Behavior Research Methods
2010, 42 (1), 273-285
doi:10.3758/BRM.42.1.273

A comparison of stimulus ratings made online and in person: Gender and method effects

Diana A. Barenboym, Lee H. Wurm, and Annmarie Cano
Wayne State University, Detroit, Michigan

In Experiment 1, separate samples rated nouns on danger, using either an online survey or the same survey in person. In Experiment 2, a single sample rated words on familiarity, using both methods. Women's in-person and online ratings correlated significantly better than men's. In-person ratings correlated significantly better with existing norms in 4 of 8 instances. There were significant effects of condition on mean ratings and completion times. Ratings from participants who withdrew from the experiment correlated significantly less well with existing norms than did ratings from those who completed the whole experiment, in 12 of 16 instances. Analysis of existing data showed that a different statistical conclusion is reached depending on whether in-person or online ratings are used. Furthermore, the categorization of 17.9% (Experiment 1) and 5.3% (Experiment 2) of the items as *high* or *low* depends on which ratings are used. Ratings gathered in person and online cannot be freely substituted.

The Internet has had a profound effect on how psychological science is conducted, and recent years have seen an explosion in the use of the Internet for data collection (e.g., Birnbaum, 2004; Skitka & Sargis, 2006). Research done online is faster, easier, and less expensive to conduct. It also allows for samples that are more diverse in terms of ethnicity, geographical location, education level, age, and socioeconomic status (e.g., Birnbaum, 2004; Gosling, Vazire, Srivastava, & John, 2004).

Internet administration of surveys is not without problems (e.g., Birnbaum, 2004; Buchanan, 2007), but growing numbers of researchers are converting to online methods, with the assumption that the results they obtain will be interchangeable with those they would have obtained in the lab. However, this is an underexamined assumption that could lead to problems (Cronk & West, 2002; Hewson & Charlton, 2005; see also Skitka & Sargis, 2006).

There is a growing literature contrasting computerized/Internet and paper-and-pencil survey administration (for reviews, see, e.g., Buchanan, 2007; Richman, Kiesler, Weisband, & Drasgow, 1999). This literature is only tangentially related to our present purpose, which is to examine the comparability of data collected over the Internet and those collected in the traditional way in cognitive psychology, because for several decades this traditional way of collecting data has been computerized but laboratory based. The issue, therefore, is one of testing environment: Independently of the computer program used to collect the ratings, are different ratings made when a participant must come to the lab to make them? Some of the reasons this could be expected to make a difference will be discussed below.

Testing Environment

In a small number of studies, the testing environment has been manipulated independently of the form of the survey. Cronk and West (2002) assigned a morality questionnaire to be completed via the Internet, either in class or at home, and with paper and pencil, again either in class or at home. Neither factor had a significant effect on the mean ratings, although variance and dropout were significantly higher when the Internet task was performed at home (as compared with all the other conditions).

Chuah, Drasgow, and Roberts (2006) randomly assigned participants to one of three conditions: paper-and-pencil (in groups of 50 or more), Internet (in a computer lab in groups of 10–18), or Internet (alone, whenever and wherever the participant wanted to complete the experiment). They analyzed ratings on 53 items from personality inventories, using item response theory. Significant *differential item functioning* was found for 19 item comparisons, but these were described as likely being due to the large number of comparisons made. Analysis of mean differences showed no significant effects when Bonferroni corrections were applied. The authors concluded that testing condition had no effect on the ratings.

Dandurand, Shultz, and Onishi (2008) contrasted two computerized versions of a problem-solving experiment. The Internet version was modeled on a lab-based study reported in Dandurand, Bowen, and Shultz (2004); some minor modifications were intended to make the task seem more "attractive" and reduce dropout. The task itself was a complex problem-solving situation in which the participant had to figure out how to determine the lightest and

L. H. Wurm, lee.wurm@wayne.edu

heaviest of a set of 12 objects by using a scale a limited number of times.

There were three conditions: Some participants were told whether their answers were correct (reinforcement); some watched demonstrations of solutions (imitation); and some read instructions on how to solve problems of this type (explicit). In the 2004 study, the participants in the reinforcement condition performed significantly worse than those in the other conditions. The 2008 study replicated this, although there was also a significant main effect of testing environment: Internet participants were less accurate than those in the lab (56% vs. 66%). They were also more likely to drop out of the study, despite the efforts to make the task seem more attractive. Thus, there are mixed findings regarding the impact of the testing environment on task performance.

Possible Nonequivalence of Samples

If differences are observed between results collected via the Internet and those collected in person, they might be due to nonequivalence of samples: Most comparisons of testing methods have either accidentally or by design used noncomparable samples of participants, and Internet surveys are not likely to produce representative samples (Epstein & Klinkenberg, 2001).

Some researchers have attempted to remedy the problem. Joinson (1999), for example, randomly assigned participants to an Internet or pen-and-paper condition (crossed with a two-level anonymity factor). They then filled out several personality measures. On social anxiety and social desirability, there were significant differences between the mean scores for the paper-and-pencil version and the Internet version. On self-esteem, the difference was marginally significant ($p = .07$). We will return to the anonymity manipulation below.

Denscombe (2006) assigned groups of 15-year-old students from the same school to complete either a Web-based or a paper-and-pencil version of a survey of health-related behaviors. There was an unambiguously statistically different pattern of responding on only 1 of 23 questions. However, the design was unbalanced (81% of the participants were assigned to the paper-and-pencil condition), and there may have been issues of statistical power as well.

Three studies have used precisely the same participants in multiple administration conditions. Silverstein et al. (2007) compared an Internet-based neurocognitive test battery with an established computerized (but non-Internet) battery, using the same 50 participants in both conditions. The results suggested that the new battery was perfectly usable: "Results indicated comparability across the two batteries" (p. 940).

Whitaker and McKinney (2007) had participants complete Internet and paper-and-pencil versions of a job satisfaction instrument. They found that job satisfaction ratings were quite comparable across methods (demonstrating *measurement invariance*), but they also found that a correlation between age and job satisfaction held only for the paper-and-pencil data.

Norris, Pauli, and Bray (2007) also made use of a single sample of participants in two measurement conditions. They examined negative affect as a function of state anxiety. They found a significant mean difference in computerized (not via Internet, but in a laboratory room) and paper-and-pencil measures of state anxiety. This difference was observed after (but not before) assignment of a grade for coursework.

The results of these five studies are mixed, with three showing measurable meaningful effects of testing condition even with equivalent samples. Whitaker and McKinney (2007) and Norris et al. (2007) demonstrated that even after completely removing what is often the largest single source of variance in psychological studies (i.e., individual differences), a *method* effect can be observed. However, in none of these studies was the form of the instrument held constant across testing conditions. This complicates their interpretation.

Although not the primary focus of the present study, we explored the issue of nonequivalence of samples in two experiments. One used different samples for two rating conditions. The other used a single sample. In the latter experiment, potential nonequivalence became relevant because of dropout.

What Is Being Rated?

In nearly every study that we are aware of, researchers have investigated characteristics of the raters themselves (e.g., personality characteristics, job satisfaction, health behaviors, anxiety, memory). Although this is not a shortcoming of the existing research, it does leave a gap in the knowledge base in an area that is of central importance in many areas of psychology. In many areas of cognitive psychology, for example, it is common practice to ask participants to rate various characteristics of *stimuli* that are to be used in subsequent experiments (objects, events, words, and so on). It is possible that differential administration is more relevant for ratings of items, because participants may implicitly believe that there are correct and incorrect answers. Even if there is no a priori reason to expect that participants will perform differently in in-person versus online studies, one cannot rule out a possible method effect on stimulus ratings, because the existing findings on self-ratings are quite mixed.

