Behavior Research Methods
2008, 40 (2), 428-434
doi: 10.3758/BRM.40.2.428

Comparing online and lab methods in a problem-solving experiment

Frédéric Dandurand, Thomas R. Shultz, and Kristine H. Onishi
McGill University, Montreal, Quebec, Canada

Online experiments have recently become very popular, and—in comparison with traditional lab experiments—they may have several advantages, such as reduced demand characteristics, automation, and generalizability of results to wider populations (Birnbaum, 2004; Reips, 2000, 2002a, 2002b). We replicated Dandurand, Bowen, and Shultz's (2004) lab-based problem-solving experiment as an Internet experiment. Consistent with previous results, we found that participants who watched demonstrations of successful problem-solving sessions or who read instructions outperformed those who were told only that they solved problems correctly or not. Online participants were less accurate than lab participants, but there was no interaction with learning condition. Thus, we conclude that online and Internet results are consistent. Disadvantages included high dropout rate for online participants; however, combining the online experiment with the department subject pool worked well.

The Internet has revolutionized the way in which people communicate and retrieve information. Powerful communication tools are transforming many scientific disciplines, including experimental and clinical psychology. The Web allows access to much wider populations—as well as to populations that were previously difficult to reach—in an inexpensive, fast, and convenient way. In clinical psychology, for example, psychological testing and assessment can be done online (see, e.g., Buchanan, 2002).

In experimental psychology, the Web is increasingly being used as an alternative to traditional lab settings for running experiments. In the present article, we review the major advantages and disadvantages of online experiments in comparison with traditional lab experiments; then we compare online versus lab results in problem solving—an area in which this comparison has received less attention.

Online Experiment Pros and Cons

There are many potential advantages for doing an experiment online as opposed to in the lab (see Birnbaum, 2004, for a review of pros and cons). First, experimental procedures can be automated, thus reducing costs and the amount of time spent managing the experiment (Reips, 2002a). This also increases the uniformity of the procedure across participants and may reduce demand characteristics (Reips, 2002a). Second, online experiments can be done in a wider array of settings—not just in the highly constrained setting of the lab (Reips, 2000)—and can include 24-h access (Reips, 2002a), considerations that can increase participants' comfort (Salgado & Moscoso, 2003). Third, ethical standards can be maintained because the experiment is publicly available for criticism and the possibility for the coercion of participants is reduced (Reips, 2002a). Finally, online accessibility allows the targeting of specific audiences (through mailing lists or newsgroups) and broadens the participant pool to Web users, rather than, for example, undergraduate students at a particular university, which may allow increased generalizability of the results (Reips, 2000).

There are also disadvantages to running an experiment online rather than in the lab. First, the environments will be more variable, including noise, lighting, and technical aspects of the equipment. Effects of this variability may be reduced by asking participants to do the study in a particular sort of environment and by checking for statistical outliers. Second, online experiments are vulnerable to multiple submissions. This seems to be generally rare (Reips, 2000), but it may be more likely when participants have strong opinions about the topic (see, e.g., Konstan, Rosser, Ross, Stanton, & Edwards, 2005). The risk of multiple submissions can be reduced by asking for personal information, using password protection or an IP address verification (Reips, 2002b), and by reducing external incentives, such as winning money or a prize. Finally, there may be biases in the final sample: Only interested and motivated participants may start (self-selection) and complete the experiment (Reips, 2002a), and there is evidence that online experiments have higher dropout rates than do those run in the lab. For example, fewer than 20% of the people who reached the first page finished an online experiment (O'Neil, Penrod, & Bornstein, 2003). However, self-selection and dropout may not be restricted to online experiments, since lab experiments that recruit volunteers may face the same issues. It may be possible to reduce dropout through prize or monetary incentives (Bosnjak & Tuten, 2003), but incentives sometimes have

F. Dandurand, frederic.dandurand@mail.mcgill.ca

the opposite effect (O'Neil & Penrod, 2001). Pilot testing of instructions and providing contact information for questions (Michalak & Szabo, 1998) may also reduce dropout. In order to reduce adverse effects of dropout, Reips (2000) also recommends getting dropout to occur before the random assignment to conditions by using, for example, warm-up tasks.

