2003

The World Wide Web and the laboratory: A comparison using face recognition

Mitchell M. Metzger, *Ashland University*
Valerie L. Kristof
Yoest J. Donald

The World Wide Web and the Laboratory:

A Comparison Using Face Recognition

Mitchell M. Metzger
Ashland University
Department of Psychology
401 College Avenue
Ashland, OH 44805
USA

Valerie L. Kristof and Donald J. Yoest, Jr.
The Pennsylvania State University
Department of Psychology
Sharon, PA 16146
USA

Address Correspondence To:

Dr. Mitchell Metzger
Ashland University
Department of Psychology
401 College Avenue
Ashland, OH 44805

419-289-5008 (office)
419-289-5665 (fax)
mmetzger@ashland.edu

Abstract

In recent years, a growing number of psychological researchers have turned to the World Wide Web (WWW) as a resource to access participants in experimental studies. While there are benefits to this approach in conducting psychological research (e.g., access to a potentially large subject pool and faster data collection), there are also concerns regarding this medium (e.g., the validity of the data). In recent years, data collected on-line has been validated by comparing it to data collected in the traditional laboratory setting. This study attempted to build on these previous reports by comparing face recognition data collected on the web and data collected in a laboratory. In two separate experiments, data collected from WWW participants did not statistically differ from data collected with undergraduate college students in a classroom setting. These findings strongly suggest that the WWW may be a viable alternative for researchers conducting face recognition experiments.

The World Wide Web and the Laboratory:

A Comparison Using Face Recognition

The dramatic growth of the Internet and WWW in recent years has prompted psychologists to examine this resource for its potential in collecting experimental data, [1-6] as there are many reasons why using the WWW would be beneficial to psychologists. For instance, people anywhere on the globe with access to a computer and Internet connection could serve as participants, thus making cross-cultural research, international collaborative research, and experiments requiring large numbers of subjects quite feasible. In addition, provided that geographically dispersed participants were completing experiments at any given time, researchers would be able to collect data over a shorter time period. By posting studies on the WWW, experimenters can present graphics, audio files, video files, and other multimedia stimuli, thereby expanding the types of experiments some researchers can conduct. Finally, posting experiments on the WWW requires the participant to access a computer to complete the study, which may eliminate or reduce the need for researchers to purchase, update, and maintain costly laboratory equipment. In essence, there are many arguments why WWW resources would be helpful in the collection of experimental data.

There are definite benefits to collecting data on-line, but there are also drawbacks and limitations. The most important concern surrounding data collected on the WWW is the lack of control the researcher has over the experimental conditions, and thus, its potential effect on the validity of the data. Contrary to data collected in the laboratory, data collected using the WWW is vulnerable to a wider variety of environmental influences. For instance, one on-line participant may be alone in a quiet environment, while another takes part in a group at a crowded campus computer lab. The user's computer speed may also influence how smoothly the program

runs, depending on the type of experiment that is delivered and methodology used in the study.[7] Furthermore, there are no guarantees that WWW participants are fully engaged in the experiment. That is, a participant's focus may be divided among several different computer applications during the time when they should be devoting all of their attention to the on-line experiment. Each of the aforementioned problems could very well lead to invalid data. Obviously, these issues must be addressed prior to the acceptance of the WWW as a reliable and valid tool for conducting experimental research.

In an attempt to validate data collected on the WWW, one straightforward method has been adopted. To determine if on-line data is an accurate reflection of laboratory data, one can simply compare data collected on the WWW with that collected in a laboratory setting. Many researchers have reported success with this method, while testing different procedures and methods. One such study demonstrated that a WWW version of a personality questionnaire yielded results similar to those from students completing the survey in paper and pencil form. [8] Another study examined the perception of dominance from schematic facial stimuli, and also reported that data collected on the WWW mimicked traditionally-collected laboratory data. [9] Krantz and his colleagues compared the results of laboratory and WWW data on the perception of female attractiveness. [10] Similar to the aforementioned studies, Krantz also reported that there were no differences between data collected over the WWW as compared to that collected in the laboratory climate.

These findings, along with other reports suggesting similar trends, [11] provide strong empirical evidence that the WWW may be used as a viable tool for collecting psychological data. However, with all of the studies that have demonstrated valid results in WWW samples, certain methodological techniques have yet to be fully studied, such as face recognition procedures.

