Running Head: Validating Internet Experiments Word Count: 3819

Evaluating the Validity of Internet Experiments: Verbal Overshadowing as a Case Study

Andrew Brand and Ulrike Hahn

School of Psychology, Cardiff University

Correspondence to: Andrew Brand, School of Psychology, Cardiff University, Cardiff, CF10 3YG, UK. Tel: +44 29 20875353. Email: branda@cardiff.ac.uk

Abstract

As part of an inquiry into the validity of internet experiments, the verbal overshadowing effect, the phenomenon that describing a face can impair its subsequent recognition, was investigated in a lab and an internet experiment. The data obtained from both the experiments were analysed in different ways. While the statistically most powerful analysis suggested that the verbal overshadowing effect did not differ between the lab and internet the remaining analyses give considerable reason to believe that they in fact do. This case study illustrates how studies with small effect sizes, the very ones for which internet experimentation seems most attractive, will be inherently difficult to validate.

Introduction

Researchers have increasingly begun to use the internet for conducting psychology experiments. This trend is reflected in the emergence of web sites such as the APS List of Psychological Research on the Net and the Web Experimental Psychology Lab which provide links to online psychology experiments. Internet experiments have both a major practical and theoretical advantage over lab experiments (Reips, 1996, 2000, 2002). Firstly, they are more cost effective with regard to both time and money. Secondly, they enable a larger number of participants to be run. This is important because lab experiments typically involve a small number of participants and hence tend to have low statistical power (e.g., see Erdfelder, Faul & Buchner, 1996; Cohen, 1962, 1994; Sedlmeier & Gigerenzer, 1989). Thus, approximately half of all psychology studies are thought to have reached false conclusions about the null hypothesis (e.g., see Schmidt & Hunter, 2002).

However, the validity of the data obtained from internet experiments may be questionable because they are conducted in a less controlled environment (e.g., participants in an internet experiment may be listening to music or be talking to someone). But several recent studies have shown high agreement between data obtained from internet and lab experiments. For instance, a survey by Musch and Reips (2000) reported high consistency between data from 18 internet experiments and their lab replications. The majority of internet experiments surveyed by Musch and Reips (2000) used text based stimuli but pictorial stimuli have also been used. Pagani and Lombardi (2000) carried out

an internet experiment using facial stimuli and replicated their experiment in the lab to assure that different types of software and hardware configurations used by participants did not substantially affect the interpretation of their stimuli. Again, the data obtained from the lab replication was highly consistent with the data obtained from the internet experiment. Hence, their findings suggest that even with an experiment involving pictorial stimuli the potential noise induced by different types of software and hardware configurations is negligible.

Unlike the validity of internet experiments involving within-subjects designs, the validity of internet experiments involving between-subjects design is more contentious. This is because the power of detecting a between-subjects effect will be diminished by the between-subjects differences in machines and context. However, Klauer, Musch and Naumer (2000) conducted a lab replication of a between-subjects internet experiment and found that the data closely corresponded. Similarly, McGraw, Tew and Williams (2000) conducted an internet experiment involving a between-subjects design which replicated the findings of a previously established effect. Consequently, McGraw et al (2000) concluded that large sample sizes would cancel out Web-induced noise in experiments involving between-subjects designs.

However, internet experiments involving between-subjects designs can also be adversely affected by dropout rates. For as Reips (1996, 2000, 2002) points out a selective dropout in one condition would undermine the findings from the experiment. Moreover, even large sample sizes could not cancel out the adverse effects of differential dropout rates

between conditions. Hence, differential dropout rates between conditions poses the greatest threat to the validity of internet experiments using between-subjects designs.

Although the findings from research investigating the validity of internet experiments have been encouragingly positive there is still need for caution. This is because only very few studies (e.g. Pagani & Lombardi, 2000; Klauer et al, 2000; Krantz, Ballard & Scher, 1997) have directly statistically compared lab and internet results. Moreover, the studies which have conducted direct statistical comparisons have typically compared lab and internet data using correlational and linear regression analyses. These analyses simply correlate the internet means for each conditions with the lab means for each condition. Consequently, these analyses are only feasible when the experiment yields sufficiently large number of data points either in the form of many conditions or repeated measures.

