

**DATA MEMO**

**BY:** PIP Director Lee Rainie (202-419-4500)  
**DATE:** July 2005

*The average American internet user is not sure what podcasting is, what an RSS feed does, or what the term “phishing” means*

Large numbers of internet users do not know the basic definition of some of the hottest new internet innovations and one of the most serious online dangers.

In a nationwide phone survey between May 4 and June 7, the Pew Internet & American Life Project asked internet users if they knew what certain internet terms meant. The results showed that some terms were well known, but that the terms “podcasting” and “RSS feeds” were not familiar to a majority of internet users and that “phishing” is still a foreign term to many:

<b>What online Americans do and don't know about internet-related terms</b>			
<i>The question: Please tell me if have a good idea what the term means, or if you aren't really sure what it means:</i>			
	<i>Have a good idea</i>	<i>Not really sure</i>	<i>Never heard the term</i>
<b>Spam</b> (unsolicited bulk commercial email)	88%	10%	3%
<b>Firewall</b> (a program or device that filters information coming through an internet connection and blocks packets that the user has identified as threatening)	78%	19%	3%
<b>Spyware</b> (programs that are loaded onto a computer without consent and that track users' online activities)	78%	19%	3%
<b>Internet cookies</b> (computer code that is placed on a hard drive when internet users go to Web sites and allow the sites to identify the computer if it returns to the site)	68%	27%	5%
<b>Adware</b> (software that is bundled with free files and programs that is loaded on a computer and can use information about a computer users preferences to provide targeted advertising to them)	52%	39%	9%
<b>Internet “phishing”</b> (unsolicited emails that attempt to acquire an internet user's sensitive information, such as credit card numbers by pretending to be a trustworthy person on business)	29%	55%	15%
<b>Podcasting</b> (audio files that are downloaded from Web sites loaded onto MP3 players such as iPods and played at the convenience of users)	13%	64%	23%
<b>RSS feeds</b> (a file format that allows syndication of Web content, including blog postings, to those who have subscribed to the material)	9%	65%	26%

Source: Pew Internet & American Life Project May 4-June 7, 2005 survey. N=1,336 internet users. Margin of error is ±3%.

The findings are another reminder that new and exciting technology developments that seize the interest of industry officials and journalists such as podcasting and RSS feeds usually take a while to register in the wider public. In addition, it is also clear that public awareness of emerging online threats like those posed by phishing scams takes a while to emerge. A previous survey by the Pew Internet Project showed that 35% of email users had received phishing-type solicitations.<sup>1</sup>

As a rule, online men are more likely than women to be aware of these terms. Those with college degrees are more aware than those with high school diplomas, and heavier internet users are more familiar with the lexicon than lighter users. Older internet users are less likely than younger ones to report a good grasp of these terms.

<b>Different levels of awareness</b>						
<i>The percent in each group of internet users who have a good idea what the term means</i>						
	<i>Men</i>	<i>Women</i>	<i>Age 18-29</i>	<i>Age 65+</i>	<i>Broadband at home</i>	<i>Dial-up at home</i>
Spam	90%	87%	88%	85%	94%	85%
Firewall	83	73	77	59	88	69
Spyware	82	74	82	55	88	75
Internet cookies	74	62	67	52	81	59
Adware	60	44	55	31	65	43
Internet "phishing"	36	23	55	31	35	25
Podcasting	16	11	12	5	16	11
RSS feeds	12	6	12	5	10	9

*Source: Pew Internet & American Life Project May 4-June 7, 2005 survey. N=1,336 internet users. Margin of error is ±3%.*

<sup>1</sup> Report available at: [http://www.pewinternet.org/pdfs/PIP\\_Spam\\_Ap05.pdf](http://www.pewinternet.org/pdfs/PIP_Spam_Ap05.pdf)

**Q18** I'm going to read you some Internet terms you may or may not be familiar with. As I read each one, please tell me if you have a good idea what the term means, or if you aren't really sure what it means.