That is, some studies have used face stimuli in "line up" procedures or in matching procedures. [12] However, no WWW studies have utilized a traditional face recognition procedure in which a series of faces are presented during the study phase and a subsequent series of faces that are presented during a test phase. Given that face recognition studies using this general procedure are common in the experimental psychology literature, it would be advantageous to determine if these procedures could be accurately studied using the WWW. That is, would the results obtained in a WWW sample produce data comparable to that collected in the laboratory?

Thus, while prior studies suggest that the WWW is a feasible tool for psychologists to utilize, more studies testing a greater variety of procedures and paradigms must be established before the acceptance of WWW data can be generalized across many psychological phenomena. Therefore, the current study attempted to expand the types of procedures that have compared WWW and laboratory data by conducting a memory experiment using a face recognition procedure. By comparing the results from subjects participating in an on-line memory experiment against the participants in a laboratory setting, it may be possible to determine whether web data is similar to laboratory data for face recognition.

We chose to post our experiments at PsychExperiments (http://psychexps.olemiss.edu), a colaboratory for web experiments located at the University of Mississippi. [3,4] As with all experiments posted at this web site, participants must first download the Shockwave web player (available in Windows or Macintosh) prior to participating. In addition, the user also downloads the entire experiment file onto their computer before they can begin the experiment. While this may translate into longer wait times for participants with slower Internet connections, it reduces the concern that the user's bandwidth (the speed of the user's Internet connection) might influence the running of the experiment. That is, this procedure helps assure that the experiment,

once started, runs smoothly in its entirety. Importantly, once the file has been downloaded, the user is further instructed to terminate all other applications that are currently running before proceeding with the experiment. The end result is that, once the experiment is downloaded and the subject initiates the program, the experiment should run comparably on machines with different processor speeds. This allows for a more accurate manipulation of variables such as exposure time, as McGraw and his colleagues have stated that they can reproduce results that are accurate to the millisecond with certain procedures. [1]

In the current study, we attempted to replicate previous findings in the face recognition literature in a WWW sample, and compare the data from these participants to data collected in a structured laboratory setting. Several variables were investigated in the present study. We manipulated the distinctiveness of the face stimuli (experiments 1 & 2), the type of recognition test administered to subjects (experiments 1 & 2), the exposure duration of faces during the study phase (experiment 2), and WWW vs. laboratory testing (experiments 1 & 2). From previous research it is well known that performance is superior for distinctive-looking, rather than typical faces, [13, 14] and that performance on face recognition tasks improves with increased exposure duration during the study phase. [15, 16] McKelvie has also reported data suggesting that distractor-free tests (DF) produced results similar to conventional tests (CV) in face recognition procedures. [17] In a CV test, participants are shown target faces (those seen during study) along with distractor faces (those not viewed during study) during the recognition test. In a DF test, participants are shown the same stimuli during the recognition test that they viewed during the study phase, with no distractors present. (With both tests, the participant simply indicates which faces they remember from the study phase.) It has historically been argued that CV tests are necessary to "keep people honest", and that performance would reach 100% if distractors were

not present during recognition testing. However, prompted by earlier findings from word recognition experiments, [18-20] McKelvie demonstrated in a face recognition test that DF and CV testing produced similar results. [17]

In short, two experiments were designed to compare WWW data and laboratory data, while manipulating the above variables. Experiment 1 constituted a 2 × 2 × 2 design, as distinctiveness, type of test (DF vs. CV), and mode of testing (WWW vs. laboratory) were manipulated. Experiment 2 sought to replicate the findings of experiment 1 and expand the findings to include the exposure time variable (here, subjects viewed faces for either 1 or 5 seconds each during the study phase). Thus, experiment 2 consisted of a 2 × 2 × 2 × 2 design. Based on earlier research, our predictions were that each of the above manipulations would replicate previous results, and that the WWW data would not differ from the laboratory data.

Experiment 1

Method

Participants. A total of 124 participants completed the experiment, with 58 participating on-line and 66 participating in the traditional laboratory setting. The WWW participants consisted of 22 males and 36 females with a mean age of 24.9 years. The laboratory subjects consisted of 21 males and 45 females with a mean age of 22.7 years. The WWW data was collected over a seven week period, and the laboratory data was collected on six separate testing occasions. WWW participants were recruited by sending email notices to psychology department heads of small colleges and universities in Pennsylvania and Ohio, and we requested they inform professors and students at their respective campuses about our experiment. Laboratory participants were recruited from introductory psychology courses at the Pennsylvania State University.