Reliability of Online Experiments

Are the conclusions that are drawn on the basis of online and in-lab data samples similar? Following Reips (2002a), researchers have directly compared the results from the two locations for questionnaires, intelligence testing, and biases in syllogistic reasoning, and they have generally found the results to be the same for the two settings (Gosling, Vazire, Srivastava, & John, 2004; Krantz & Dalal, 2000; Meyerson & Tryon, 2003; Musch & Klauer, 2002; Preckel & Thiemann, 2003; Riva, Teruzzi, & Anolli, 2003). However, in some cases, minor differences have been reported. For instance, in a study about organizational attitudes of employees, online participants tended to be more cynical and to judge their organization more harshly than did lab participants (Eaton & Struthers, 2002).

In the present study, we directly compared the results from both online and lab versions of a complex problem-solving task (Dandurand, Bowen, & Shultz, 2004). The literature provides evidence that relatively short and simple experiments can be run reliably online, but less is known about longer, more cognitively demanding tasks in which distraction and motivation may play a larger role.

The Gizmo Problem-Solving Task

We used a problem called the *gizmo problem-solving task*, which consists in finding—with three uses of a scale—the one gizmo that is either heavier or lighter than the rest of a set of 12 gizmos. After figuring out information about gizmos' weights on the basis of weighing evidence, participants were instructed to label those gizmos, using a color selector tool. Participants had to work on problem trials for 30 min, completing as many trials as possible. Figure 1 presents a screen shot of the program used to present the task. Variants of this class of problems are well-known logical-mathematical tasks (Halbeisen & Hungerbühler, 1995), and a version of this class of problems, called the *coin problem*, was used in a psychology experiment on hints (Simmel, 1953).

In the lab experiment, there were three learning groups (Dandurand et al., 2004). We found that participants who were only told whether their answers were correct (reinforcement learning) were less accurate than those who watched demonstrations of correct solutions (imitation learning) or those who read instructions on how to solve these types of problems (explicit learning).

In the present study, we had two main questions. First, would we find the same pattern of effects for learning