**Based on internet users [N=1,336]**

	YES, HAVE GOOD IDEA WHAT TERM MEANS	NOT REALLY SURE WHAT TERM MEANS	(VOL) NEVER HEARD TERM	DON'T KNOW/ REFUSED
a Firewall	78	19	3	*
b Internet cookies	68	27	5	0
c Spyware	78	19	3	*
d Adware	52	39	9	0
e Internet phishing, spelled with a P-H at the beginning	29	55	15	*
f Spam	88	10	1	*
g Podcasting	13	64	23	*
h RSS feeds	9	65	26	*

## Methodology

### SUMMARY

The Spyware Survey, sponsored by the Pew Internet & American Life Project, obtained telephone interviews with a nationally representative sample of 2,001 adults living in continental United States telephone households. The survey was conducted by Princeton Survey Research International. Interviews were done in English by Princeton Data Source, LLC from May 4 to June 7, 2005. Statistical results are weighted to correct known demographic discrepancies. The margin of sampling error for the complete set of weighted data is  $\pm 2.3\%$ .

Details on the design, execution and analysis of the survey are discussed below.

### Sample Design

The sample was designed to represent all continental U.S. telephone households. The telephone sample was provided by Survey Sampling International, LLC (SSI) according to PSRAI specifications. The sample was drawn using standard *list-assisted random digit dialing* (RDD) methodology. *Active blocks* of telephone numbers (area code + exchange + two-digit block number) that contained three or more residential directory listings were selected with probabilities in proportion to their share of listed telephone households; after selection two more digits were added randomly to complete the number. This method guarantees coverage of every assigned phone number regardless of whether that number is directory listed, purposely unlisted,

or too new to be listed. After selection, the numbers were compared against business directories and matching numbers purged.

### **Contact Procedures**

Interviews were conducted from May 4 to June 7, 2005. As many as 10 attempts were made to contact every sampled telephone number. Sample was released for interviewing in replicates, which are representative subsamples of the larger sample. Using replicates to control the release of sample ensures that complete call procedures are followed for the entire sample.

Calls were staggered over times of day and days of the week to maximize the chance of making contact with potential respondents. Each household received at least one daytime call in an attempt to find someone at home. In each contacted household, interviewers asked to speak with the youngest adult male currently at home. If no male was available, interviewers asked to speak with either the youngest or oldest female at home based on a random rotation.<sup>2</sup> This systematic respondent selection technique has been shown to produce samples that closely mirror the population in terms of age and gender.

### **WEIGHTING AND ANALYSIS**

Weighting is generally used in survey analysis to compensate for patterns of nonresponse that might bias results. The interviewed sample of all adults was weighted to match national parameters for sex, age, education, race, Hispanic origin and region (U.S. Census definitions). These parameters came from a special analysis of the Census Bureau's 2004 Annual Social and Economic Supplement (ASEC) that included all households in the continental United States that had a telephone.

Weighting was accomplished using Sample Balancing, a special iterative sample weighting program that simultaneously balances the distributions of all variables using a statistical technique called the *Deming Algorithm*. Weights were trimmed to prevent individual interviews from having too much influence on the final results. The use of these weights in statistical analysis ensures that the demographic characteristics of the sample closely approximate the demographic characteristics of the national population. Table 1 compares weighted and unweighted sample distributions to population parameters.

---

<sup>2</sup> This is part of a continuing experiment to see what effect, if any, asking for the youngest instead of the oldest female has on sample demographics.