Materials. The stimuli consisted of black and white photographs of college seniors from a different geographical location. Prior to the experiment, 200 photographs (100 male and 100 female) were scanned into a computer and randomly presented to 47 independent observers who rated each face on a seven-point distinctiveness scale (the scale ranged from 1 = very typical, to 7 = very distinctive). From these ratings, the 32 most distinctive faces (sixteen male and sixteen female) and the 32 most typical faces (sixteen male and sixteen female) were chosen as stimuli for the experiment. A t-test was computed on the rating scores of the distinctive ($\underline{M}$ = 4.84) and typical ($\underline{M}$ = 2.77) faces to ensure that they were reliably different from each other (t_{62} = 31.84, $p<.01$). One-half of the distinctive and one-half of the typical faces were randomly chosen to serve as "targets" for the experiment, while the remaining faces were used as "distractors". (That is, the 32 target and 32 distractor faces both consisted of 8 female distinctive faces, 8 male distinctive faces, 8 female typical faces, and 8 male typical faces.) For those participating in the laboratory experiment, these images were presented using PowerPoint slideshow software, through a Pentium computer, and were projected onto a 10-foot screen with a ceiling mounted projector. The faces were approximately 61cm tall × 30cm wide, and all subjects had a clear view of the stimuli as the experiment took place in a room with stadium-type (elevated) seating. As participants were seated in three different rows in the experimental room, the distance of each person from the projector screen varied. However, the average viewing distance was 675cm; therefore, on average, the stimuli subtended a visual angle of 5.2°. Those participating on the WWW logged onto the PsychExperiments website (http://psychexps.olemiss.edu) and completed the experiment on personal computers. The slideshow for the experiment was written using Macromedia's Authorware 5.1 software, and was similar to the presentation viewed by participants in the laboratory condition. The face stimuli were approximately 5.75cm tall ×

3.81cm wide, and assuming an average viewing distance of 76.2cm (30 inches), the visual angle subtended for the WWW participants was 3.9°. (There is, of course, no way of accurately determining how far subjects sat from the monitor as they participated in the experiment.)

Procedure. Before the experiment, those in the laboratory condition signed appropriate consent documentation, which was followed by procedural instructions. Participants in the on-line condition first accessed the PsychExperiments home page where they were visually directed to the current experiment. By clicking the "I agree" button at the end of the consent document, they provided individual consent to participate. As with the laboratory subjects, informed consent was a prerequisite to participating in the on-line experiment.

Before the experiment began, participants in both conditions were instructed to provide several pieces of demographic information, including their age and gender, all while remaining anonymous. Once this information was entered, instructions for the procedure of the experiment were presented on the computer screen for WWW participants, and were verbally described to the laboratory participants. The instructions given to both laboratory and WWW participants were identical, as both groups of subjects completed the same experiment, but under different viewing conditions. Regardless of experimental condition, the experiment was divided into two phases: study and recognition testing. During the study phase, participants were presented with "target" faces that they would be required to later identify, and their memory was assessed during recognition testing. Throughout study and recognition testing, one-half of the faces presented were distinctive and one-half were typical. In addition, one-half of these faces were female and one-half were male. Participants were randomly divided into one of two conditions, and were given either a CV or a DF test during recognition testing. Specifically, the study phase consisted of 32 individually presented faces (16 typical and 16 distinctive) for 2 seconds, with a

1 second inter-stimulus interval (ISI) between each face. The recognition test immediately followed the study phase. For those in the DF condition, the same 32 faces were presented during the recognition test for a period of 5 seconds each. If a participant remembered the face from familiarization, they responded "old", and if they did not remember the face from familiarization they responded "new". For those in the CV procedure, the 32 faces that were viewed during familiarization were randomly mixed with 32 novel (distractor) faces (one-half female/male and one-half distinctive/typical) and also presented for 5 seconds each. These participants also responded "old" if they remembered a face from familiarization and "new" if they did not remember the face. (Those in the laboratory condition circled their response on a prepared answer sheet, while those participating on -line clicked on an "old" or "new" button for each response.) During the CV test, the probability of an "old" face on any given trial was .50, and participants were blind to the testing conditions (DF or CV). They were simply informed that another series of faces would be presented during recognition testing, and they were to indicate which ones they remembered from the study phase. The dependent variable for this procedure was Hit Rate (HR – percentage of targets correctly identified) for both distinctive and typical faces. Other measures, such as false alarms and d prime (d´), were not measured as false alarms are not produced in the DF condition (no foils are present during DF recognition testing). Two variables were between groups (laboratory vs. WWW and DF vs. CV test), and one variable was within groups (distinctive vs. typical faces).

