Stimulus norming
It is too soon to close down brick-and-mortar labs

Lee H. Wurm and Annmarie Cano
Wayne State University

Psycholinguists grapple with an ever-increasing list of control variables, in addition to any that are of theoretical interest. Some variables are subjective constructs like familiarity, concreteness, and semantic or affective connotations. Historically researchers approached these by having participants come to a laboratory and provide ratings for each stimulus, but the use of the Internet in data collection has increased in recent years and is likely to continue doing so. In the context of stimulus norms, the equivalence of lab-based and Internet methodologies has not been extensively examined. We discuss some of the pros and cons of online stimulus norming and touch on several issues to consider. We also highlight some important differences between norms obtained online and those obtained in-person.

Keywords: stimulus ratings, online ratings, data collection by internet

Many variables of interest to researchers who use words as stimuli can be computed objectively, like word length and orthographic/phonological neighborhoods, but several others cannot. This latter category includes subjective constructs like familiarity, concreteness/imageability, semantic or affective connotations, and so on. In the past researchers would attempt to get at these subjective constructs by asking participants to come to a laboratory and to provide ratings for each stimulus.

The use of the Internet in data collection has increased dramatically in recent years (e.g., Birnbaum, 2004). This makes a great deal of sense given how easy it has become, but the question of whether this methodology can simply be substituted for the historical one has not been extensively examined (Cronk & West, 2002; Hewson & Charlton, 2005; see also Skitka & Sargis, 2006).

Terms such as norms and ratings are used imprecisely and inconsistently across different areas of research. We will follow what seems to be current practice...
in psycholinguistics. The term ratings generally refers to the raw data, i.e. the individual stimulus judgments made by each participant. These are distilled in some way to produce norms. In many instances the norm value for an item is simply the mean of that item’s individual ratings.

Any problem affecting ratings is likely to affect the norms created from those ratings. With norms the issue is particularly troubling though, because referring to values as norms implies that they will be made available to other interested researchers who can treat them as stable, established values. Such norms are often used to select stimuli with certain characteristics or included as predictor variables in statistical analyses. As we will see, there is reason to be cautious about assuming that norms created from online ratings can be used interchangeably with those created from ratings made in a traditional lab setting.

How this approach can advance knowledge

With online stimulus norming it is easy to obtain participant samples as diverse as the researcher wants, in terms of age, gender, ethnicity, native language, and even physical location. If a researcher wants to know, for example, whether speakers of German located in Germany have the same word associations as speakers of German located in China, he or she can simply check the IP address from which the data arrive. Some caution must be exercised though, because widely-available software can be used to disguise IP addresses.

Stimulus norming of this type can be done very quickly and with minimal investment of time from laboratory personnel. Data can be collected at all hours of the day and on all days of the week, without any supervision. Use of other kinds of laboratory resources is similarly reduced to nearly zero.

One final advantage derives from the anonymity made possible by the Internet. Although most researchers’ lexical processing interests would not be classified as sensitive, if a researcher does happen to be interested in gathering ratings or lists of taboo words for example, he or she can have increased confidence in the candor of the responses if they are made anonymously. We will have more to say about anonymity below.

Key domains of application

There is a tremendous amount of research contrasting paper-and-pencil versus computerized administration of surveys (for a review, see e.g., Buchanan, 2007). The question of interest to lexical processing researchers is different in two ways.
First, ratings have already been computerized for decades in lexical research, so the comparison of interest is “computerized ratings made in the lab” versus “computerized ratings made off-site.” This question has been examined in a small number of studies that have included comparisons of multiple computerized versions of the same task. No clear consensus has emerged from the results.

Chuah, Drasgow, and Roberts (2006) had participants complete self-ratings on several personality dimensions. Two of the conditions were performed over the Internet: Some participants made the ratings alone and others made the ratings in a computer lab in groups of 10–18 people. There was some evidence of what the authors called differential item functioning for 19 statistical comparisons, but the authors concluded that these were due to chance and the very large number of comparisons that were made.

Silverstein et al. (2007) examined performance on a neurocognitive test battery, comparing an established computerized version with a new Internet-based version. They used a within-subjects design with the same 50 participants in both conditions. The authors’ conclusion was that “Results indicated comparability across the two batteries...” (p. 940).