Furthermore, the correlational and linear regression analyses are based *only* on the lab and internet means for each conditions and do not consequently take into account the variances and samples sizes for each condition. Hence, such analyses might obscure crucial differences between lab and internet data. For example, in light of the high correlation obtained, Krantz et al (1997) concluded that there was no significant difference between the data they obtained in the lab and on the internet. But at the same time, Krantz et al (1997) reported that the statistical power to detect the effects of their independent variables and hence the significance levels of these manipulations were consistent between their lab and internet experiments. But if this is true then the conclusion that there was no difference between the lab data and the internet data is

clearly false, since the internet experiment, which involve 316 more participants than the lab experiment, should have considerably *more* power to detect the effects of the independent variables and hence should achieve considerable higher levels of significance than the lab experiment. In other words, the correlational and linear regression analyses mask considerably difference in variability between these lab and internet experiments.

A factorial ANOVA analysis where the study type (lab or internet) is used as a between-subjects factor is amore statistically powerful and reliable technique of comparing lab and internet data precisely because it takes into account variability. Nevertheless, this approach has its own problem, since a failure to reject the null hypothesis and hence conclude that there is a difference between lab and internet data may be due to lack of power rather than the true absence of a significant difference. Consequently, “successful” validation of an internet experiment (i.e., no difference between the lab and internet data) may be simply due to lack of power.

In this paper, we further examine the theoretical and practical implications of these validational issues in the context of a case study by conducting a verbal overshadowing experiment both on the internet and in the lab and comparing their data using a factorial ANOVA.

The verbal overshadowing effect provides a good example of an effect which might benefit from internet experimentation. This is because, according to Meissner's and

Brigham's (2001) meta-analysis involving the data from 29 studies, the verbal overshadowing effect size is small (Z_r=-0.12). Therefore a large number of subjects will be required detect to it. For example, a power analysis shows that in order to have 50% power an experiment would require 267 subjects.

The experimental paradigm for investigation verbal overshadowing involves a between-subjects design which originates from Schooler and Engstler-Schooler (1990). Participants initially view a target face and then perform an unrelated filler task, after which they either describe the target face or perform an unrelated control task. Then they identify the target face from a line-up of faces consisting of several distractors. The finding which is indicative of verbal overshadowing is that participants who described the target face are significantly *less* accurate in identifying the target face than the subjects who performed the unrelated control task. The counterintuitive nature of this finding and its practical implications for eyewitness testimony have subsequently motivated a considerable amount of further research into the verbal overshadowing effect (e.g., see Dodson, Johnson, & Schooler, 1997; Fallshore & Schooler, 1995; Finger & Pezdek, 1999; Meissner, Brigham, & Kelley, 2001; Ryan & Schooler, 1998; Schooler & Engstler-Schooler, 1990; Westerman & Larsen, 1997).

To increase the informativeness of the comparison between the internet and the lab verbal overshadowing experiments, three dependent variables were measured in both the experiments: accuracy, response latencies and confidence. Unlike measuring accuracy and confidence, measuring response latencies in the internet experiment may seem

problematic. However, precise timing to the nearest millisecond is not required and furthermore the results of McGraw et al (2000) suggest that an effect measured in tens of millisecond can be successfully replicated in an internet experiment.

Method

Participants

Lab

The lab sample involved 146 students from Cardiff University who either received course credit or payment for their participation. The sample consisted of 73% (107) females and 27% (39) males. The average age of participants was 21 years old.

Internet

Participants were recruited via the APA Psychology Experiments on the Internet (http://psych.hanover.edu/Research/exponnet.html) and Web Lab for Experimental Psychology (http://www.psychologie.unizh.ch/genpsy/Ulf/Lab/WebExpPsyLab.html).