Results

Mean Hit Rate and standard deviations for each of the experimental conditions can be seen in Table 1. As expected, the ANOVA computed on distinctiveness was significant

---

Insert Table 1 about here

[F(1,120) = 130.60, p<.001], as participants had reliably higher HR scores for distinctive (M = .78) compared to typical (M = .63) faces. This finding is not surprising, as laboratories using a variety of face stimuli and experimental conditions have previously recorded this same result. [13,14] Furthermore, there were no observed differences between those participating in either the DF or CV testing conditions. This ANOVA was not significant [F(1,120) = 1.064, p>.05], suggesting that participants in the DF condition performed comparably to those in the CV condition. This finding replicates other investigations that have measured the similarities between these two modes of testing, [17,20] and lends support to McKelvie's finding that DF tests, within certain parameters, may be reliable indicants of face recognition performance. (That is, this experiment utilized 32 faces during the recognition test for those in the DF condition, and 64 faces during the test for those in the CV condition. It is possible that DF and CV tests might produce different results if the number of stimuli were different from what was presented here.) The main thrust of this experiment was to measure whether possible differences between WWW and laboratory testing existed, and statistical analysis of the data suggests that participants in both conditions performed comparably. The ANOVA [F(1,120) = .72, p>.05] did not yield any differences between laboratory and WWW participants, suggesting that this data collected over the WWW did not statistically differ from data collected in the laboratory setting. Furthermore, there were no significant interactions among any of the variables (p>.05).

This experiment provides new and important contributions to researchers investigating the validity of WWW data. Participants performed similarly on DF and CV tests, for both

distinctive and typical faces, regardless of whether completing the experiment on-line or in person. The absence of significant interactions between the mode of testing and other variables suggests that WWW and laboratory participants responded comparably to their laboratory counterparts in this experiment. These results indicate that the WWW can be a viable alternative for collecting face recognition data, and lends growing support the notion that the WWW should be taken advantage of by psychological researchers. However, because of the novelty of these findings, a second experiment was designed. We sought to replicate the results obtained in experiment 1, acquire a larger sample size, and include an additional variable to assess a more complex experimental design. We also used a different method of determining the target and distractor faces that is more consistent with current studies of face recognition.

Experiment 2

The purpose of experiment 2 was to replicate the results of the first experiment with a larger sample, and to also expand the procedure to include another variable. We accomplished this by manipulating the same variables as in experiment 1, and also included the variable of exposure duration (1 vs. 5 seconds) of stimuli during the study phase for each of the experimental groups. As increased exposure to stimuli has been shown to increase face recognition accuracy, [15,16] it was predicted that participants, regardless of group assignment, would perform better with longer exposure to the stimuli during the study phase. The addition of this variable created a 2 × 2 × 2 × 2 design. This allowed us to test whether WWW data collected with a slightly more complex experimental design would still mimic data collected in the traditional laboratory setting. Our predictions for the second experiment were the following: (1) WWW and laboratory data would not differ, (2) DF and CV testing would produce similar results, (3) participants would perform better on distinctive than typical faces, and (4) those in

the 5 second exposure condition would perform better than participants viewing faces for 1 second during study. Based on the results of experiment 1, we did not expect to observe any significant interactions among any of the variables in this experiment.

Method

Participants. A total of 254 participants completed the experiment, with 174 in the WWW condition and 80 in the laboratory condition. The on-line participants consisted of 78 males and 96 females with a mean age of 22.9 years. The laboratory participants consisted of 29 males and 51 females with a mean age of 28.8 years. The WWW data was collected over a seventeen week period, and the laboratory data was collected on four separate testing occasions. The WWW participants were again directed to the PsychExperiments web site by email solicitation to psychology department heads of local colleges and universities (different from those solicited in experiment 1), and the lab participants were recruited from introductory psychology courses at the Pennsylvania State University.