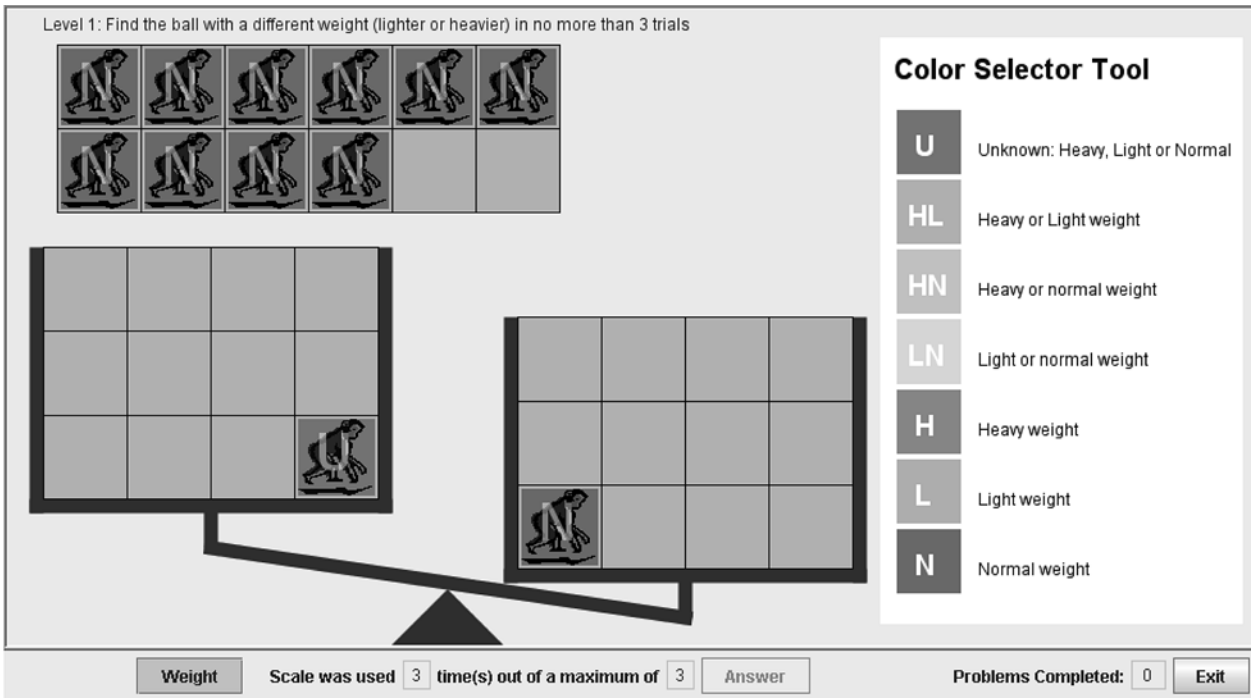

Figure 1. The gizmo problem-solving task. Participants were asked to find—with three uses of a scale—the one gizmo that was either heavier or lighter than the rest of a set of 12 gizmos. The screen is divided into three sections. On the top left is the bank for the 12 gizmos. At the bottom left is a balance scale. On the right is the "color selector tool" that participants used to label gizmos to keep track of their hypotheses about their possible weights. In this example, we see the third weighing in which 11 gizmos have been determined to be of normal weight. Here, 10 are unused and are in the bank, and 2 are on the balance scale. The weight of the gizmo on the left had been unknown; however, because (after the third weighing) the right gizmo is found to be heavier than the left one, the participant can now deduce that this gizmo is of light weight.

group in the online results as was previously found in the lab? Second, would dropout be higher online than in the lab in this rather difficult task? In addition, we explored some demographic characteristics of our final sample of online participants. Thus, we compared the previously collected lab results with new data collected online, treating location (lab or online) as an independent factor.

METHOD

Design

In order to directly compare lab and online experiments, we manipulated the location in which the experiment was run: either in the lab or online over the Internet. We had participants from two populations—undergraduates and others (see the Participants section). Finally, there were three learning conditions.

Reinforcement. Participants in this group got feedback indicating whether they were correct in determining the target gizmo and its weight.

Imitation. Participants in this group watched the program successfully solving five problem trials. The trials presented were randomly selected from all 24 possible trials (12 gizmos × 2 weights).

Explicit. Participants in this group read instructions on how to solve gizmo problems. We matched the amount of information and the time of presentation with that of the demonstrations in the imitation group.

For each trial, we measured accuracy (correct or incorrect identification of the target gizmo and its weight) and response time (RT; time to complete a trial of three weighings). For participants who did not complete the experiment, we noted at which step they dropped out (see Table 1) and which learning group they were in.

Software

The gizmo problem-solving task was implemented using Java—a general-purpose programming language that is well suited for distributing code online to most standard Web browsers. The lab version was a Java application running on a dedicated computer, with data stored on the local hard disk. The online version was a Java applet for use with a Java browser plug-in. The lab program was modified into a client-server architecture to allow data to be transmitted and stored on a Web server. Server-side processing was performed using Perl scripts.

The online version was essentially the same as the lab version, with the following differences.

1. In an attempt to limit dropout, we increased task attractiveness by changing the experiment name from the ball weighing experiment (Dandurand et al., 2004) to gizmo problem solving, as well as by displaying gizmos that would change on each trial, instead of circular balls on every trial.
2. Demographic data were collected online, but not in the lab.
3. The consent form was a Web page online and on paper in the lab.
4. The online consent form asked participants to agree with rules that were implicitly enforced in the lab. For example, the use of paper and teamwork were not allowed.
5. Online participants could optionally request further information about the experiment, which was sent by e-mail after completion of participation. We sent those participants a document with information about the study results. We included this option in order to increase participant commitment and reduce dropout.

Finally, some participants in the lab experiment reported enjoying the task and wishing to play some more. Therefore, in an attempt to limit multiple submissions, we provided a play-only version online with unrestricted access. This version was identical to the experimental version but did not collect any experimental data. Visitors were asked to participate first before accessing the play-only version (available at lnsclab.org/html/BallsWeightExperiment/PlayVersion/play.html), and they were disqualified as potential participants if they said that they had played before.

Pilot testing. Online experiments need to be pilot tested, particularly because participants cannot ask the experimenter questions. The software and instructions were developed in two phases: A lab version was first designed, and it was later ported to an online version. The program's user interface and the instructions for the lab version went through prototyping and incremental pilot testing and updating with 5 users (usually enough to identify about 80% of usability problems; Virzi, 1992). After this initial testing, no interface problems that required fixing were discovered. After porting the program as an applet, 3 testers assessed Web page content and functionality.