The internet sample involved 273 participants which consisted of 64% (164) females and 35 % (89)[1] males . The average age of a participant was 27 years old.

Apparatus and Materials

The experiment was written using HTML and Javascript. The experiment conducted in the Lab was run on a PC running Windows NT using Internet Explorer 5.0. The screen resolution was 800 x 600 and the colour depth was 16 bit.

The target photo and six-face photo lineup used by Finger and Pezdek (1999) were used as stimulus material. The target photo was of a 31 years old caucasian male with his head turned at 45 degrees. The six-face photo lineup consisted of six full-frontal photos, one of the target face and five of verbally similar distractor faces.

The internet and lab study both involved a between-subjects design. The dependent variables were the identification accuracy which was either correct or incorrect, the identification confidence which was rated on a 1 (low) to 9 (high) scale and identification response latency which was measured in milliseconds.

Procedure

The procedure for the internet and lab study were identical and as follows:

Participants initially view the target face for 2 seconds. They then perform an unrelated filler task (sliding block puzzle) for 5 minutes, after which they either listed the countries

of the world or verbally described the target face for 5 minutes, finally they identify the target face from a six-face photo lineup and rate their confidence.

Results and Discussion

The results from the lab and internet experiment are summarized in Table 1.

Table 1 about here

A 2 (Study type) x 2 (Verbalization) factorial ANOVA was conducted for each dependent measure (i.e. accuracy, confidence and response latencies). The three main findings from the factorial ANOVAs were as follows: Firstly, participants in the internet experiment were significantly more accurate than participants in the lab experiment, $F(1,415) = 5.47$, $p < 0.05$. Secondly, participants in the internet experiment were significantly more confident than participants in the lab experiment, $F(1,415) = 22.96$, $p < 0.001$. Thirdly, participants in the internet experiment were significantly faster than participants in the lab experiment, $F(1,390^2) = 18.12$ $p < 0.001$.

One might reasonably expect performance in the lab to be superior to performance on the internet. This is because the lab environment is more strictly controlled and hence participants are less likely to be distracted. However, our findings reflect the contrary since participants in the internet experiment were more accurate, more confident and responded faster than participants in the lab experiment. The most plausible explanation

of our findings is that participants in the internet experiment are more motivated than participants in the lab experiment. This is because, unlike the lab participants, who must participate for course credit requirements or who participate for money, internet participants participation is truly voluntary. Furthermore, unmotivated internet participants are more likely to dropout than participants who perform the experiment in the lab (Reips 1996, 2000, 2002).

It would be unwise to maintain that this difference in performance between internet participants and lab participants would undermine the validity of the internet data. In that respect, it is more reasonable to maintain that the data obtained from the internet may be preferable to the data obtained from the lab. Moreover, this difference in performance could be considered theoretically irrelevant because the differences in accuracy, confidence ratings and response latencies between the control and verbalization conditions did not vary as a function of the study type, since the verbalization by study type interaction was non-significant for accuracy ($F = 1.04$, $p > 0.05$), confidence ($F < 1$, $p > 0.05$) and response latencies ($F < 1.70$, $p > 0.05$). Therefore, with regard to investigating the verbal overshadowing effect, findings derived from the factorial ANOVA analyses suggest that the difference between the lab and internet data is relatively innocuous.

However, from the perspective of an investigator examining verbal overshadowing, the crucial matter of interest is whether or not a verbal overshadowing effect was found in each experiment. Consequently, we analysed the data from the lab and internet

experiment separately by conducting the planned comparisons of interest which are indicative of verbal overshadowing. The pattern of results obtained from these separate analyses did not provide a different interpretation of the confidence and response latencies data. However, it did provide a radically different interpretation of the accuracy data, for the verbal overshadowing effect was detected in the lab experiment (i.e., participants in the verbalization condition were significantly less accurate than participants in the control condition) but it was not detected in the internet experiment.

