** **
Remote login	**0.5909*** **	**0.770501**	**0.3721*** **	**-0.3718*** **
Spam	**0.5895*** **	**0.844139**	**0.3637*** **	**-0.3511*** **
Boolean expression	**0.6887*** **	**0.865298**	**0.3512*** **	**-0.2058** **
ISP	0.2401	**0.728378**	**0.3455*** **	**-0.3041*** **
"bcc" option in email	**0.7915*** **	**0.879233**	**0.3412*** **	**-0.3681*** **
Cookie	**0.5833*** **	**0.864487**	**0.3197*** **	**-0.3494*** **
Natural language	0.2366	**0.769852**	**0.3024*** **	-0.1625
Mirror site	**0.7398*** **	**0.891701**	**0.2915*** **	**-0.2267***
Flaming	**0.8224*** **	**0.892490**	**0.2772** **	**-0.3220*** **
Message thread	**0.7203*** **	**0.807312**	**0.2766** **	**-0.3034*** **
XML	**0.5704*** **	**0.849796**	**0.2707*** **	**-0.2866*** **
Meta-search engine	**0.5406*** **	**0.888293**	**0.2687*** **	-0.1905
Usenet	**0.5284*** **	**0.604538**	**0.2494***	**-0.2525***
Server	0.1542	0.210150	**0.2453***	-0.1675
Open attachment	N/A	N/A	**0.2381***	-0.1052
Click-through	**0.6198*** **	**0.760315**	**0.2289***	**-0.2128***
Image map	**0.6648*** **	**0.759624**	**0.2247***	**-0.2773** **
Proximity operators	**0.4559** **	**0.586734**	**0.2159***	-0.0277
Meta-tag	**0.7665*** **	**0.911197**	**0.2012***	-0.1867

Weblog	N/A	-	**0.2004***	-0.1449
DNS parking	**0.7590*** **	**0.901940**	0.1858	-0.1636
Modem	N/A	N/A	0.1490	-0.1085
P3P	**0.7235*** **	**0.999986**	0.1447	-0.1692
Filtering software	0.2770	0.486559	0.1379	**-0.3090*** **
Spider	**0.8585*** **	**0.926497**	0.0903	**-0.2157***

References

Bandura, Albert. 1977. "Self-Efficacy: Toward a Unifying Theory of Behavioral Change." *Psychological Review* 84:191-215.

DiMaggio, Paul, Eszter Hargittai, Coral Celeste, and Steven Shafer. 2004. "Digital Inequality: From Unequal Access to Differentiated Use." in *Social Inequality*, edited by K. Neckerman. New York: Russell Sage Foundation.

Dutton, W.H. and Ronald E. Anderson. 1989. "Computers and literacy: differing perspectives in the social sciences." *Social Science Computer Review* 7:1-6.

Fallows, Deborah. 2004. "The Internet and Daily Life." Pew Internet and American Life Project, Washington, D.C.

Hargittai, E. 2003. "How Wide a Web? Inequalities in Accessing Information Online." Ph.D. Thesis, Sociology Department, Princeton University, Princeton, NJ.

Hargittai, Eszter. 2002a. "Beyond logs and surveys: In-depth measures of people's Web use skills." *Journal of the American society for information science and technology perspectives* 53:1239-1244 http://www.eszter.com/research/a09-methods.html.

—. 2002b. "Second-Level Digital Divide: Differences in People's Online Skills." *First Monday* 7 http://www.firstmonday.org/issues/issue7_4/hargittai/index.html.

—. 2004. "Life Beyond Google." in *BBC News*. London.

Howard, P.E.N. and S. Jones. 2003. *Society Online*. Thousand Oaks, Calif: Sage Publications.

Lynch, Scott. 1999. "Bayesian Estimation of Polychoric and Polyserial Correlations via Markov Chain Monte Carlo Simulation Algorithms: A Simulation Study Comparing Posterior Mean and Maximum Likelihood Estimators." Institute of Statistics and Decision Sciences, Duke University, Durham, NC.

Mossberger, Karen, Caroline J. Tolbert, and Mary Stansbury. 2003. *Virtual Inequality: Beyond the Digital Divide*. Washington, D.C.: Georgetown University Press.

Rieh, Soo Young. 2004. "On the Web at Home: Information Seeking and Web Searching in the Home Environment." *Journal of the American Society for Information Science and Technology* 55:743-753.

Shashaani, Lili. 1994. "Gender-Differences in Computer Experience and its Influence on Computer Attitudes." *Journal of Educational Computing Research* 11:347-67.

Spink, A., B.J. Jansen, D. Wolfram, and T. Saracevic. 2002. "From E-Sex to E-Commerce: Web Search Changes." *IEEE Computer* 35:107-109.

Spink, Amanda, Dietmar Wolfram, and Major B. J. Jansen. 2001. "Searching the Web: The Public and Their Queries." *Journal of the American Society for Information Science and Technology* 52:226-234.

Wellman, Barry and Caroline Haythornthwaite. 2002. *The Internet in Everyday Life*. Oxford: Blackwell Publishers.