RT measurements in Java. We computed RTs using the Java applet. Java was not designed for highly accurate timing—although timing errors are typically less than 100 msec—and there are techniques for improving accuracy, should it be necessary (Eichstaedt, 2001). Since RTs in this experiment are on the order of 2 min, timing errors are likely to be small (less than 0.1%), and thus unlikely to be problematic.

Procedure

The experimental procedure included accepting a consent form, reading instructions, performing a warm-up task (with only 3 gizmos), performing multiple trials of the experimental task (with 12 gizmos) for 30 min, and reading debriefing information. Participants were assigned to one of the learning conditions (reinforcement, imitation, or explicit) in a round-robin fashion, starting with groups with the fewest completed participants. On each trial, the program randomly selected a target gizmo and randomly assigned it either a heavy or a light weight. The participant was asked to determine—with three uses of the scale—which of the 12 gizmos was heavier or lighter than the others.

Table 1
Online Visit Outcomes for the First 7.5 Months of Online Availability

Participation Outcome	Count	Percentage
No Participation		
Loaded first page only	326	54.3
First and about the experiment	17	2.8
Played only	33	5.5
No Java Runtime Env. plug-in	98	16.3
Total	474	79.0
Dropout (Nontechnical)		
Consent form	27	4.5
Personal info	31	5.2
Warm-up task (3 gizmos)	24	4.0
Experimental task (12 gizmos)	17	2.8
Total	99	16.5
Completed the Experiment	27	4.5

Note—Count indicates the number of visitors (out of 600) who dropped out at a particular stage or completed the experiment. Percentage is calculated out of 600.

RESULTS

Participants

Participants in the online experiment were recruited through links on experiment directories, posts in newsgroups, and the psychology department's subject pool. Participants from the subject pool received course credit. The other online participants were not compensated.

Participants in the earlier lab experiment were recruited through the psychology department's subject pool or by using printed ads and personal contacts, and they were eligible to win one performance-based $50 prize. Only subject-pool participants were directly compensated with course credit.

A total of 126 participants took part in the experiment: 63 in the lab and 63 online. Each learning condition (reinforcement, imitation, and explicit) had 42 participants: 21 from the lab and 21 from online. The participants were from two populations: 83 undergraduates—mostly from the subject pool (53 in the lab, 30 online)—and 43 others, including 10 graduate students and engineers tested in the lab and 33 unselected Web users in the online experiment.

Inclusion criteria were the same for both laboratory and online samples. First, participants had to successfully complete the warm-up trials within 30 min. Second, they had to perform above chance (1/24 = 4.2%). Third, they had to be attentive, as indicated by not having any excessively long individual trials. In the lab, data from 3 participants who could not complete the warm-up trials and from 2 participants whose overall accuracies were close to chance were excluded. Online, data from 2 participants were excluded for inattentiveness, and no participants were excluded for low accuracy. Although no online participant worked on warm-up trials for the complete 30-min session, participants who had difficulties understanding instructions or completing warm-up trials would probably have dropped out before reaching the experimental task (see Table 1).

In the online version of the experiment—but not in the laboratory version—we asked about the participants' gender, age, and self-reported prior knowledge of the task before participation. A total of 23/63 (36.5%) of the participants were male, and 63.5% were female. Participants' average reported age was 23.8 years (range: 18–68). Subject-pool participants were younger ($M = 20.9$ years) than unselected Web participants ($M = 26.5$) [$t(61) = 2.92$, $p < .01$]. Most participants had no prior knowledge of the task (92.1%), whereas 4.8% reported some and 3.2% reported good knowledge of the task. Preliminary analyses for accuracy found no reliable main effects or interactions with gender or prior knowledge (Fs < 1.19, ps $> .47$) and no correlation with age ($r = -.17$, $p > .18$). Analyses of RT found no main effects or interactions with gender or prior knowledge (Fs < 1.05, ps $> .49$); however, there was a significant correlation with age ($r = .30$, $p < .02$), indicating that older participants were slower to respond than were younger ones. Since we were primarily interested in accuracy, we collapsed across gender, age, and prior task knowledge in subsequent analyses.