We are aware of only two studies in which ratings of stimuli have been compared across different conditions. Balota, Pilotti, and Cortese (2001) had three samples of participants rate the subjective familiarity of words: undergraduates from the university community, healthy older adults, and an Internet sample that ranged in age from 14 to 84 years. Each word was rated by either 30 or 32 participants. Ratings for the undergraduates and healthy older adults were administered with paper and pencil, and undergraduates were tested in groups of 7–40. All of the correlations between these samples' ratings were high (all three rs $> .91$), but the Internet sample gave significantly higher ratings than did either of the other two (both ps $< .004$). The ratings from the undergraduate sample have become a standard reference in psycholinguistics, and they were used for comparison in Experiment 2 of the present study.

Lahl, Göritz, Pietrowsky, and Rosenberg (2009) gathered ratings of German nouns via the Internet. Each noun was rated by approximately 24 participants on one of three psycholinguistic variables: concreteness, valence, or arousal. They did not collect any in-person ratings themselves but compared their results with two existing sets of norms. The authors concluded that there was "good agreement" between their obtained ratings and these norms. However, the correlations between the obtained ratings and those from these existing databases ranged from .58 to .93, averaging .84. Unexplained variance thus ranged from 14% to 66%, averaging nearly 30%.

Correlations this low suggest that if one were using a factorial design and classifying items into groups (e.g., high and low arousal), one would end up with different sets of items depending on whether Internet ratings or existing norms were being used. In the present article, we will demonstrate that this happens. More and more commonly, though, this factorial approach is being abandoned in favor of using the mean item ratings as a continuous predictor variable in regression analyses of data from subsequent experiments that use the stimuli. Differential rating performance in different settings is a potential problem for this approach, too; the correlation coefficients from Lahl et al. (2009) indicate substantial variation in the estimation of the item means. We will provide evidence below that this can lead to differing statistical conclusions in regression analyses: We would reject the null hypothesis using one set of ratings and fail to reject it using the other set of ratings, even though the two sets of ratings were very highly correlated.

One other study that may fit into this category was conducted by Krantz, Ballard, and Scher (1997), who compared ratings of female attractiveness made via the Internet with those gathered in the lab. Although the study was not described as such, we can imagine a situation in which the images shown to participants would be potential stimuli for subsequent studies. The authors found no significant effects of testing condition on mean ratings and obtained correlations above .9 between the Internet ratings and the in-person ratings.

The results of the small number of studies reviewed in this section are thus also mixed. Complicating matters, in all of the studies, the form of the rating instrument varied along with the testing location. One additional factor that may have influenced the results is that, in each of the studies, the Internet participants (probably) completed the task alone, whereas the in-person participants were tested in groups (except for Balota et al.'s [2001] sample of healthy older adults, who were tested at home). We will have more to say about this issue below.

The present study adds to the literature in which stimulus ratings gathered in different settings have been compared. We collected ratings made in a typical laboratory setting, in which a single participant at a time rated a number of stimuli on some dimension. We also collected ratings using an identical rating program, but under what we assume to be typical Internet procedures. We will refer to the latter condition as *online*. The laboratory testing condition will be called *in person*, even though it, too, was technically done online; the identical rating program was used, but participants had to come to the lab to participate (see the Method section of Experiment 1).

In Experiment 1, we used two separate samples that were not restricted on gender. In Experiment 2, we used one approximately gender-balanced sample for both parts of the experiment.

Analytic Strategy

There are different things that might be meant when researchers speak of results being comparable across rating methodologies or locations. The underlying factor structure of the data or other psychometric indices may be relevant if researchers are validating constructs or developing instruments (e.g., Buchanan et al., 2005; Hewson & Charlton, 2005; Meyerson & Tryon, 2003). These were not important for our purposes.

The mean difference in the ratings may be assessed as a way of concluding whether or not the rating method has an effect. Pearson correlation coefficients inform researchers about the extent to which items have the same standard score on two scales. This information is logically independent of the mean rating, and in fact, potential stimuli can switch their relative positions (and thus perhaps even end up in different "groups" if one were doing such categorizing) without having any effect on the means (see, e.g., Buchanan, 2007; Meyerson & Tryon, 2003). In the present study, we used assessments of scale means and correlations to compare the results from different rating conditions. We also compared our ratings with existing norms in Experiment 2, assessing both mean differences and correlations where appropriate.

We included participant gender in our statistical models, since differences in the performance of subgroups of participants have not been explored extensively (Whitaker & McKinney, 2007). There is some evidence of differential gender effects in how participants make ratings such as these (e.g., Hewson & Charlton, 2005; Ullman et al., 2002) and also in how they respond to them (e.g., Wurm, Whitman, Seaman, Hill, & Ulstad, 2007).

Differences in motivation are often cited as a possible reason for differences between online and offline performance (Buchanan, 2007). We did not include any direct measures of motivation, but we did have available what we believed to be a proxy in Experiment 2: completion times. Completion times are not often available in the published literature, but if they are found to depend on the rating condition, it could be of interest to researchers.

Finally, we analyzed data from any participants who failed to complete any of the sessions for possible differences from the data of people who finished the experiment. Dropout is a problem in many research projects, and although we did not expect it to be a major problem in these simple rating tasks, any patterns that began to emerge could inform future research efforts aimed at examining participant motivation, nonequivalence of samples, and so on.

We analyzed the ratings (and completion times when available) with a multilevel linear mixed-effects ANCOVA, with participant and item as crossed random effects (Baayen, Davidson, & Bates, 2008). It simultaneously included participants and items as random effects, replacing the separate by-subjects and by-items analyses usually seen in psycholinguistic research. It is also more powerful than the traditional approach of collapsing observations across participants or across items. Unless otherwise noted, all of the analyses in this study were of this type, in which there were multiple observations per participant.

There is a debate about the appropriate *df* for these analyses. The *p* value typically produced by software packages is based on the upper bound of *df*, which is equal to the number of participants times the number of items, minus the number of parameters in the model. This *p* value has been shown to be somewhat anticonservative, and an alternate method for its calculation has been developed (Baayen et al., 2008). In the present study, we report the traditional *p* value, but all of our significant effects reached significance by both methods.

In comparing correlation coefficients, we usually computed a *z* score, using the method outlined by Meng, Rosenthal, and Rubin (1992) for comparing correlated correlations. This method was not appropriate for all of our comparisons; those where we did not use it will be marked as Fisher *r*-to-*z* transformations.

EXPERIMENTS 1A AND 1B
Danger Ratings

We have conducted several studies in which we correlated participants' mean ratings of the subjective danger and usefulness of word referents with subsequent performance on cognitive tasks such as naming and lexical decision (e.g., Wurm, 2007; Wurm & Seaman, 2008; Wurm & Vakoch, 2000). The danger ratings have always been collected via a computerized, laboratory-based rating methodology. To determine whether online ratings would be comparable to those collected by our typical method, we ran two concurrent experiments. They were identical except that ratings were gathered online in Experiment 1A and in the lab in Experiment 1B.

Method

Participants. Fifty-five students (46 women) from the psychology participant pool at Wayne State University participated. All were native speakers of English. Extra credit in a psychology course was offered in exchange for participation. Twenty-five (22 women) participated online, and 30 (24 women) in person. None dropped out before completing the entire experiment. Gender was independent of condition [$\chi^2(1) = 0.54, p = .46$].

Materials. Stimuli were 117 first constituents of noun–noun compound words used as part of another study (e.g., the *cow* in *cowboy*). These were all fairly common nouns ranging in length from three to eight letters.

Procedure. Two separate studies were posted on a department Web page, the only method by which studies may be advertised in our department. The ratings were made within the same online study environment whether the participants made the ratings online or in person. The rating program was, in fact, identical. For the online version, the participants logged onto the experiment Web site, read the consent form and instructions, and completed the experiment. They were free to do this whenever and wherever they wanted to. The in-person session was identical, except that the participants had to come to the lab to participate. They called from a waiting room and were met and escorted to the lab by the first author. They then were seated at a computer and accessed the rating program, with the experimenter typing in the invitation code. After making sure that the program was running, the experimenter exited the testing room and closed the door.