Materials. Forty-eight photographs of males were used as stimuli in this experiment, which were taken from the same set previously rated by 47 independent observers. The 24 most distinctive faces ($\underline{M}$ = 4.76) and the 24 most typical faces ($\underline{M}$ = 2.91) were chosen as stimuli, and these stimulus sets were reliably different from each other ($t_{46} = 22.43$, $p<.01$). To determine target and distractor faces for the experiment, both the distinctive and typical faces were divided into two sets of 12 each (A & B), which did not differ from each other ($p>.05$). Approximately one-half of the participants viewed set A faces as targets and set B faces as distractors, and vice-versa for the other one-half of participants. This counterbalancing procedure ensures that each face is used as a target by approximately one-half the subjects, and used as a distractor for the other one-half of subjects. Similar to experiment 1, WWW participants completed the experiment at the

PsychExps web site while laboratory participants completed the experiment in a classroom at the Pennsylvania State University.

Procedure. The same procedure utilized in experiment 1 was used in this experiment, with the addition of an exposure (1 vs. 5 seconds) variable imposed on each of the experimental groups. As in experiment 1, one-half of the faces were distinctive and one-half were typical during study and recognition testing. Approximately one-half of the subjects viewed 24 faces for 1 second each (with a 1 second ISI) during the study phase, and one-half of subjects viewed the same stimuli for 5 seconds each (with a 1 second ISI) during study. (For approximately one-half of the participants, the target faces consisted of stimuli from set A, and the other half of participants saw target faces from set B.) Participants were also randomly divided into either CV or DF test during the recognition testing phase of the experiment, and the recognition test immediately followed the study phase. For those in the DF condition, the same 24 faces were presented during the recognition test for a period of 5 seconds each. For participants in the CV procedure, the 24 faces that were viewed during study were randomly mixed with 24 novel (distractor) faces and also presented for 5 seconds each (one-half of participants saw target faces from set A and distractor faces from set B, and vice-versa for the other one-half of participants). During the CV test, the probability of a target face on any given trial was .50. As in experiment 1, participants were blind to group assignment. If a participant remembered the face from study, they responded "old", and if they did not remember the face from study they responded "new". (Those in the laboratory condition responded with paper and pencil and those in the WWW condition responded by clicking a button on the screen.) Participants were blind to the testing condition (as in experiment 1), as they were simply told that another series of faces would be presented and they were to identify the faces they remembered seeing during the study phase.

As in experiment 1, the dependent variable for this procedure was HR for both distinctive and typical faces. Three variables were between groups (laboratory vs. WWW, DF vs. CV test, and 1 vs. 5 second exposure duration), and one variable was within groups (distinctive vs. typical face).

Results

Each of the results observed in experiment 1 were replicated in the second experiment (see Table 2 for means and standard deviations). Analysis of HR confirmed a significant effect

---

Insert Table 2 about here

---

for distinctiveness [$F(1, 246) = 72.8$, $p<.01$], as participants demonstrated reliably higher HR for distinctive ($\underline{M} = .73$) compared to typical ($\underline{M} = .62$) faces. Furthermore, there were no reliable differences between CV or DF testing [$F(1, 246) = 1.10$, $p>.05$], which provide additional support that DF testing may be a reliable alternative to CV testing under certain stimulus parameters. The exposure duration variable imposed in experiment 2 produced a significant result [$F(1, 246) = 22.67$, $p<.01$], as participants who viewed faces for 5 seconds ($\underline{M} = .72$) during study demonstrated reliably higher HR scores than subjects who viewed the faces for 1 second ($\underline{M} = .63$) during the study phase. This result also replicates a number of earlier studies that have reported better face recognition accuracy with longer viewing times during the study phase. [15,16] Importantly, there were no differences between the WWW and laboratory data [$F(1, 246) = .09$, $p>.05$], replicating the results of experiment 1. As in experiment 1, there were no significant interactions among any of the variables ($p>.05$).