Data Preparation

Each participant worked on trials for 30 min, completing an average of 15.9 trials (range: 4–53). Trials took an average of 134.3 sec. Accuracy for each participant was the proportion of trials solved correctly out of the total number of trials completed. We applied an arcsine transformation to these proportions to stabilize variance (Hogg & Craig, 1995). RT was calculated from the presentation of a new trial to the pressing of the answer button. In order to increase normality, we applied a log transformation to the RT of each trial for each participant; then we calculated the mean log time for each participant.

Participation and Dropout for Non-Subject-Pool Participants

We begin by describing some of the characteristics of our final sample. The online experiment was available only to unselected (non-subject-pool) Web users for 7.5 months. During that time, there were exactly 600 visitors (loading of the welcome page) to the experiment Web site after bot and spider hits were manually discarded. Of these, 376 (62.7%) did not try to begin the experiment, and an additional 98 (16.3%) were unable to participate because they did not have a Java plug-in.[1] A total of 126 people began the experiment, and, of these, 27 (21.4%) completed it. Thus, 4.5% of the initial 600 visitors completed the experiment. Table 1 summarizes these participation outcomes.

By contrast, in the lab, 65/68 (95.6%) of the participants completed the experiment (the other 3 did not complete the warm-up task but did not technically drop out), although it was explicitly stated on the consent form that participants were allowed to withdraw without penalty.

In order to complete the online experiment, we recruited participants for the online study from the psychology department's subject pool. Students selected experiments in which to participate from a list on a departmental Web site, and they received extra credit in a course for writing a report about their participation experience. During the 2-month period it took to run 28 participants from the subject pool, 8 additional unselected Web users also completed the experiment, and an additional 23 participants dropped out after completing the warm-up task, for a total of 63 completed online participants.

Finally, we determined where our online participants were from. Twenty-eight of the 63 participants (44%) were from the subject pool and thus lived in Montreal. The continent or country from which the remaining 35 participants logged into the experiment is shown in Figure 2. By contrast, all lab participants were from the local community.

Dropout Across Groups

In order to examine whether dropout was different across learning groups, we considered the 103 participants who reached a point at which the learning groups diverged (completing the study or dropping out after finishing the warm-up task) and constructed a contingency table of completion or dropout (Table 2). Dropout did not differ by learning group [$\chi^2(2, N = 103) = 0.47$, $p > .79$].

Problem-Solving Results

Finally, we turn to the question of whether the online results essentially replicate the results found in the lab. Mean accuracy across participants by learning group and location is presented in Figure 3. We performed a three-factor ANOVA on accuracy (arcsine transformed), with location (lab, online), learning group (reinforcement, imitation, and explicit), and population (undergraduates, others) as between-subjects factors. The variables of location and population were somewhat overlapping because undergraduates made up 53 out of 63 of the lab participants and 30 out of 63 of the online participants. However,

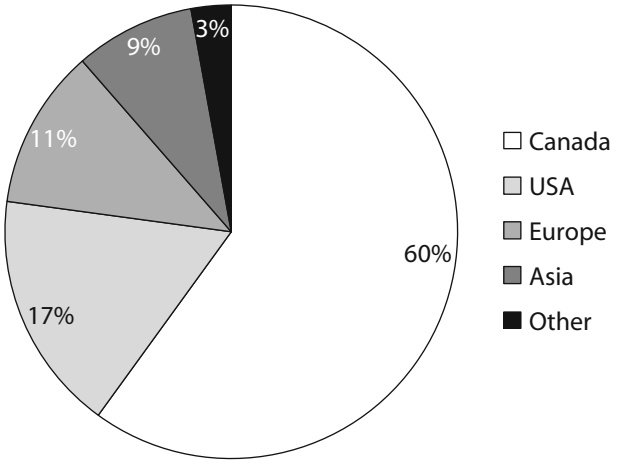

Figure 2. Percentage of the 35 unselected Web participants from different geographic areas.

since there was neither a main effect of population (Fs < 2.93, $p > .09$) nor interactions with population (Fs < 1), we attribute the differences we find between the present results and those of Dandurand et al. (2004) to the variable of location.