In both conditions, after reading the on-screen consent form, the participants were shown the following text:

> In this rating questionnaire, you will rate how DANGEROUS FOR HUMAN SURVIVAL several words are. Numbered choices will appear on-screen for you to use in making your rating. Take as much time as you need and just use your own judgment. There are no right or wrong answers. If you do not recognize a word, press the 0 key.

Words appeared 25 to a screen, with each word having the rating scale appear just below it. The 8-point scale had end points labeled *Not at all dangerous for human survival* (1) and *Extremely dangerous for human survival* (8). Trials on which a rating of zero was given were not analyzed.

The participants clicked a button to indicate that they were finished with the current screen. They were then shown the next 25 words (until the final group, which had only 17). Location of words in this list of 117 items was randomized for each participant. Credit was granted by the software as soon as the student closed the survey upon completion and went toward a course that they indicated at the time of signing up.

Results and Discussion

We calculated the mean rating for each item and condition. These means are available from the authors upon request. Sixty-five of the 117 items had a mean difference of at least 0.5 point in one direction or the other. The range of mean differences was 2.5 (ranging from 1.0 in one direction to 1.5 in the other, and representing 31% of the 8-point scale). Thus, a particular item's mean rating might differ *by nearly a third of the rating scale*, depending on the method used to obtain the ratings.

This *range of mean differences* measure is intended to give some indication of the risk of misclassification if one were using ratings to create categories of items, as is often done in psycholinguistic studies. In the present study, we did not intend to classify or categorize items at all, because we were primarily interested in testing hypotheses related to method differences. However, by way of illustration, we used a median split to group the 117 items into *high-* or *low-danger* categories, as is common. Twenty-one of the items (17.9%) had their classification change from high to low danger (or vice versa), depending on whether the in-person or online ratings were used to make the classification. Thus, one's choice of method for gathering ratings can produce differing lists of items. As will be seen below, one's choice of method can also determine whether a researcher concludes that he or she has a significant effect in regression analyses.

The mean ratings (logged because they were positively skewed) did correlate highly, as would be expected [$r(115) = .63$ for men; $r(115) = .89$ for women; both $ps < .001$]. A Fisher *r*-to-*z* transformation showed that the correlation for women was significantly different from that for men ($z = 5.14, p < .001$).

We analyzed the log danger ratings with a multilevel linear mixed-effects ANCOVA, with participant and item as crossed random effects (see Table 1). Ratings given on-

Table 1
Summary of the Multilevel ANCOVA for Variables Predicting Log Danger Ratings

Variable	Regression Coefficient (B)	Standard Error of B	t
Main Effects			
Gender (M)	0.12	0.12	1.08
Condition (OL)	0.19	0.09	2.24*
Interaction			
Gender × condition	0.05	0.24	0.29

Note—The label in parentheses shows which level of the two-level factor is to have the coefficient added to the ratings estimate. The default level of these factors (F, IP) is determined alphabetically by the statistical software. "IP" stands for in person, and "OL" stands for online. M, male; F, female. *$p < .05$.

line were significantly higher than those given in person (M = 2.27 vs. 1.86 on the 8-point scale; SEM = 0.36 and 0.30, respectively). Although small in absolute terms, this is a difference of 22%. Participant gender was quite imbalanced (84% women), so we reran the analysis without it in the model. The main effect of condition remained significant ($B = 0.18$, $SE\ B = 0.09$; $t = 2.14$, $p < .05$).

As was mentioned in the introduction, differing item means could also affect the outcomes of regression analyses. As one example of this, we took the existing data from Fischer (2007), a study that gathered data on a number of variables of psycholinguistic interest for 85 of the 117 items in the present experiment. We examined the question of whether the morphological family size (the number of derived words and compound words that contain a particular morpheme; see Schreuder & Baayen, 1997) of the first constituent of a compound word (e.g., the *cow* in *cowboy*) can be predicted by the rated subjective danger of that constituent. The answer depends on which ratings are used (for in-person ratings, $B = 0.15$, $SE\ B = 0.10$, $t = 1.49$, $p = .14$; for online ratings, $B = 0.23$, $SE\ B = 0.11$, $t = 2.03$, $p < .05$). It is important to note that the correlation between the in-person and online ratings for these 85 items is very high ($r > .93$). Nevertheless, this demonstrates very clearly that the statistical conclusion one reaches can depend on whether one uses ratings gathered online or in person.

We thus have clear evidence that participant ratings of word characteristics can differ substantially as a function of the setting in which they are given. We also have evidence that the correlation between online and in-person ratings depends significantly on rater gender. The sample of men, though, was small ($n = 9$). We have also seen that the classification of a word as high or low on danger (using a median split) depends on which ratings are used, in nearly one out of every five instances. Finally, we have demonstrated that the decision about whether a particular effect is significant in a regression analysis can also depend on which ratings are used.

Perhaps the biggest shortcoming of Experiment 1 was the use of two separate participant samples in the two administration modes, which leaves us unable to conclude that testing condition per se has an effect on ratings (for an extreme example of why this issue is so crucial, see Birnbaum, 1999).

EXPERIMENT 2
Familiarity Ratings

In addition to using a single sample of participants, in Experiment 2, we gathered survey completion times and more evenly balanced participant gender. We also chose a different variable for rating (familiarity). We had no reason to expect that this change would have any implications for the comparison of data between conditions. Neither of the rating tasks had a great deal of personal relevance for the participants, so we did not expect a major influence of variables such as social desirability or online disinhibition. Any differences we observed would likely require a different explanation. We chose familiarity because it is an intuitive and easy judgment for participants to make. In addition, because familiarity is one of the most extensively studied variables in psycholinguistics, there were data available for comparison.

Our analyses proceeded along the same lines as in Experiment 1. We computed correlation coefficients for the ratings themselves and made comparisons of mean ratings. We could also now compare mean completion times. From this, we might expect to learn something about the conscientiousness of the participants, or how much care and effort they put into the task in different conditions. Other statistical tests were possible here that were not possible in Experiment 1, including the relation of our data to established databases (in terms of correlations and means) and an analysis of data from participants who did not complete both parts of the experiment (again in terms of correlations, mean ratings, and mean completion times).

Method

Participants. We were contacted by 240 students from the psychology participant pool at Wayne State University (137 women and 103 men) about enrollment in the study, 139 of whom provided at least partial data. All were native speakers of English. Extra credit in a psychology course was offered in exchange for participation.

Materials. We randomly selected 300 words from among those that are common to a number of widely used databases containing familiarity norms: the MRC Psycholinguistic Database (Wilson, 1988), the Bristol norms (Stadthagen-Gonzalez & Davis, 2006), the Hoosier norms (Nusbaum, Pisoni, & Davis, 1984), and the Balota norms (Balota et al., 2001).

Procedure. The experiment was posted on a department Web page. The information posted made it clear that the participants had to complete two sessions, one in person and one online, in order to receive credit. Students contacted the researcher directly and were randomly assigned to complete either the online or the in-person phase first. If a person was assigned to the online phase first, they were e-mailed a password by which to log in and complete Session 1. If a person was assigned to the in-person phase first, they were given a lab appointment for Session 1.

The procedural specifics for the online and in-person conditions were identical to those in Experiment 1, except for the change in what was being rated. All the ratings were made within the same online program. Regardless of condition, after reading the on-screen consent form, the participants saw the following text:

> Words differ in how commonly or frequently they are encountered. Some words are encountered very frequently, whereas other words are encountered infrequently. The purpose of this study is to have you rate a list of words with respect to this dimension. We believe that your ratings will be important to future

> studies involving word recognition. The rating scale you should use will be displayed on-screen at all times.