Finally, Dandurand, Shultz, and Onishi (2008) compared lab-based and Internet-based versions of a computerized problem-solving experiment. Internet participants were found to be 10% less accurate, and were also more likely to drop out of the study.

The results of these studies do not allow any firm conclusion to be reached about the role of the testing environment. In two of the three studies the form of the instrument or survey varied with testing environment, further complicating their interpretation.

The second major difference between the key question for lexical processing researchers and those driving the existing research is that we generally need participants to rate aspects of potential stimuli, whereas existing studies usually have had participants rate some aspect of themselves (personality characteristics, job satisfaction, etc.) or perform some sort of problem-solving. There is, though, a small literature comparing ratings of potential stimuli across administration conditions.

Krantz, Ballard, and Scher (1997) compared mean ratings of female attractiveness made via the Internet to those gathered in the lab. While this was not described as a study in which potential stimuli were being rated, it is of course possible to imagine that the images shown to participants might be potential stimuli for subsequent studies. Krantz et al. found very high correlations between the two sets of mean ratings and concluded that the same psychological processes drive the ratings in both settings.

Balota, Pilotti, and Cortese (2001) had participants rate the subjective familiarity of words. Three groups of participants were contrasted: A group of
Online stimulus norming

undergraduates, a group of healthy older adults, and a diverse group ranging in age from 14–84 years who completed the task on the Internet. Although mean ratings correlated very strongly for the three groups (all \( r > .91 \)), the Internet sample did have a significantly higher mean than the other two samples (both \( p < .004 \)). Balota et al. emphasized the strong correlations and downplayed the importance of mean differences to some extent, but they did offer two possible explanations. The first was education level. 45% of their web-based sample indicated a higher education level than the “Some college” category, but less than 10% of their in-person samples indicated this much education. Higher education could be associated with higher word familiarities. Second, the web-based sample made a single rating of familiarity, while the in-person sample made separate ratings for “written, said, and read” familiarity prior to giving an overall rating. The authors thought it possible that the effect of making these individual ratings first could have lowered the subsequent overall rating.

Recent work shows that neither of these explanations is likely to be sufficient. Barenboym, Wurm, and Cano (2010, Experiment 1) and Wurm, Cano, and Barenboym (2010; see below) both replicated the Balota et al. (2001) finding of a higher online mean. Education level was not assessed in these studies, but because all participants in both studies were from a university subject pool, nearly all would probably be at the “Some college” level. In addition, online and in-person participants used the identical rating procedure, except for location. Thus, while the finding of a higher mean with online ratings appears replicable, determination of its cause awaits further research.

Lahl, Göritz, Pietrowsky, and Rosenberg (2009) used the Internet to obtain ratings for German words on concreteness, valence, and arousal. Mean ratings were then compared to existing norms. Correlations averaged .84. This means that there was substantial unexplained variance between the two sets of norms (nearly 30% on average). In some areas of psychological testing, correlations such as this would be considered acceptable or perhaps even good, but for simple judgments of word attributes, this seems low to us. As we will see below, it is considerably worse than similar correlations reported in two more recent studies, and it is cause for some concern because in many areas of psychology, such norms are used to create stimulus groupings on the rated dimension.

However, as Barenboym et al. (2010) point out, correlations tell only part of the story. Even if item means vary substantially, as long as the relative positions of the means remains consistent, the correlation will be high. Furthermore, interpretation of many studies is complicated somewhat by the fact that the form of the instrument used to make the ratings usually varies along with the testing location. Two recent studies remedied this situation by using the identical rating procedure across different testing environments, and by presenting a fuller picture
of the results. They presented not only the correlations, but statistical comparisons of the mean ratings, and tallies of the percentage of items whose categorization would change depending on which ratings were used.

Barenboym et al. (2010, Experiment 1) asked participants to rate the subjective danger of words. The study demonstrated that correlation coefficients will indeed be more impressive than .84 if the identical rating procedure is used, but that even with very high correlations, each set of ratings would produce different item sets if they were used to create stimulus groups. The authors found that 18% of items would change from a high to a low danger group using a median split. One can argue about the wisdom of using median splits to arrive at stimulus groups, but it remains common practice in many areas of research. More importantly, in a separate regression analysis that did not use any stimulus groupings at all, Barenboym et al. found that the statistical significance of danger as a predictor variable depended on whether the norms were based on online or in-person ratings (p < .05 vs. p = .14). These effects emerged even though the two sets of norms correlated .93.