This difference in conclusions derived from the lab and internet data is also reflected in the effect sizes for accuracy. Unlike, the accuracy effect size from the internet experiment (Z_r=-0.04), the accuracy effect size from the lab experiment (Z_r=-0.14) was in close agreement with the effect size derived from Meissner's and Brigham's (2001) meta-analysis (Z_r=-0.12). The fact that the internet findings are based upon a much larger sample size than the lab findings further suggests that the discrepancy between the lab and internet findings is substantial. This is because the experiment with the larger sample size and hence greater statistical power should be more likely to detect the effect and consequently provide a more accurate estimate of the true effect size.

The internet experiment's inability to detect the verbal overshadowing effect is most plausibly due to the difference in dropout rates between the conditions (e.g., see Reips, 1996, 2000): there was a statistically significant greater dropout rate in the verbalization condition than in the control condition, χ^2(1, N=600[3]) = 3.50, p (one-tailed[4]) < 0.05.

Furthermore, accuracy was considerably higher for the verbalization condition on the internet than for the verbalization condition in the lab.

There are three ways in which differential dropout may have inflated accuracy for the verbalization condition in the internet experiment. Firstly, participants who have a poor memory of the target face may be more likely to dropout when required to verbally describe the face than when required to perform the control task. Hence, participants who have a good memory for the target face are more likely to submit their data in the verbalization condition. Secondly, participants who are less motivated may be more likely to dropout when required to verbally describe the face than when required to perform the control task, since verbally describing the target face requires more cognitive effort. Hence, highly motivated participants are more likely to submit their data in the verbalization condition. Thirdly, participants who have a low level of verbal expertise may be more likely to drop out when required to verbally describe the target face than when required to perform the control task. Consequently, participants who have a high level of verbal expertise and are therefore less susceptible to verbal overshadowing (e.g., see Melcher and Schooler, 1996; Ryan and Schooler, 1998) are more likely to submit their data in the verbalization condition.

The differential dropout rate explanation provides a compelling reason for the difference in p-value based conclusions for accuracy in the lab and internet studies and taking into account that the size of the verbal overshadowing effect for the internet study was over one third smaller than the lab study we might be inclined to conclude that there is a

significance difference between the lab and internet data for accuracy. However, since the verbalization by study type interaction was non-significant for accuracy in the factorial ANOVA analysis we are compelled to conclude that there is in fact no difference.

However, the most likely explanation for why the verbalization by study type interaction was not significant is lack of statistical power; the observed effect size of the verbalization by study type interaction was small (Z_r=0.06) and hence the power for detecting the interaction was low (18%). In practical terms, to have an 80% chance of detecting the interaction, one would need to test 2597 participants. Furthermore, this assumes that there are equal sample sizes in each condition and hence at least 1000 participants would have to be tested in the lab!

Lack of sufficient power to detect an interaction between the experimental variable and the study type is not unique to our investigation, where a meta-analysis (Meissner and Brigham, 2001) has shown the effect to be small. It is also not unique to internet experiments which have significantly different dropout rates between conditions rather it is a general problem for validating *all* small effects. This is because if the effect under investigation is small then both the effect sizes for the lab and internet studies will also tend to be small. Hence, the difference between the effect size for the lab and internet study, which corresponds to the interaction between the experimental variable and the study type, will be by definition, small[5].

So, paradoxically the main reason for using internet experiments to investigate small effects (i.e., the inability to obtain large sample sizes in the lab) is also the main reason for why the validity of internet experiments investigating small effects cannot be effectively assessed. This is because effective assessment of the validity of the internet experiment would require an unfeasibly large sample size to be run in the lab. Moreover, the sample size required in the lab experiment in order to validate the internet experiment would be substantially larger than the sample size required to *only* investigate the effect in the lab.