**Table 1: Sample Demographics**

	Parameter	Unweighted	Weighted
<u>Gender</u>			
	Male	48.1	43.5
	Female	51.9	56.5
<u>Age</u>			
	18-24	12.6	9.6
	25-34	18.0	13.3
	35-44	20.3	17.6
	45-54	19.3	19.7
	55-64	13.4	15.7
	65+	16.4	24.0
<u>Education</u>			
	Less than HS Grad.	15.1	10.3
	HS Grad.	35.8	34.4
	Some College	23.3	25.0
	College Grad.	25.8	30.3
<u>Region</u>			
	Northeast	19.2	18.2
	Midwest	23.0	26.1
	South	36.0	36.9
	West	21.8	18.7
<u>Race/Ethnicity</u>			
	White/not Hispanic	71.7	79.8
	Black/not Hispanic	10.8	9.7
	Hispanic	11.9	5.9
	Other/not Hispanic	5.6	4.7

**Effects of Sample Design on Statistical Inference**

Post-data collection statistical adjustments require analysis procedures that reflect departures from simple random sampling. PSRAI calculates the effects of these design features so that an appropriate adjustment can be incorporated into tests of statistical significance when using these data. The so-called "design effect" or *deff* represents the loss in statistical efficiency that results from systematic non-response. The total sample design effect for this survey is 1.12.

PSRAI calculates the composite design effect for a sample of size  $n$ , with each case having a weight,  $w_i$  as:

$$deff = \frac{n \sum_{i=1}^n w_i^2}{\left( \sum_{i=1}^n w_i \right)^2} \quad \text{formula 1}$$

In a wide range of situations, the adjusted *standard error* of a statistic should be calculated by multiplying the usual formula by the square root of the design effect ( $\sqrt{deff}$ ). Thus, the formula for computing the 95% confidence interval around a percentage is:

$$\hat{p} \pm \left( \sqrt{deff} \times 1.96 \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} \right) \quad \text{formula 2}$$

where  $\hat{p}$  is the sample estimate and  $n$  is the unweighted number of sample cases in the group being considered.

The survey's *margin of error* is the largest 95% confidence interval for any estimated proportion based on the total sample—the one around 50%. For example, the margin of error for the entire sample is  $\pm 2.3\%$ . This means that in 95 out every 100 samples drawn using the same methodology, estimated proportions based on the entire sample will be no more than 2.3 percentage points away from their true values in the population. It is important to remember that sampling fluctuations are only one possible source of error in a survey estimate. Other sources, such as respondent selection bias, questionnaire wording and reporting inaccuracy, may contribute additional error of greater or lesser magnitude.

## RESPONSE RATE

Table 2 reports the disposition of all sampled telephone numbers ever dialed from the original telephone number sample. The response rate estimates the fraction of all eligible respondents in the sample that were ultimately interviewed. At PSRAI it is calculated by taking the product of three component rates:<sup>3</sup>

- Contact rate – the proportion of working numbers where a request for interview was made – of 80 percent<sup>4</sup>
- Cooperation rate – the proportion of contacted numbers where a consent for interview was at least initially obtained, versus those refused – of 48 percent
- Completion rate – the proportion of initially cooperating and eligible interviews that were completed – of 91 percent

Thus the response rate for this survey was 35 percent.

---

<sup>3</sup> PSRAI's disposition codes and reporting are consistent with the American Association for Public Opinion Research standards.

<sup>4</sup> PSRAI assumes that 75 percent of cases that result in a constant disposition of "No answer" or "Busy" over 10 or more attempts are actually not working numbers.

**Table 1: Sample Disposition**

	<u>Final</u>
Total Numbers dialed	11,514
Business	1,083
Computer/Fax	884
Other Not-Working	1,945
Additional projected NW	803
<hr/>	
Working numbers	6,799
<b>Working Rate</b>	<b>59.0%</b>
No Answer	109
Busy	44
Answering Machine	866
Callbacks	87
Other Non-Contacts	258
<hr/>	
Contacted numbers	5,434
<b>Contact Rate</b>	<b>79.9%</b>
Initial Refusals	2,070
Second Refusals	766
<hr/>	
Cooperating numbers	2,598
<b>Cooperation Rate</b>	<b>47.8%</b>
No Adult in HH	20
Language Barrier	369
<hr/>	
Eligible numbers	2,209
<b>Eligibility Rate</b>	<b>85.0%</b>
Interrupted	208
<hr/>	
Completes	2,001
<b>Completion Rate</b>	<b>90.6%</b>
<b>Response Rate</b>	<b>34.6%</b>