The results of this experiment replicate and extend the results of experiment 1, suggesting that the WWW data collected in the present experiment was similar to the data collected in the traditional laboratory setting. Experiments 1 and 2 differed in several ways, including the type and number of faces used in the experiment, the number of participants recruited, and the procedure for establishing target and distractor faces. Even with these changes in methodology between experiments the pattern of results are identical to each other, thereby suggesting that these findings are quite reliable. These data, when combined with earlier studies demonstrating the validity of data collected on the WWW, provide strong evidence that the WWW is a viable tool for psychological researchers to explore. [2,6,10]

General Discussion

In two experiments, which manipulated two (experiment 1) and three (experiment 2) other variables, data collected on the WWW were compared to data collected from a laboratory sample. This pattern of results strongly suggests that the WWW may be a viable tool for researchers conducting face recognition studies. These data, when compiled with previous reports comparing web and laboratory data provide compelling evidence that the WWW can be a useful, reliable, and valid resource for psychological researchers. [6,8,10] While there are drawbacks to using this medium for experimental purposes (i.e. lack of experimental control), accurate data can be acquired, under proper conditions, as demonstrated by this and other studies. The strengths of WWW data (i.e. access to a larger subject pool and speed of data collection) should be reason enough to promote this line of research, and will allow researchers to ascertain the promise and/or limits of this rich resource.

Importantly, each of the facial recognition variables manipulated in these experiments produced results consistent with earlier findings. In our experiments, distinctive faces were

better remembered than typical faces, and participants performed similarly on DF and CV recognition tests. Experiment 2 demonstrated that subjects who viewed faces for 5 seconds performed better (i.e. showed higher HR scores) than subjects who viewed faces for 1 second during the study phase. While every result obtained was as predicted, it is critical to note that none of these variables interacted with the mode of testing (WWW vs. laboratory). This suggests that on-line participants performed comparably to laboratory participants on each of the manipulated variables. To our knowledge, these findings are the first to demonstrate a comparison of WWW and laboratory data from a traditional face recognition experiment (but see another report of an on-line face memory experiment that used a matching procedure [12]).

We acknowledge, however, as we did not find any significant differences between our WWW sample and the laboratory sample, that caution must be exercised in interpreting these null results. That is, perhaps an inadequate manipulation was set up between the conditions which prevented us from detecting any differences between the WWW and laboratory samples. To address this issue, effect size was analyzed. In experiment 1, effect size (Cohen's *d*) was $d = .13$, suggesting that the distribution of our two samples shared a great degree of overlap. An analysis of the experiment 2 data was similar, in that $d = .02$, suggesting that the distribution of our two samples overlapped almost completely. By examining effect size, we can more confidently suggest that our sample distributions (WWW and laboratory) were very similar to each other. As Cohen defined small effect sizes as those in the $d = .20$ range, [21] we believe that our data demonstrate that the distribution of scores we collected in our experiments were not different from each other.

McGraw and his colleagues have reported compelling evidence indicating that web-delivered experiments provide data that mirrors findings from laboratory-delivered experiments.

[3,4] While it is acknowledged that the experimenter loses a greater degree of control over the experimental environment in the web-based format, this is countered by the larger sample sizes that achievable with web-based studies. In essence, the larger sample sizes that can be generated with web-based experiments more than compensate for the loss of control a researcher has with this method of delivery. The results from our experiment 2 support this notion, as there were more than double the number of participants in the WWW condition (N=174) as compared to the laboratory condition (N=80). (It is interesting to note that the number of on-line participants (N=58) in experiment 1 was actually fewer than the number of laboratory participants (N=66), yet our data still suggests that both groups performed comparably.) More specifically, it is believed that the extra "noise" created by WWW delivery of experiments is thought to translate into greater variability, which is then countered by large sample sizes. It is noteworthy to mention that variability in the WWW and laboratory samples from experiments 1 and 2 were virtually the same (see Tables 1 and 2). It is not clear why the on-line participants did not show greater variability, as would be expected, compared to the laboratory subjects in the present study. Perhaps one reason for the similar variability scores is because the participants were essentially the same for both conditions. The laboratory participants consisted of students enrolled in psychology courses at the Pennsylvania State University, and the WWW participants were largely recruited from students at other institutions also enrolled in psychology courses. We do not have data indicating the proportion of on-line participants that were recruited from other psychology course sections; however, since we actively sought out these subjects (by contacting department heads), we assume that they made up the majority of participants completing the experiment. It is possible that more variability would have been observed in the WWW groups had these subjects consisted of individuals that were not undergraduate students.

For this and other reasons, future studies should focus on populations other than college students. It would be advantageous to compare a group of non-student participants who complete these experiments on-line and compare them to a group of college students participating in the traditional laboratory setting. This would allow one to determine if WWW data (since it will come from a variety of individuals) is truly comparable to data collected in the college laboratory.