Words appeared 25 to a screen, with each word having the rating scale appear just below it. The rating scale used was from Balota et al. (2001). The participants were asked to rate "how frequently you encounter" each word on a scale from 1 to 7 (1 = *never*, 2 = *once a year*, 3 = *once a month*, 4 = *once a week*, 5 = *once every 2 days*, 6 = *once a day*, and 7 = *several times a day*).

The participants clicked a button to indicate that they were finished with the current screen. They were then shown the next 25 words. Location of the words in this list of 300 items was randomized for each participant. Credit was granted by the software as soon as the student closed the survey upon completion and went toward a course that they indicated at the time of signing up.

The software stored not only the rating given, but also the time (in minutes) taken to complete the task and participant gender. Timing began when the participants saw the consent form and ended when they clicked to submit their final page of ratings.

One week from the date of completion for Session 1, the participants received an e-mail telling them how to perform Session 2. The participants who had completed Session 1 in the lab were e-mailed a password to complete Session 2 wherever and whenever they wanted to. The participants who had completed Session 1 online made an appointment to come to the lab for Session 2.

Results and Discussion

We obtained complete data for 113 participants (58% women) and incomplete data from another 26 participants. Table 2 presents a summary of the participant contacts. A χ^2 test of independence showed that the distribution for women differed from that for men [$\chi^2(2) = 10.91, p < .01$]. Women were much more likely to agree to participate but to fail to complete Session 2 (.16 vs. .04), whereas men were more likely to decline to participate at all (.50 vs. .36).

We will begin with data from the participants who completed both sessions, before presenting analyses of some of the incomplete data. We calculated the mean rating for each item and condition. These means are available from the authors upon request. The range of mean differences across items was 0.9 (ranging from 0.4 in one direction to 0.5 in the other direction and representing 13% of the 7-point scale). Although this was smaller than the analogous effect in Experiment 1, it nevertheless resulted in 16 items (5.3%) having their categorization as high or low on familiarity (again using a median split) depend on which ratings were used to make the categorization. This was a smaller percentage than the 17.9% observed in Experiment 1 but is still rather impressive when we consider that *exactly the same people* provided the two ratings in the present experiment.

As was expected, mean ratings correlated extremely highly [$r(298) = .98$ for men; $r(298) = .99$ for women; both ps $< .001$]. Although both of these values were extremely high, a Fisher r-to-z transformation showed that they were significantly different from each other ($z = 2.64$, $p < .01$). In Experiment 1, too, we saw that the women's correlation was stronger than the men's.

Table 2
Number of Participants Contacting the Authors About the Study and Their Relative Frequency

	Women		Men	
Category	No.	Freq.	No.	Freq.
Completed both phases	66	.48	47	.46
Completed only one phase	22	.16	4	.04
Inquired but declined to participate	49	.36	52	.50
Total contacts	137	1.00	103	1.00

Table 3
Summary of the Multilevel ANCOVA for Variables Predicting Familiarity Ratings

Variable	Regression Coefficient (B)	Standard Error of B	t
Main Effects			
Gender (M)	0.45	0.16	2.86^{**}
Session (Two)	0.05	0.01	5.85^{***}
Condition (OL)	−0.04	0.01	-4.77^{***}
Interactions			
Gender × session	0.01	0.02	0.38
Gender × condition	0.04	0.02	2.57^{*}
Session × condition	−0.13	0.31	−0.43
Three-way	−0.07	0.64	−0.13

Note—The label in parentheses shows which level of the two-level factor is to have the coefficient added to the ratings estimate. The default level of these factors (F, One, IP) is determined alphabetically by the statistical software. "IP" stands for in person and "OL" stands for online. M, male; F, female. $^{*}p < .05$. $^{**}p < .01$. $^{***}p < .001$.

A multilevel linear mixed-effects ANCOVA with participant and item as crossed random effects showed that all the predictors had significant main effects on the familiarity ratings (Table 3). The gender coefficient was substantially larger than either of the other main effects. This underscores the importance of assessing gender effects, which is typically not done in research of this type. Men's and women's ratings correlated very highly [$r(298) = .96$, $p < .001$], but the means differed significantly: Men gave significantly higher ratings than did women (4.24 vs. 3.79; *SEM*s = 0.17 and 0.14, respectively). Familiarity ratings were significantly higher in Session 2 than in Session 1 (4.00 vs. 3.95; *SEM*s = 0.15 and 0.16, respectively).

As in Experiment 1, condition (online vs. in-person) had a significant main effect. However, whereas in Experiment 1 the in-person danger ratings were significantly lower than those given online, in Experiment 2 familiarity ratings given in person were significantly higher than those given online (3.99 vs. 3.95; *SEM*s = 0.16 and 0.15, respectively). The apparent discrepancy could be due to differences in the dimension rated (danger vs. familiarity), but analyses presented below suggest a different explanation.

Condition also interacted with gender: The difference between in-person and online mean ratings was more than twice as large for women as for men. Note that this interaction is over and above the main effects, including the substantial main effect of gender. Post hoc analyses showed that the condition effect was significant for women ($B = -0.06$, $SE\ B = 0.01$; $t = -5.25$, $p < .001$), but not for men ($B = -0.01$, $SE\ B = 0.01$; $t = -1.11$, $p = .27$).

The online condition produced both the shortest and the longest completion times. For online ratings, the range of times was 7–100 min (1.95–4.61 in the log units that were analyzed), whereas for in-person ratings, the range of times was 11–54 min (2.40–3.99 in log units). We analyzed these completion times, after logging them to achieve normality. This analysis was similar to the one

Table 4
Summary of the Multilevel ANCOVA for Variables Predicting Log Completion Times

Variable	Regression Coefficient (*B*)	Standard Error of *B*	*t*
Main Effects			
Gender (M)	−0.10	0.07	−1.48
Session (Two)	−0.12	0.04	−2.95**
Condition (OL)	0.11	0.04	2.59*
Interactions			
Gender × session	−0.05	0.08	−0.56
Gender × condition	−0.17	0.08	−2.04*
Session × condition	0.00	0.13	0.03
Three-way	0.20	0.27	0.76

Note—The label in parentheses shows which level of the two-level factor is to have the coefficient added to the time estimate. The default level of these factors (F, One, IP) is determined alphabetically by the statistical software. Completion times were log transformed prior to analysis. "IP" stands for in person and "OL" stands for online. M, male; F, female. $^{*}p < .05$. $^{**}p < .01$.

for the ratings, except that it was not possible to include item as a random effect. Table 4 shows the results of this analysis.

The mean log completion time was significantly shorter for the in-person condition and for Session 2. Condition also interacted significantly with gender. As with the ratings, post hoc analyses showed a significant condition effect for women ($B = 0.18$, $SE\ B = 0.05$; $t = 3.42$, $p < .001$), but not for men ($B = 0.01$, $SE\ B = 0.07$; $t = 0.10$, $p = .92$).

The results of Experiment 2 thus far demonstrate that the correlation between in-person and online ratings was significantly higher for women than for men, even though both were extremely high. Furthermore, mean ratings were lower when given online, but only for women. Mean completion times were longer online, but again only for women. The differential condition effect for women on the ratings in particular has implications for how one chooses to collect such data.

Relation to other familiarity databases. The solid boxes in Tables 5 (women) and 6 (men) show the correlations between the online and in-person ratings and those from a number of other databases. The rating scale and instructions for the present study were based on the Balota et al. (2001) study, and that is the study with which our ratings correlated most strongly. Note that all the correlations with the Hoosier database were quite low. The familiarity scale used by Nusbaum et al. (1984) had different anchor points and labels, and as a result, 216 of our 300 words got a rating of 7 (on their 7-point scale), and none had a mean of less than 4.08.