One might wonder whether Barenboym et al. (2010) overstated the problem, and then exacerbated it by using non-separated categories. That is, high and low are contiguous categories divided at the median, and thus tiny differences can potentially push a stimulus out of one category and into the other. However, the average difference in this experiment was not small — the mean online rating was 22% higher than the mean in-person rating (p < .05). In addition, Wurm et al. (2010) have shown that such effects are present even with well-separated stimulus categories. They extended the research of Barenboym et al. (2010) by examining additional common techniques for stimulus grouping. As expected there was an extremely high correlation between online and in-person familiarity norms (r = .97). Consistent with the findings of Barenboym et al. (2010) and Balota et al. (2001), the mean online rating was higher than the mean in-person rating (p < .001), and consistent with the findings of Barenboym et al. (2010), the two sets of norms resulted in different stimulus groupings. The percentage of stimuli whose categorization depended on which norms were used was 7% for a median split approach, 11% for an extreme quartiles approach, and 13% for an equal-thirds approach. Note that the extreme quartiles approach involves separated categories.

Wurm et al. (2010) also demonstrated that the statistical conclusion reached about whether familiarity is a significant predictor in data analyses depended in many cases on whether in-person or online norms were used, echoing another finding of the Barenboym et al. (2010) study. In both Wurm et al. (2010) and Barenboym et al. (2010) the authors also checked whether their in-person norms and their online norms had significantly different correlations with established existing norms: the MRC Psycholinguistic Database (Coltheart, 1981; Wilson, 1988); the Bristol norms (Stadthagen-Gonzalez & Davis, 2006); the Hoosier norms
Online stimulus norming

(Nusbaum, Pisoni, & Davis, 1984); and the Balota norms (Balota et al., 2001). In five out of the 12 comparisons made, the in-person norms correlated significantly better than the online norms. The reverse was never true.

A final indication that norms derived from in-person ratings are probably to be preferred to those derived from online ratings comes from simulations Wurm et al. (2010) ran. They demonstrated that even with repeated replications of a single rating methodology (i.e. comparing online norms to new online norms; or in-person norms to new in-person norms), some numbers of items are likely to change categories and some statistical conclusions will change. However, they showed that the problems would be significantly worse with repeated online replications than with repeated in-person replications, because of the higher within-item variability for the online ratings. We will return to this issue below.

Currently available hardware and software

There are many programs, commercially available and free, that can accommodate online surveys. Among the more common ones in current use are SONA systems (http://www.sona-systems.com/), Zoomerang (http://www.zoomerang.com/), SurveyMonkey (http://www.surveymonkey.com/), and Remark (http://www.gravic.com/remark/websurvey/), but there are many others. In addition, simple surveys that would suffice for many psycholinguistic applications could be created with any web-page editor and hosted anywhere there is space for simple web-pages and the data that are generated. These programs can be run on virtually any computer that has access to the Internet and an even somewhat recent browser.

IP addresses can be checked to determine the region or country in which a participant is located (e.g., by use of a Java script). Checking of IP addresses has also been done to limit multiple participation (e.g. Buchanan et al., 2005), although this is becoming a more complex issue with dynamic allocation of IP addresses. There are also differences in how Internet service providers operate. Some assign a unique IP address to each user on each request, and others put hundreds or thousands of users behind a single address. Gosling, Vazire, Srivastava, and Oliver (2004) and Birnbaum (2004) discuss possible solutions to the potential problem of unwanted multiple participation (see also Chuah et al., 2006).

Dependent variables

A limited number of dependent variables has been examined to date, which is not surprising given the size of the literature on this topic. As mentioned above,
multiple studies have looked at subjective familiarity (Balota et al., 2001; Barenboym et al., 2010; Wurm et al., 2010). Barenboym et al. (2010) also looked at danger ratings, while Lahl et al. (2009) gathered ratings on concreteness, valence, and arousal. Krantz et al. (1997) examined ratings of female attractiveness for front- and side-view photographs. Barenboym et al. (2010) also examined completion times for their rating surveys.