For the role of location, we found that online participants ($M = .56$) were significantly less accurate than lab participants ($M = .66$) [$F(1,114) = 6.10$, $p < .02$]. There was a main effect of learning condition [$F(2,114) = 4.05$, $p < .03$], which arose from the reinforcement group ($M = .52$) being less accurate than the imitation ($M = .67$; $p < .01$) and explicit ($M = .66$; $p < .03$) groups (based on Tukey HSD post hoc tests). Imitation and explicit groups were not statistically different from each other ($p > .92$).

We also performed a one-way ANOVA with learning condition as a factor, using online participants only ($n = 63$). We found a main effect of learning condition [$F(2,60) = 3.33$, $p < .05$]. LSD post hoc tests revealed that the reinforcement group ($M = .46$) was less accurate than the imitation ($M = .62$; $p < .04$) and explicit ($M = .64$; $p < .03$) groups, but that accuracies of the imitation and explicit groups were not different from each other ($p > .87$)—just as was found in the lab (Dandurand et al., 2004).

Mean RTs per trial are presented in Table 3. An ANOVA with the factors of location, learning group, and population revealed a main effect of population [$F(1,114) = 4.10$, $p < .05$], suggesting that subject-pool participants ($M = 125$) were faster than others ($M = 152$). Using an ANCOVA, we found that the effect of population disappears [$F(1,60) < 1$] when age is included as a covariate [$F(1,60) = 5.4$, $p < .05$]. Therefore, although there was an effect of population on RTs, it can be mostly attributed to differences in group ages: Subject-pool participants were younger, and younger participants were faster. No other main effects or interactions were significant (Fs < 2.66, ps > .10). Thus, as in Dandurand et al. (2004), there was no effect of learning condition on the average time to complete trials.

To summarize, the online accuracy results replicated the results from the lab experiment for learning condition: Participants in the imitation and explicit learning groups were more accurate than those in the reinforcement group. This result was found in spite of the finding that the online participants were significantly less accurate than the lab participants. We did not find any statistically reliable effects of location or learning group on the speed of completing trials, although subject-pool participants—probably because of their younger age—completed trials more quickly than did other participants. In addition, the online data yielded a wider geographical diversity but were plagued with very high dropout (79%). Fortunately, dropout rates did not differ significantly across experimental groups. Finally, we assessed demographic variables collected online, and none of them predicted significant differences on measured performance, with one exception: Older participants took more time to complete problem trials.

DISCUSSION

As we have seen, lab results of learning condition have been replicated online: Reinforcement-learning participants were less accurate as compared with the imitation and explicit groups, who received more and better information about how to solve this type of problem in the form of demonstrations or instructions. Because we are interested in the effect of learning condition, the main effect of location on accuracy does not affect our interpretation. But why were online participants significantly less accurate than lab participants?

On the basis of informal feedback that we received from lab participants, the task was judged as being fun and interesting, but also as being very difficult. Because high concentration is required, such a difficult task is perhaps more prone to adverse effects of distraction, contributing to the decreased accuracy in the online version of the experiment. Online participants may have simultaneously been working on other tasks or have gotten sidetracked more easily. Furthermore, lab participants were encouraged to do as well as they could to win a performance-based prize, whereas online participants were not given this incentive. In addition, participants who had personally met with the experimenter might have felt more committed to the experiment, and thus have expended more effort. However, this decreased accuracy may better reflect problem-solving skills in more typical environments (unlike in the lab setting). More work needs to be done to determine which factors predict performance differences, particularly the role of cognitive difficulty.

Table 2
Contingency Table of Online Experiment Dropout During the Experimental Task (After the Warm-Up) and of Completed Participations

Learning Group	Reinforcement	Imitation	Explicit	Total
Dropped out	12	12	16	40
Completed	21	21	21	63
Total	33	33	37	103

Note—The table includes participants from Table 1 (27 completed and 17 dropouts) and 59 additional participants (28 completed subject-pool participants, 8 completed Web users, and 23 dropouts).