The by-condition differences in these correlations appear small, but several were significant. For both men and women, the in-person ratings correlated significantly better with the Balota norms than did the online ratings. For women, there were no other differences, but for men, two of the other three comparisons reached significance, also favoring in-person ratings.

Because we used the instructions and rating scale of Balota et al. (2001), it is appropriate to compare our mean ratings with theirs. In all four cases (in person and online, for both women and men), *t* tests showed that our mean ratings were significantly lower than Balota et al.'s mean of 4.31 (men, in person, $t = -3.02$, $p < .05$; men, online, $t = -3.95$, $p < .001$; women, in person, $t = -18.63$, $p < .001$; women, online, $t = -22.20$, $p < .001$).

To summarize this subsection, for both men and women, the in-person familiarity ratings correlated significantly better with the Balota norms than the online ratings did. For men, in-person ratings correlated significantly better for two of the other existing databases as well. For both genders and for both online and in-person ratings, the mean ratings for our participants were significantly lower than the mean rating in the Balota norms.

Comparison of completers versus noncompleters. Twenty-four participants (21 women) completed the online phase but failed to keep their subsequent lab appointments. Their data may inform us about the performance of nonequivalent samples, since a χ^2 goodness-of-fit test showed that the gender composition of the completers (58% women) differed from that of the noncompleters (85% women) [$\chi^2(1) = 6.04$, $p < .05$]. The groups did not differ on age [$M = 23.5$ for completers, $SEM = 1.18$;

Table 5
Bivariate Correlation Matrix for Several Measures of Familiarity, Female Participants

	In Person	Online (Withdrew)^{a}	Balota	Bristol	Hoosier	MRC^{b}
Online	.99	.94	.87	.81	.29	−.66
In person		.94	.88^{c}	.81	.29	−.66
Online (withdrew)^{a}			.85^{d}	.78^{d}	.30	−.64
Balota				.83	.31	−.70
Bristol					.23	−.66
Hoosier						−.43

Note—Boxes show the correlations between the online and in-person ratings and those from other databases. ^{a}These data are from women who did not complete the whole experiment (see the Comparison of Completers Versus Noncompleters section). ^{b}MRC familiarity ratings were inverse and square root transformed to achieve normality (Tabachnick & Fidell, 2001). ^{c}Coefficient differs from the "Online" coefficient ($p < .05$) directly above it. ^{d}Coefficient differs from both the "In Person" and "Online" coefficients directly above it (both *p*s $< .05$).

Table 6
Bivariate Correlation Matrix for Several Measures of Familiarity, Male Participants

	In Person	Online (Withdrew)[a]	Balota	Bristol	Hoosier	MRC[b]
Online	.98	.33	.89	.81	.33	−.68
In person		.31	.90[c]	.81	.36[c]	−.70[c]
Online (withdrew)[a]			.25[d]	.20[d]	.10[d]	−.24[d]
Balota				.83	.31	−.70
Bristol					.23	−.66
Hoosier						−.43

Note—Boxes show the correlations between the online and in-person ratings and those from other databases. [a]These data are from men who did not complete the whole experiment (see the Comparison of Completers Versus Noncompleters section). [b]MRC familiarity ratings were inverse and square root transformed to achieve normality (Tabachnick & Fidell, 2001). [c]Coefficient differs from the "Online" coefficient ($p < .05$) directly above it. [d]Coefficient differs from both the "In Person" and "Online" coefficients directly above it (all ps $< .001$).

$M = 24.3$ for noncompleters, $SEM = 1.26$; $t(77) = 0.50$, $p = .62$].

We compared the data from these 24 participants with the data from participants who completed both phases of the experiment, using only the online data from those who completed the online phase first. The columns and rows labeled "Online (Withdrew)" in Tables 5 and 6 show the correlations between each item's mean rating for those participants who withdrew and the other mean ratings discussed above. For women (Table 5), ratings from the participants who withdrew correlated .94 with the ratings given by the participants who completed both phases of the experiment, regardless of condition. Both of these correlations were significantly weaker than the .99 correlation between online and in-person ratings for women who completed the experiment ($z = 15.37$, $p < .001$). For men (Table 6), ratings from the participants who withdrew correlated much more weakly (.33 and .31) with the ratings given by the participants who completed both phases of the experiment. Both of these correlations were significantly weaker than the .98 correlation between online and in-person ratings for men who completed the experiment (the smaller z was 11.74, $p < .001$). We will have more to say about these very low correlations, and the small number of participants on which they are based, below.

Several significant differences emerged in the correlations with existing databases, too. The correlations for ratings made by women who withdrew (shown in the dotted box in Table 5) were significantly weaker than the correlations for ratings made in person, for two of the four databases (smaller $z = 2.53$, $p < .05$). In these same two cases, the correlations in the dotted box were also significantly weaker than those for ratings made online by women who completed the experiment (smaller $z = 2.04$, $p < .05$). The results for male noncompleters were even more dramatic. The correlations for ratings made by men who withdrew (shown in the dotted box in Table 6) were astonishingly weak. For all eight comparisons, the coefficients in the dotted box are significantly weaker (smallest $z = 3.54$, $p < .001$).

As in our analysis above, we analyzed the ratings with a multilevel linear mixed-effects ANCOVA, with participant and item as crossed random effects (Table 7). The gender difference observed in the full data set above was seen again here, with men giving higher ratings. Completion status was not significant, nor was the interaction of these two factors.

We log transformed the completion times to achieve normality and analyzed them with an ordinary multiple regression analysis (Table 8). Note that in this analysis, there was only one observation per participant in the data object (i.e., the person's completion time for the only session being considered), and so the traditional p value was not anticonservative here.

Table 7
Summary of the Multilevel ANCOVA for Variables Predicting Familiarity Ratings

Variable	Regression Coefficient (B)	Standard Error of B	t
Main Effects			
Gender (M)	0.49	0.20	2.47*
Withdrew (Yes)	0.14	0.21	0.70
Interaction			
Gender × withdrew	0.06	0.55	0.11

Note—The label in parentheses shows which level of the two-level factor is to have the coefficient added to the ratings estimate. The default level of these factors (F, No) is determined alphabetically by the statistical software. M, male; F, female. *$p < .05$.

Table 8
Summary of the Regression Analysis for Variables Predicting Log Completion Times

Variable	Regression Coefficient (B)	Standard Error of B	t
Main Effects			
Gender (M)	−0.32	0.11	−3.07**
Withdrew (Yes)	−0.22	0.11	−2.00*
Interaction			
Gender × withdrew	−0.69	0.28	−2.49*

Note—The label in parentheses shows which level of the two-level factor is to have the coefficient added to the time estimate. The default level of these factors (F, No) is determined alphabetically by the statistical software. Completion times were log transformed prior to analysis. M, male; F, female. *$p < .05$. **$p < .01$.

Table 9
Mean Completion Times (With Standard Errors of the Means) As a Function of Participant Gender and Completion Status (in Minutes)

	Completed Both Phases[a]		Completed Online Only	
Gender	*M*	*SEM*	*M*	*SEM*
Women	30.43	2.68	26.71	2.25
Men	24.04	2.14	10.33	1.45

Note—Completion times were log transformed prior to analysis. [a]Only the data from the online phase are included, and only if that phase was Session 1.

Both main effects were significant. The participants who withdrew performed the task more quickly than those who returned to complete the experiment, and men performed the task more quickly than women. Over and above the main effects of gender and completion status, there was also a significant interaction between the two. Men who withdrew performed Session 1 in less than half the time, on average, as compared with the other participants (Table 9). Post hoc analyses showed that the effect of completion status on completion time was significant both for women ($B = -0.10$, $SE\ B = 0.01$; $t = -13.99$, $p < .001$) and for men ($B = -0.78$, $SE\ B = 0.01$; $t = -60.52$, $p < .001$).