Commonly explored independent variables

In addition to the obvious variable of rating method (on- or off-site), Barenboym et al. (2010) included participant gender and found several large main effects and interactions affecting both the familiarity ratings made and survey completion times. Balota et al. (2001) found an age difference in estimates of subjective frequency specific to perceptual modality (i.e. reading vs. hearing vs. writing vs. speaking), and Krantz et al. (1997) found age differences in the female attractiveness ratings given to front-view photographs. Balota et al. (2001) also examined the effects of objective word frequency, meaningfulness, and neighborhood size on participants’ mean subjective familiarity ratings. They found that word frequency and meaningfulness were much stronger predictors than was neighborhood size. This study also included a preliminary speculative examination of self-reported education level.

New independent variables and new opportunities for the approach

A number of additional independent variables are of potential importance not only for theoretical reasons but to gain a better understanding of the rating process itself. For example, supervision/proctoring (i.e. the presence of an experimenter) was manipulated in Chuah et al. (2006) and Cronk and West (2002), but the presence or absence of an experimenter was perfectly confounded with the presence or absence of other participants. In fact, in nearly all of the studies reviewed above, in-person participants were tested in groups while Internet participants were assumed to have completed the task alone. Exceptions to this are the Barenboym et al. (2010) and Wurm et al. (2010) studies, in which all raters in all conditions performed the computerized task alone. It will be important to demonstrate that reliable data can be obtained in the absence of supervision and other participants, if the full benefits of Internet norming are to be realized.

Perceived anonymity is another variable worth exploring. Joinson (1999) included an anonymity manipulation, telling half of the participants that their
responses were anonymous and could not be linked to them. However, testing of all participants was done with an experimenter and other participants co-present. We thus have reason to wonder how successful the manipulation was, and we cannot be sure because no manipulation check was used. It will be useful to try to separate the effects of supervision/proctoring from those of anonymity. The full benefits of data collection by Internet are available only under conditions that make supervision impossible, but even with Internet administration it is possible to vary anonymity to some extent (Barenboym et al., 2010).

Another crucial issue has to do with possible non-equivalence of samples. Epstein and Klinkenberg (2001) asserted 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) and concluded further that Internet surveys are unlikely to yield representative samples. Sometimes non-equivalence might be desired, for example if researchers are interested in norms from people other than college sophomores, but unintended non-equivalence is always undesirable.

One potentially interesting kind of non-equivalence involves willingness to travel to a lab. Barenboym et al. (2010) found striking performance differences between participants who completed both an in-lab and an Internet phase of their Experiment 2, and those who completed the Internet phase and then dropped out of the study. Those who dropped out were more likely to be female, which matters because Barenboym et al. (2010) found several large gender main effects and interactions in their familiarity ratings. Familiarity ratings from participants who dropped out had significantly weaker correlations with established norms, compared to ratings from participants who finished the study. Those who dropped out also made their ratings significantly faster than those who finished. This was particularly dramatic for men, and male drop-outs were the only subgroup to have a mean rating significantly higher than that from an established set of norms (Balota et al., 2001). All other subgroups had means significantly lower than Balota et al. (2001).

Future research should investigate the extent to which people who volunteer for lab-based as opposed to Internet studies represent distinct populations. Barenboym et al. (2010) reported that

“…our experience in recruiting participants for these two experiments suggests the presence of two virtually non-overlapping 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 there were two required phases, several of these participants asked if they could do just the online phase” (p. 283).
The willingness-to-travel-to-a-lab dimension could be operationalized, and its relationship to several other variables examined (e.g., conscientiousness, motivation).

There are also a number of practical considerations that arise with online stimulus rating. These are mostly the same as those facing researchers who collect ratings in the lab, but one might imagine that the lack of experimenter co-presence could encourage certain kinds of bad behavior. Institutional Review Boards (IRBs) sometimes unintentionally exacerbate this problem by mandating that participants be given the option of not answering each question. This opens the door for an irresponsible person to simply refuse to answer all of the questions, completing the entire survey with one click of the “I’m Finished” button. This undesirable outcome can be avoided by requiring a separate click for every item, but including “I prefer not to answer” as one of the choices. Researchers should verify ahead of time that this method is acceptable to their own IRBs.