It is obvious that participants tested in the laboratory were treated differently than those tested over the WWW. Participants in the laboratory condition viewed larger stimuli (larger visual angle subtended) in a group setting, viewed the stimuli using PowerPoint, and made their responses with pencil and paper. Those in the WWW condition individually viewed smaller stimuli (smaller visual angle) presented with Authorware, and made responses by clicking boxes on the computer screen. While some may argue that these differences in methodology produced confounding variables that interfere with the interpretation of our data, our intent was to produce two very different conditions in which the stimuli were presented. That is, the intent of this study was to determine if the data we collected on the web was similar to the data we collected in the traditional laboratory fashion. Our logic was similar to early studies [10] which demonstrated that various tests administered on-line produced data similar to tests completed in the laboratory. In these early studies, participants tested in the laboratory viewed stimuli that were slightly different from the stimuli viewed by those tested over the WWW. In the case of survey research comparing laboratory and WWW samples, [8] laboratory participants did not complete the questionnaires on the computer (to make the conditions as similar as possible and reduce confounding variables), but rather did the survey in the traditional paper and pencil fashion. By

intentionally producing two different conditions as we did in this study, we can more confidently conclude that WWW testing produces data similar to that of the traditional laboratory.

As discussed throughout this study, an important criticism of web-delivered experiments is the lack of control the researcher has over the experimental participant. This may be especially relevant for paradigms that test memory, as in the current study. That is, participants may view any test (memory or otherwise) as a personal evaluation and want to perform well. By doing an on-line experiment, users could take advantage of the chance to start over before submitting their data at the end of the experiment. That is, if someone completed the experiment and felt that they performed poorly, they could simply fail to submit their data for that session and complete the experiment a second, or third, time until they felt their performance was satisfactory. We have no data which would indicate whether this happened, but based on the performance of WWW and laboratory subjects in the current experiments, it is unlikely that this scenario occurred. Since both groups scored comparably across each variable that we measured, we assume that the WWW participants did not start over if they felt their performance was not satisfactory. Had a large number of on-line participants taken advantage of this loophole and started over, the performance of the WWW group should have been better than that of the laboratory group. This possibility is yet another reason why the large sample sizes acquired in web-delivered experiments is an important feature of this testing medium. The considerable number of participants cancels out the noise created by the few users who may have participated more than once (see [22] for a discussion of this issue).

Only the passage of time will tell if the WWW will be fully received as a viable source of research by a majority of psychologists, for there are many who remain weary about utilizing this resource for the varied reasons mentioned above. Great advances have been made by those using

this medium, and the number of studies reporting similar findings between web-delivered experiments and laboratory experiments is encouraging. Future studies should incorporate a wider variety of experimental designs and paradigms to determine if certain methodological procedures should be readily accepted, or ruled out, for testing on the WWW. It is quite likely that researchers will be able to accurately test certain experimental designs on-line, while other (yet to be identified) designs will best be tested in the traditional laboratory setting.

Author Note

The results from experiment 1 were presented at the annual meeting of the American Psychological Society in Toronto, Ontario, Canada in June 2001. This research was supported, in part, by a research and development grant awarded to the first author by the Pennsylvania State University, and these data were collected while the first author was employed by PSU. We thank Ken McGraw, Mark Tew, and John Williams at the University of Mississippi for their patience and assistance during the completion of this study, and for the excellent training in Authorware. We also thank Jason Machan for his helpful assistance with the data analysis.
Request for reprints may be sent to Mitchell Metzger, Ashland University, Department of Psychology, 401 College Avenue, Ashland, OH, 44805 (Email: mmetzger@ashland.edu)

References

1. Azar, B. A web experiment sampler. APA Monitor 2000; 31(4):46-7.

2. Birnbaum, MH. (Ed.). (2000). *Psychological experiments on the Internet*. San Diego: Academic Press.

3. McGraw KO, Tew MD, Williams JE. The integrity of web-delivered experiments: Can you trust the data? Psychological Science 2000a; 11(6):502-06.

4. McGraw KO, Tew MD, Williams JE. (2000b) PsychExps: An online psychology laboratory. In Birnbaum MH, ed. *Psychological experiments on the Internet*. San Diego: Academic Press, pp. 219-33.