Because of the small number of men in the noncompleters group, we performed a nonparametric test on the ranks of the completion times, using the Mann–Whitney *U* test. This test makes no assumptions about the normality of the underlying data or the relative sizes of the samples, but it does account for imbalances in these relative sizes. This analysis, too, showed that the effect of completion status on completion times was significant for men ($z = -2.71$, $p < .01$).

Finally, *t* tests showed that for women, the noncompleters' mean was significantly lower than the Balota et al. (2001) mean [$t(299) = -14.88$, $p < .001$]. For men, the noncompleters' mean was significantly *higher* than the Balota et al. mean [$t(299) = 2.42$, $p < .05$].

To summarize this subsection, noncompleters and completers differed in their gender composition but not in their ages. Noncompleters' ratings correlated significantly less well with existing norms than did completers' ratings, in 12 of the 16 comparisons. Female noncompleters had a significantly lower mean rating than the Balota norms, but male noncompleters had a significantly higher mean. Noncompleters finished the task significantly more quickly than completers; this effect was especially dramatic in the case of men, but it was significant for women, too.

Can we learn anything from 3 participants? We must be careful not to overinterpret the results of the completers versus noncompleters analyses, because there were so few men in the noncompleter group. This small group is an intriguing puzzle, though. These participants were not identifiable as outliers during data screening. Their completion times were all within 5 min of each other (note the *SEM*s in Table 9), but these were not the shortest times in the data set. The distribution of logged completion times was quite normal (skew = 0.16, kurtosis = −0.35; both *p*s > .55). Nor could these participants be identified as outliers on the basis of their means or variances: None had either the highest or the lowest mean rating, and none had either the highest or the lowest variance.

The concern about a sample size of 3 remains, but additional analyses argue that the responses of male noncompleters truly were different from those of other subsets of participants and that the results were not simply due to the small sample size. The first row of Table 10 shows the correlation coefficients for ratings from our male noncompleters and existing norms (copied from Table 6). For row 2 of the table, we took 10 random samples of size $n = 3$ from the men who completed the experiment. For each of these random samples, we calculated the correlation between the online ratings and existing norms. The values shown are the averages of those 10 coefficients. Rows 3 and 4 were calculated in an analogous way for the female participants.

For all databases, the coefficient shown in row 1 differed significantly from all of the others (smallest $z = 2.12$, $p < .05$). In fact, not a single one of the 120 new correlation coefficients computed in constructing Table 10 was as weak as its corresponding value in row 1. It is thus not the case that any sample of size $n = 3$ will produce the extremely poor correlations that our male noncompleters did. There must be some other explanation. Prudence dictates that these findings be replicated with larger samples, but this is an intriguing avenue for future research.

GENERAL DISCUSSION

The present study adds to the small literature looking at the effects of testing condition on ratings of potential stimuli, rather than on people's ratings of themselves. We

Table 10
Comparison of Bivariate Correlations for Very Small Random Samples of Participants

	Balota	Bristol	Hoosier	MRC[b]
Online, men who withdrew	.25[c]	.20[c]	.10[c]	−.24[c]
Online, men who did not withdraw[a]	.64	.58	.25	−.49
Online, women who withdrew[a]	.69	.63	.25	−.52
Online, women who did not withdraw[a]	.71	.66	.23	−.53

[a]Correlations shown are averages computed over 10 random samples of size $n = 3$. [b]MRC familiarity ratings were inverse and square root transformed to achieve normality (Tabachnick & Fidell, 2001). [c]Coefficient differs from all three coefficients below it ($p < .05$).

did not have strong reasons to believe that testing condition would be more relevant for ratings of items than for self-ratings, except insofar as the participants might believe that there were correct and incorrect answers for ratings of words. However, we still were not able to rule out a possible condition effect on stimulus ratings, because even the existing findings on self-ratings are so equivocal.

In Experiment 1, with two separate samples, we found a high correlation between mean online and in-person ratings for women and a significantly weaker correlation for men. In Experiment 2, with a single sample, the correlations were extremely high—but even here, the correlation for women was significantly stronger than that for men. We thus have evidence that the relationship between online and in-the-lab performance is stronger for women, whether these comparisons are between subjects (Experiment 1) or within subjects (Experiment 2).

Even with the strong correlations, there was substantial variation in the means obtained in the two conditions for particular items (representing 31% of the rating scale in Experiment 1 and 13% of the scale in Experiment 2). Such differences could easily result in misclassifications if one were constructing groups of items in different ranges on the scale (e.g., high vs. low). We showed that even in Experiment 2, with identical participants making the ratings at both times and with the correlation between the two sets of ratings being nearly perfect ($r = .99$), this happened more than 5% of the time. In Experiment 1, which we think more nearly approximated real-world testing situations (see below), items were differentially categorized as high versus low nearly 18% of the time, even though the two sets of ratings correlated .93. A less-than-perfect correlation, no matter how strong it appears, is thus no guarantee that one set of ratings can be used in place of another.

If the mean values are instead to be used as continuous predictor variables in regression models, mean differences such as these are likely to affect the conclusions of regression analyses. In our analysis of the data from Fischer (2007), we showed that by using one set of ratings, it is possible to predict morphological family size, but if we use the other set of ratings, it is not possible. Unfortunately we did not have access to a similar data set with which to test the Experiment 2 ratings in this way, but our finding will nevertheless be of interest to researchers who use ratings to classify items for further use or analysis. Clearly, additional research is needed to determine the extent to which shifting item classification occurs across different item types and methods, but these results suggest that researchers should consider how their method might influence classification even if the ratings derived from different methods are highly correlated.

We also found interactions between gender and testing condition in Experiment 2. We do not have a compelling explanation for these interactions based on gender roles, technophobia, or personality characteristics. Nevertheless, these findings will be of interest to researchers who recruit from undergraduate psychology participant pools, which typically consist of a majority of women. The differential condition effect for women (on the mean ratings, in particular) has implications for how one chooses to collect rating data.

In Experiment 2, we were able to assess the correlations between our data and existing norms. Correlations varied as a function of testing condition in half of the comparisons. For both men and women, ratings made in person correlated with the Balota et al. (2001) database significantly better than did ratings made online. For men, two of the other three comparisons were also significant, both again favoring the ratings made in person. These findings make sense, since none of these other norms were collected via the Internet. These results suggest that one cannot simply assume equivalence across testing conditions. More widespread dissemination of norms gathered online would help researchers make more appropriate comparisons of studies.

Our study also has something to say about nonequivalence of samples because noncompleters differed significantly from the completers in terms of gender composition. This matters because gender had a main effect in three of the five analyses, and when it was significant, it tended to have a very large coefficient. This makes sense in that gender is necessarily a between-subjects effect, whereas the other effects were usually within-subjects effects (as was noted above, individual differences constitute the single largest source of variation in many psychology studies). Gender often interacted, too, over and above its significant main effect.

The participants who completed both phases of Experiment 2 and those who dropped out produced dramatically nonequivalent data. Noncompleters' correlations with existing databases were significantly lower than the corresponding correlations for completers, in 12 of 16 instances. Further analyses showed that noncompleters (especially men) made their ratings significantly more quickly than completers, too. Male noncompleters were also the only subgroup to have a significantly higher mean rating than the Balota et al. (2001) database.

As was mentioned above, we must be careful about these noncompleter results, because there were so few men in the noncompleter group. Supplemental analyses showed, though, that a small sample size does not by itself produce the pattern of results seen for our male noncompleters (see Table 10). This intriguing issue awaits additional research.