Our general approach has been to mostly ignore the possibility of noncompliance, increase the sample size, and hope that the effects of any bad subjects even out in the end. There are several alternative approaches that might be considered, but for a task like rating, for which a participant’s accuracy cannot be calculated, it is difficult to justify excluding data. Still, although none of these methods is by itself perfect, it is often the case that certain participants show up as possibly questionable under multiple methods. Data from such participants could perhaps be excluded.

One approach is to include a small numbers of catch trials, which might be non-critical items that nearly everybody should give a similar rating (e.g., air should be rated as highly useful for human survival). Another kind of catch trial involves repeating a small number of non-critical items early and late in a session, to see if they receive similar ratings on both encounters. Survey completion times or individual item RTs that are either impossibly fast or unusually slow can be indicators that the participant was not doing the task as intended, and the variance of each participant’s ratings can also be informative. Zero variance probably indicates noncompliance, and unusually large variance might indicate random responding, depending on the distribution of the items to be rated.

A slightly more time-consuming method involves checking the correlation between each individual participant’s ratings and the norms calculated over all participants. This is an unsophisticated approach that ignores questions regarding the assumptions that underlie the use of correlation coefficients, but in general we would expect these correlations to be positive and at least moderately strong. Sometimes a strongly negative correlation is observed, indicating that the participant may have misunderstood the rating task. Correlations nearer to zero are perhaps cause for concern about random responding. With luck, potentially
problematic participants will appear as clear outliers in a histogram or quantile-quantile plot. Here, too, the distribution of items to be rated matters, as disagreement on a small number of items can dramatically change correlation coefficients under some circumstances (if the distributions are badly skewed, for example).

Figure 1 shows such histograms and quantile-quantile plots for two unpublished data sets from our lab. In the top row, three individuals whose ratings correlate < .5 with the overall norms appear as possible outliers. The data set represented in the bottom row shows an even clearer outlier, in this case a person whose correlation is < .1 with the norms. This latter data set also shows the difficulty encountered if one wished to use a fixed cut-off for exclusion, such as $r < .5$. Indeed there probably can be no clear cut-off in the context of subjective stimulus ratings. Any given person’s rating are as valid as anyone else’s ratings, provided that the instructions were understood and the ratings were made in good faith.

**Figure 1.** Histograms and quantile-quantile plots of Pearson correlation coefficients ($r$) for two sets of rating data. Each data point shows the correlation between an individual rater’s data and the overall mean ratings.
Conclusion

In this paper we have outlined some of the issues to consider with online stimulus norming, and reviewed the small literature that has already examined certain aspects of it. The results of these studies suggest that significant and potentially serious differences exist between stimulus norms obtained via the Internet and those obtained by computer in the lab. Nevertheless online stimulus norming is likely to become increasingly common, if for no other reason than its ease. This is perfectly understandable but researchers should be aware of the implications of this choice rather than simply assuming equivalence. There is much work to be done to understand where method differences arise and why, as well as what should be done about them.

As we learn more about these issues, we will also gain a clearer understanding of the extent to which the differences we have highlighted apply to other kinds of psycholinguistic research that might be conducted online. Wurm et al. (2010) concluded that greater variance in the online ratings contributed disproportionately to the method differences they found, and that experiment-to-experiment variation will be significantly worse with online methods than with lab-based methods.

What causes online ratings to be more variable? Based on very limited data, our best guess about the disadvantages of online norming is that they stem from some combination of variations in willingness to travel to a lab, participant conscientiousness, perceived supervision, and distractions in the testing environment. In the lab, a relatively distraction-free environment can be maintained, and the co-presence of the experimenter probably reduces the harmful effects of low conscientiousness. We cannot think of many psycholinguistic applications for which it would be acceptable to surrender these benefits.

References


Author’s address

Lee Wurm
Department of Psychology
Wayne State University
5057 Woodward Avenue (7th floor)
Detroit, MI 48202
USA
lee.wurm@wayne.edu