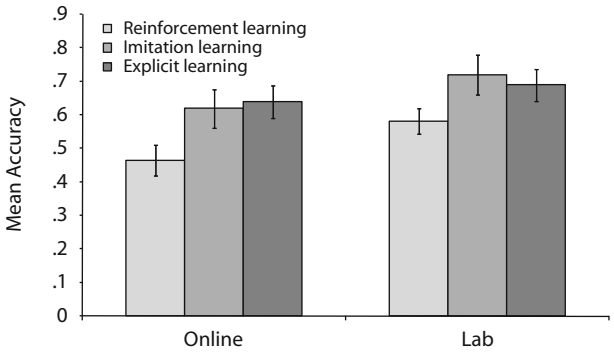

Figure 3. Mean accuracy across participants by learning group for the present online and for the previously collected lab results. Note that because arcsine-transformed accuracies were used for the ANOVA, the error bars (standard errors) of the untransformed data presented here are not indicative of statistical significance.

The length of our experiment—between 45 and 60 min depending on learning group—may have also contributed to dropout. This length was possibly too long for the attention and motivation span of typical Web users. However, incentives successfully limited dropout: All online participants who were recruited through the subject pool and compensated with course credit completed the experiment. As was discussed, offering compensation may encourage multiple submissions, but the subject pool was a relatively safe method, since the registration procedure was the same for all subject-pool participants (whether online or in the lab) and involved the identification of students (name and ID). However, recruiting participants using the subject pool does not realize the full potential of sample diversity obtainable online.

Finally, using the online procedure saved time and resources (no research assistant was needed to run participants). However, the online version of the program was more complex to build than the lab version because of the client-server architecture and its use of server-side Perl scripts. The use of Java in the earlier lab version facilitated the transfer of the experiment to the Internet, but many popular tools for designing and running lab experiments do not necessarily have online equivalents.

Further research—possibly a meta-analysis—is also required to resolve an apparent contradiction in the literature. On the one hand, population samples accessible over the Internet tend to be more diverse than those typically obtained in the lab. Furthermore, environmental characteristics (noise, light, etc.) tend to vary more online than in the lab. As a result, we might expect differences in conclusions drawn from the two settings. When such differences occur, online conclusions might be preferred because they generalize better to wider populations and settings (Reips, 2000). On the other hand, direct comparison of lab and online results as a validation technique assumes no such difference and, in fact, assumes that the lab results are valid and that we are uncertain about the online method. However, it is unclear what we should do if differences are found between online and lab settings: Should we accept online conclusions or not? Further research is also required to better understand the kinds of tasks and settings for which we expect differences between online and lab results and what those differences mean: failure of the online method, or a less selective sample?

Table 3
Mean Response Times to Solve Problem Trials (in Seconds) and Standard Errors Across Participants by Learning Group for the Current Online and for the Previously Collected Lab Results

	Online		Lab	
Learning Group	*M*	*SE*	*M*	*SE*
Reinforcement	151	18	125	13
Imitation	156	14	102	11
Explicit	164	15	108	9

In conclusion, in our complex problem-solving study, we found that the online experiment reproduced the pattern of data found in the lab. There was a wider geographical sample of participants, as well as increased flexibility and savings associated with automation. We also found that online experimenting combined well with a university subject pool. On the negative side, dropout was high, possibly due to the difficulty of the task in combination with its length. Thus, without incentives, the length and difficulty of the experiment appeared excessive for running online, despite being adequate for the lab.

Future research could help answer the following questions: What sorts of experiments are amenable to being run online and under what conditions? How much time are participants willing to spend doing online experiments? And how can motivation and commitment be influenced?

AUTHOR NOTE

This work began as a project completed for a graduate seminar in Human Factors and Ergonomics, taught by D. C. Donderi in the McGill University Department of Psychology. We thank Simcha Samuel for her help with data collection. The research was supported by a McGill Major scholarship to F.D. and by a grant to T.R.S. from the Natural Sciences and Engineering Research Council of Canada. Correspondence concerning this article should be sent to T. R. Shultz, Department of Psychology, McGill University, 1205 Dr. Penfield Avenue, Montreal, PQ, H3A 1B1 Canada (for e-mail, please contact the first author: frederic.dandurand@mail.mcgill.ca).

REFERENCES

BIRNBAUM, M. H. (2004). Human research and data collection via the Internet. *Annual Review of Psychology*, **55**, 803-832.