Several effects in Experiment 2 appear to be numerically small even when statistically significant. For example, although significant at $p < .001$, the mean difference in ratings as a function of testing condition was only about one tenth the size of the mean difference in Experiment 1. This is not altogether surprising, given that the data in Experiment 1 came from two different samples of participants, whereas the data in Experiment 2 came from one sample, measured twice. Correlation differences, too, were larger in Experiment 1 than in Experiment 2, even when significant in the latter. We echo Buchanan (2007), who concluded in a review of the generally small method effects in studies like ours, that "the practical significance of these differences will vary depending on how the tests are to be used" (p. 454). Assessment of practical significance is made even harder because ours is the first study to compare stimulus ratings, holding the form of the instrument constant, with a within-subjects design.

The larger effects might be the more realistic estimates, because our experience in recruiting participants for these two experiments suggests the presence of two virtually nonoverlapping populations of participants: those who do not mind coming to the lab and those who would strongly prefer not to. Not everyone who refused to participate stated a reason, but those who did invariably said that they did not want to come for an in-lab session. In addition, despite the study description saying that there were two required phases, several of these participants asked whether they could do just the online phase.

The present study also adds to the small literature in which the same computer administrations have been compared in different settings. As was discussed in the introduction, Cronk and West (2002) found that mean ratings were the same for an Internet version of a morality questionnaire whether administered in class or at home, and Chuah et al. (2006) similarly concluded that ratings on personality inventories were equivalent whether collected via the Internet in a computer lab in groups of 10–18 or alone at home. These conclusions contrast with our own, in that we found a significant effect of testing condition in both experiments.

Epstein and Klinkenberg (2001) noted that "researchers lose almost all hope at controlling the experimental environment when they decide to collect data via the Internet" (p. 303; see also Krantz et al., 1997). In the laboratory setting, there is much more control over the environment, but even here there can be variation along dimensions such as perceived supervision and accountability. In our view, this variation offers the most likely explanation for the contrasting conclusions of the present study and earlier work.

Perceived Supervision and Accountability

Cronk and West (2002) and Chuah et al. (2006) included a manipulation of supervision or proctoring in their studies. Both studies concluded that the factor did not affect the data in important ways. This methodology sounds similar to that in the present study, except that, both in Cronk and West and in Chuah et al., the presence or absence of an experimenter was confounded with the presence or absence of other participants: In the unsupervised/unproctored conditions, participants were assumed to be alone while completing the survey, but the supervised/proctored conditions always took place in groups. The presence of an experimenter in these studies can be presumed to have increased perceived accountability, but the presence of (sometimes many) other participants would seem to have worked against this manipulation. Perhaps the effect of supervision could be diluted by the presence of these other participants.

A related but separate issue is perceived anonymity. Joinson (1999) told half of his participants that their responses were anonymous and could not be linked to them. Testing of all participants was done in groups, in the same physical setting, with an experimenter present (thus making these like the supervised/proctored conditions in the studies above). The only thing that differed was the instruction that the participants were given. The manipulation had strong effects: Social anxiety and social desirability scores were significantly higher in the nonanonymous condition, whereas self-esteem was significantly higher in the anonymous condition.

In the present study, we opted not to introduce any manipulations having to do with supervision or anonymity. Instead, both testing conditions closely approximated current operating procedures. In between-subjects designs like our Experiment 1, we believe that perceived supervision and accountability are much higher in person, especially given the fact that our participants took part in the experiment one at a time. In Experiment 2, with a within-subjects design, we observed a smaller effect of testing condition; but even here, the participants knew that they would have to do half of the study in person, so perceived supervision was probably higher for the in-person condition. Perceived accountability may have been comparable in the two conditions, insofar as the participants knew that they were expected to participate in both phases of the study and that we must, therefore, have had some way of keeping track of which ratings belonged to which participant even in the online condition. However, we did not assess the perceptions of supervision or accountability, and neither did Cronk and West (2002) or Chuah et al. (2006). Also, like Joinson (1999), we did not assess perceptions of anonymity. These should be investigated more systematically in future research comparing identical computerized administrations in different settings.

Conclusions and Directions for Future Research

The question of whether the traditional laboratory-based administration or Internet administration is to be preferred does not have a straightforward answer. In the context of gathering ratings to be used in a subsequent psycholinguistic study, it seems likely to us that the traditional in-person setting would lead to a more conscientious effort by participants, who may believe that their behavior is being observed, at least indirectly, by the experimenter or that their data can be identified as theirs (e.g., Smyth, Dillman, & Christian, 2007). Mean completion time was longer online (for women), which may appear to argue that online performance is more conscientious, but we must exercise caution in interpreting completion times because of the possibility of a curvilinear relationship. As was mentioned above, the online participants in Experiment 2 provided not only the shortest completion times, but also the longest (perhaps they actually walked away from the task for minutes at a time or took a phone call).

Internet administration of surveys is valid under many circumstances, and in some cases, it is clearly preferable. Joinson (1999, 2001) noted that there may be less serious social desirability effects with online administration and that participants may be more honest or forthcoming with information. In our study, because the participants were rating words rather than themselves, issues such as these seem unlikely to have been behind our significant condition effects.

Researchers doing pilot studies can make educated guesses about variables such as personal relevance, likelihood of socially desirable responding, and whether anonymity will be desired. Investigations with stronger anonymity manipulations (and assessments of their effectiveness) will be very useful. Teasing apart anonymity and supervision effects will also be important. The full benefits of Internet-based research (e.g., testing thousands of participants from all over the world) are available only in conditions in which supervision is impossible. Anonymity, on the other hand, can be varied to a certain extent even with Internet studies.

If psychology participant pools are to be used, gender will be a thorny issue. Whenever such participants are allowed unrestricted freedom to volunteer for a study, the sample is likely to be overwhelmingly female. We saw this in Experiment 1 (84% women) and in the self-selected noncompleters in Experiment 2 (85% women). Researchers have known for a long time that random samples from psychology participant pools are not likely to be representative of the population to which they usually want to generalize. It could be, though, that artificially creating representativeness on gender brings with it other problems. As the present study has shown, gender is a factor that can have a very large main effect and can interact with other variables.

Future research should determine whether different participants volunteer for lab-based and online studies. This relates to our observation of differences between completers and noncompleters. Noncompleters can be conceptualized in at least two non-mutually-exclusive ways. They might be representative of less responsible participants, or they might be representative of participants who favor online studies to in-person studies. Willingness to travel to a lab is almost certainly a graded dimension, and our noncompleters probably fall closer to the *avoid in-person studies* end than our completers do. Researchers could easily enough develop a measure of this willingness and look at how it relates to performance in a variety of tasks and settings. Additional work on this question, including more direct assessments of conscientiousness and motivation, will increase our understanding of the complex issues involved.

AUTHOR NOTE

We thank Sean Seaman for a valuable discussion about aspects of our experimental design. Correspondence concerning this article should be addressed to L. H. Wurm, Department of Psychology and Program in Linguistics, Wayne State University, 5057 Woodward Avenue, 7th Floor, Detroit, MI 48202 (e-mail: lee.wurm@wayne.edu).

Note—Accepted by the previous editorial team, when John H. Krantz was Editor.

REFERENCES

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory & Language*, **59**, 390-412. doi:10.1016/j.jml.2007.12.005

Balota, D. A., Pilotti, M., & Cortese, M. J. (2001). Subjective frequency estimates for 2,938 monosyllabic words. *Memory & Cognition*, **29**, 639-647.

Birnbaum, M. H. (1999). How to show that 9 > 221: Collect judgments in a between-subjects design. *Psychological Methods*, **4**, 243-249. doi:10.1037/1082-989X.4.3.243

Birnbaum, M. H. (2004). Human research and data collection via the Internet. *Annual Review of Psychology*, **55**, 803-832. doi:10.1146/annurev.psych.55.090902.141601

Buchanan, T. (2007). Personality testing on the Internet: What we know, and what we do not. In A. N. Joinson, K. McKenna, T. Postmes, & U.-D. Reips (Eds.), *Oxford handbook of Internet psychology* (pp. 447-459). Oxford: Oxford University Press.