5. Reips, UD, Bosnjak, M., eds. (2001). *Dimensions of Internet Science*. Lengerich: Pabst.

6. Senior C, Smith M. (1999). The Internet…A possible research tool? The Psychologist 1999; 12(9):442-44.

7. Reips UD. (2000). The web experiment method: Advantages, disadvantages, and solutions. In Birnbaum MH ed. *Psychological Experiments on the Internet*. San Diego: Academic Press, pp. 89-117.

8. Buchanan T, Smith JL. Using the Internet for psychological research: Personality testing on the World Wide Web. British Journal of Psychology 1999; 90:125-44.

9. Senior C, Phillips ML, Barnes J, David AS. An investigation into the perception of dominance from schematic faces: A study using the World-Wide Web. Behavior Research Methods, Instruments, & Computers 1999; 31(2):341-46.

10. Krantz JH, Ballard J, Scher J. Comparing the results of laboratory and World-Wide Web samples on the determinants of female attractiveness. Behavior Research Methods, Instruments, & Computers 1997; 29(2):264-69.

11. Krantz JH, Dalal R. (2000) Validity of web-based psychological research. In Birnbaum MH ed. *Psychological experiments on the Internet*. San Diego: Academic Press, pp. 35-60.

12. Ruppertsberg AI, Givaty G, Van Veen HAHC, Bulthoff H. (2001) Games as research tools for visual perception over the internet. In Reips, UD, Bosnjak, M, eds. *Dimensions of Internet Science*. Lengerich: Pabst, pp. 147-58.

13. Going M, Read JD. The effect of uniqueness, sex of subject, and sex of photograph on facial recognition. Perceptual and Motor Skills 1974; 39:109-10.

14. Valentine T, Bruce V. The effects of distinctiveness in recognising and classifying faces. Perception 1986; 15:525-35.

15. Reynolds JK, Pezdek J. Face recognition memory: The effects of exposure duration and encoding instruction. Applied Cognitive Psychology 1992; 6:279-92.

16. Shepherd JW, Gibling F, Ellis HD. The effect of distinctiveness, presentation time and delay on face recognition. European Journal of Cognitive Psychology 1991; 3(1):137-45.

17. McKelvie SJ. Effects of spectacles on recognition memory for faces: Evidence from a distractor-free test. Bulletin of the Psychonomic Society 1993; 31(5):475-77.

18. Faw, HW. Memory for names and faces: A fair comparison. American Journal of Psychology 1990; 103(3):317-26.

19. Ley R, Long K. A distractor-free test of recognition and false recognition. Bulletin of the Psychonomic Society 1987; 25(6):411-14.

20. Wallace WP. On the use of distractors for testing recognition memory. Psychological Bulletin 1980; 88(3):696-704.

21. Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Earlbaum Associates.

22. Musch J, Reips UD. (2000) A brief history of web experimenting. In Birnbaum MH, ed. *Psychological Experiments on the Internet.* San Diego:Academic Press, pp. 61-87.

Table 1. Mean hit rate (HR) scores for Internet and laboratory participants on both the Conventional (CV) and Distractor-Free (DF) testing conditions. (Standard Deviations are presented in parentheses)

	CV		DF	
	DISTINCT	TYPICAL	DISTINCT	TYPICAL
INTERNET (.18) (N = 58)	.81 (.12)	.65 (.17)	.76 (.14)	.65
LABORATORY (.16) (N = 66)	.78 (.14)	.64 (.16)	.78 (.12)	.59

Table 2. Mean hit rate (HR) scores for Internet and laboratory participants on both the Conventional (CV) and Distractor-Free (DF) testing conditions. 1 and 5 represent the number of seconds the face stimuli were presented during the study phase of the experiment. (Standard Deviations are presented in parentheses)

	CV		DF	
	DISTINCT	TYPICAL	DISTINCT	TYPICAL
INTERNET-1 (.17) (N = 89)	.69 (.19)	.62 (.18)	.68 (.16)	.57
LAB-1 (.17) (N = 40)	.65 (.14)	.56 (.17)	.71 (.14)	.56
INTERNET-5 (.18) (N = 85)	.78 (.16)	.68 (.15)	.74 (.18)	.64
LAB-5 (.21) (N = 40)	.83 (.16)	.67 (.16)	.78 (.19)	.64