BOSNJAK, M., & TUTEN, T. L. (2003). Prepaid and promised incentives in Web surveys. An experiment. *Social Science Computer Review*, **21**, 208-217.

BUCHANAN, T. (2002). Online assessment: Desirable or dangerous? *Professional Psychology: Research & Practice*, **33**, 148-154.

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.

EATON, J., & STRUTHERS, C. W. (2002). Using the Internet for organizational research: A study of cynicism in the workplace. *CyberPsychology & Behavior*, **5**, 305-313.

EICHSTAEDT, J. (2001). An inaccurate-timing filter for reaction time measurement by JAVA applets implementing Internet-based experiments. *Behavior Research Methods, Instruments, & Computers*, **33**, 179-186.

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.

HALBEISEN, L., & HUNGERBÜHLER, N. (1995). The general counterfeit coin problem. *Discrete Mathematics*, **147**, 139-150.

HOGG, R. V., & CRAIG, A. T. (1995). *Introduction to mathematical statistics*. Upper Saddle River, NJ: Prentice-Hall.

KONSTAN, J. A., ROSSER, B. R. S., ROSS, M. W., STANTON, J., & EDWARDS, W. M. (2005). The story of subject naught: A cautionary but optimistic tale of Internet survey research. *Journal of Computer-Mediated Communication*, **10**, Article 11. Retrieved 2006 from jcmc .indiana.edu/vol10/issue2/konstan.html.

KRANTZ, J. H., & DALAL, R. (2000). Validity of Web-based psychological research. In M. H. Birnbaum (Ed.), *Psychological experiments on the Internet* (pp. 35-60). San Diego: Academic Press.

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.

MICHALAK, E. E., & SZABO, A. (1998). Guidelines for Internet research: An update. *European Psychologist*, **3**, 70-75.

MUSCH, J., & KLAUER, K. C. (2002). Psychological experimenting on the World Wide Web: Investigating context effects in syllogistic reasoning. In B. Batinic, U.-D. Reips, & M. Bosnjak (Eds.), *Online social sciences* (pp. 181-212). Seattle: Hogrefe & Huber.

O'NEIL, K. M., & PENROD, S. D. (2001). Methodological variables in Web-based research that may affect results: Sample type, monetary incentives, and personal information. *Behavior Research Methods, Instruments, & Computers*, **33**, 226-233.

O'NEIL, K. M., PENROD, S. D., & BORNSTEIN, B. H. (2003). Web-based research: Methodological variables' effects on dropout and sample characteristics. *Behavior Research Methods, Instruments, & Computers*, **35**, 217-226.

PRECKEL, F., & THIEMANN, H. (2003). Online- versus paper-pencil- version of a high potential intelligence test. *Swiss Journal of Psychology*, **62**, 131-138.

REIPS, U.-D. (2000). The Web experiment method: Advantages, disadvantages, and solutions. In M. H. Birnbaum (Ed.), *Psychological experiments on the Internet* (pp. 89-114). San Diego: Academic Press.

REIPS, U.-D. (2002a). Standards for Internet-based experimenting. *Experimental Psychology*, **49**, 243-256.

REIPS, U.-D. (2002b). Theory and techniques of conducting Web experiments. In B. Batinic, U.-D. Reips, & M. Bosnjak (Eds.), *Online social sciences* (pp. 229-250). Seattle: Hogrefe & Huber.

RIVA, G., TERUZZI, T., & ANOLLI, L. (2003). The use of the Internet in psychological research: Comparison of online and offline questionnaires. *CyberPsychology & Behavior*, **6**, 73-80.

SALGADO, J. F., & MOSCOSO, S. (2003). Internet-based personality testing: Equivalence of measures and assessees' perceptions and reactions. *International Journal of Selection & Assessment*, **11**, 194-205.

SIMMEL, M. L. (1953). The coin problem: A study in thinking. *American Journal of Psychology*, **66**, 229-241.

VIRZI, R. A. (1992). Refining the test phase of usability evaluation: How many subjects is enough? *Human Factors*, **34**, 457-468.

NOTE

1. We provided a link for automatic installation to participants who did not have the Java plug-in installed on their computer, but 98 people did not install it.

(Manuscript received August 5, 2007;
accepted for publication September 28, 2007.)