Buchanan, T., Ali, T., Heffernan, T. M., Ling, J., Parrott, A. C., Rodgers, J., & Scholey, A. B. (2005). Nonequivalence of on-line and paper-and-pencil psychological tests: The case of the prospective memory questionnaire. *Behavior Research Methods*, **37**, 148-154.

Chuah, S. C., Drasgow, F., & Roberts, B. W. (2006). Personality assessment: Does the medium matter? No. *Journal of Research in Personality*, **40**, 359-376. doi:10.1016/j.jrp.2005.01.006

Cronk, B. C., & West, J. L. (2002). Personality research on the Internet: A comparison of Web-based and traditional instruments in take-home and in-class settings. *Behavior Research Methods, Instruments, & Computers*, **34**, 177-180.

Dandurand, F., Bowen, M., & Shultz, T. R. (2004). Learning by imitation, reinforcement, and verbal rules in problem-solving tasks. In J. Triesch & T. Jebara (Eds.), *Proceedings of the Third International Conference on Development and Learning: Developing social brains* (pp. 88-95). La Jolla: University of California, San Diego, Institute for Neural Computation.

Dandurand, F., Shultz, T. R., & Onishi, K. H. (2008). Comparing online and lab methods in a problem-solving experiment. *Behavior Research Methods*, **40**, 428-434. doi:10.3758/BRM.40.2.428

Denscombe, M. (2006). Web-based questionnaires and the mode effect: An evaluation based on completion rates and data contents of near-identical questionnaires delivered in different modes. *Social Science Computer Review*, **24**, 246-254. doi:10.1177/0894439305284522

Epstein, J., & Klinkenberg, W. D. (2001). From Eliza to Internet: A brief history of computerized assessment. *Computers in Human Behavior*, **17**, 295-314. doi:10.1016/S0747-5632(01)00004-8

Fischer, K. (2007). *Constituent usefulness effects on the recognition of compound words*. Unpublished honors thesis, Wayne State University.

Gosling, S. D., Vazire, S., Srivastava, S., & John, O. P. (2004). Should we trust Web-based studies? A comparative analysis of six preconceptions about Internet questionnaires. *American Psychologist*, **59**, 93-104. doi:10.1037/0003-066X.59.2.93

Hewson, C., & Charlton, J. P. (2005). Measuring health beliefs on the Internet: A comparison of paper and Internet administrations of the Multidimensional Health Locus of Control Scale. *Behavior Research Methods*, **37**, 691-702.

Joinson, A. (1999). Social desirability, anonymity, and Internet-based questionnaires. *Behavior Research Methods, Instruments, & Computers*, **31**, 433-438.

Joinson, A. (2001). Knowing me, knowing you: Reciprocal self-disclosure in Internet-based surveys. *CyberPsychology & Behavior*, **4**, 587-591. doi:10.1089/109493101753235179

Krantz, J. H., Ballard, J., & Scher, J. (1997). Comparing the results of laboratory and World-Wide Web samples on the determinants of female attractiveness. *Behavior Research Methods, Instruments, & Computers*, **29**, 264-269.

Lahl, O., Göritz, A. S., Pietrowsky, R., & Rosenberg, J. (2009). Using the World-Wide Web to obtain large-scale word norms: 190,212 ratings on a set of 2,654 German nouns. *Behavior Research Methods*, **41**, 13-19. doi:10.3758/BRM.41.1.13

Meng, X. L., Rosenthal, R., & Rubin, D. B. (1992). Comparing correlated correlation-coefficients. *Psychological Bulletin*, **111**, 172-175. doi:10.1037/0033-2909.111.1.172

Meyerson, P., & Tryon, W. W. (2003). Validating Internet research: A test of the psychometric equivalence of Internet and in-person samples. *Behavior Research Methods, Instruments, & Computers*, **35**, 614-620.

Norris, J. T., Pauli, R., & Bray, D. E. (2007). Mood change and computer anxiety: A comparison between computerised and paper measures of negative affect. *Computers in Human Behavior*, **23**, 2875-2887. doi:10.1016/j.chb.2006.06.003

Nusbaum, H. C., Pisoni, D. B., & Davis, C. K. (1984). *Sizing up the Hoosier mental lexicon: Measuring the familiarity of 20,000 words*

(Research on speech perception progress report No. 10). Bloomington: Indiana University, Department of Psychology, Speech Research Laboratory.

Richman, W. L., Kiesler, S., Weisband, S., & Drasgow, F. (1999). A meta-analytic study of social desirability distortion in computer-administered questionnaires, traditional questionnaires, and interviews. *Journal of Applied Psychology*, **84**, 754-775. doi:10.1037/0021-9010.84.5.754

Schreuder, R., & Baayen, R. H. (1997). How complex simplex words can be. *Journal of Memory & Language*, **37**, 118-139. doi:10.1006/jmla.1997.2510

Silverstein, S. M., Berten, S., Olson, P., Paul, R., Williams, L. M., Cooper, N., & Gordon, E. (2007). Development and validation of a World-Wide-Web-based neurocognitive assessment battery: WebNeuro. *Behavior Research Methods*, **39**, 940-949.

Skitka, L. J., & Sargis, E. G. (2006). The Internet as psychological laboratory. *Annual Review of Psychology*, **57**, 529-555. doi:10.1146/annurev.psych.57.102904.190048

Smyth, J. D., Dillman, D. A., & Christian, L. M. (2007). Context effects in Internet surveys: New issues and evidence. In A. N. Joinson, K. McKenna, T. Postmes, & U.-D. Reips (Eds.), *Oxford handbook of Internet psychology* (pp. 429-445). Oxford: Oxford University Press.

Stadthagen-Gonzalez, H., & Davis, C. J. (2006). The Bristol norms for age of acquisition, imageability, and familiarity. *Behavior Research Methods*, **38**, 598-605.

Tabachnick, B. G., & Fidell, L. S. (2001). *Using multivariate statistics* (4th ed.). Boston: Allyn & Bacon.

Ullman, M. T., Estabrooke, I. V., Steinhauer, K., Brovetto, C., Pancheva, R., Ozawa, K., et al. (2002). Sex differences in the neurocognition of language. *Brain & Language*, **83**, 141-143.

Whitaker, B. G., & McKinney, J. L. (2007). Assessing the measurement invariance of latent job satisfaction ratings across survey administration modes for respondent subgroups: A MIMIC modeling approach. *Behavior Research Methods*, **39**, 502-509.

Wilson, M. (1988). MRC Psycholinguistic Database: Machine-usable dictionary, version 2.00. *Behavior Research Methods, Instruments, & Computers*, **20**, 6-11.

Wurm, L. H. (2007). Danger and usefulness: An alternative framework for understanding rapid evaluation effects in perception? *Psychonomic Bulletin & Review*, **14**, 1218-1225.

Wurm, L. H., & Seaman, S. R. (2008). Semantic effects in naming and perceptual identification, but not in delayed naming: Implications for models and tasks. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **34**, 381-398. doi:10.1037/0278-7393.34.2.381

Wurm, L. H., & Vakoch, D. A. (2000). The adaptive value of lexical connotation in speech perception. *Cognition & Emotion*, **14**, 177-191. doi:10.1080/026999300378923

Wurm, L. H., Whitman, R. D., Seaman, S. R., Hill, L., & Ulstad, H. M. (2007). Semantic processing in auditory lexical decision: Ear-of-presentation and sex differences. *Cognition & Emotion*, **21**, 1470-1495. doi:10.1080/02699930600980908

(Manuscript received February 16, 2009;
revision accepted for publication October 20, 2009.)