In summary, our case study provides an example where we have good reason to believe that there is significant difference between lab data and internet data, since we obtained the effect in the lab but not on the internet. Moreover, the absence of an effect in the internet experiment can also be cogently explained by the fact that there was a significant difference in dropout rates between conditions. Nevertheless, the interaction between the type of study (lab or internet) and the experimental variable (non-verbalization or verbalization) was non-significant and consequently we were forced to conclude that there was no significant difference between lab and internet data. However, the most plausible reason for why this interaction was non-significant is lack of power and this poses a general problem for validation of any small effect. This is because if the effect being validated is small then a significant interaction between the effect obtained in lab and the effect obtained on the internet will also be necessarily small.

The dilemma then, is that determining whether or not internet experimentation provides a valid alternative to the lab will sometimes be unfeasible in exactly those cases for which it would be practically most beneficial to exploit this new methodology - namely small effects in need of large sample sizes, because, ironically, it would require running an unfeasibly large number of participants in the lab.

Conclusion

One of the main motivations for conducting an internet experiment is that it enables large samples size to be obtained. However, this main advantage may be offset. First, an increase in sample size may be offset by an increase in variance as in the case of Krantz et al (1997), where the power to detect an effect on the internet did not significantly differ from the power to detect the effect in the lab and hence the increase in sample size is practically useless. Second, validating an internet experiment in order to determine whether increase variability or other factors such as drop out rates will compromise its validity, will require running considerably more participants in the lab than simply conducting the experiment *only* in the lab. However, maybe in the future when the issues concerning the validity of internet experimentation are more fully appreciated and understood, the need to validate every internet experiment by conducting a lab replication will not be necessary and then the full benefits of internet experimentation will be reaped.

References

Cohen, J. (1962). The statistical power of abnormal-social psychological research a review. Journal of Abnormal and Social Psychology, 65, 145-153.

Cohen, J. (1994). The earth is round ($p < .05$). American Psychologist, 49, 997-1003.

Dodson, C. S., Johnson, M. K., & Schooler, J. W. (1997). The verbal overshadowing effect: Why descriptions impair face recognition. Memory and Cognition, 25, 129-139.

Erdfelder, E., Faul, F., & Buchner, A. (1996). GPower: A general power analysis program. Behavior Research Methods, Instruments, & Computers, 28, 1-11.

Fallshore, M., & Schooler, J.W. (1995). The verbal vulnerability of perceptual expertise. Journal of Experimental Psychology: Learning, Memory and Cognition, 21, 1608-1623.

Finger, K., & Pezdek, K. (1999). The effect of the cognitive interview on face identification accuracy: Release from verbal overshadowing. Journal of Applied Psychology, 84, 340-348.

Klauer, K. C., Musch J., & Naumer, B. (2000). On belief bias in syllogistic reasoning. Psychological Review, 107, 852-884.

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

McGraw, K. O., Tew, M. D., & Williams, J. E. (2000). The integrity of web-delivered experiments: Can you trust the data? Psychological Science, 11, 502-506.

Meissner, C. A., & Brigham, J. C. (2001). A meta-analysis of the verbal overshadowing effect in face identification. Applied Cognitive Psychology 15, 603-616.

Meissner, C. A., Brigham, J. C., & Kelley, C. M. (2001). The influence of retrieval processes in verbal overshadowing. Memory and Cognition, 29, 176-186.

Melcher, J. M., & Schooler, J. W. (1996). The misremembrance of wines past: Verbal and perceptual expertise differentially mediate verbal overshadowing of taste memory. Journal of Memory and Language, 35, 231-245.

Musch, J., & Reips, U. -D. (2000). A brief history of web experimenting. In M. Birnbaum (Ed.), Psychological experiments on the Internet (pp. 61-87). Orlando, FL: Academic Press.

Pagani, D. and Lombardi, L. (2000). An intercultural examination of facial features communicating surprise. In M. Birnbaum (Ed.), Psychological experiments on the Internet (pp. 169-194). Orlando, FL: Academic Press.

Reips, U.-D. (1996, October). Experimenting in the world wide web. Paper presented at the Society for Computers in Psychology conference, Chicago.

Reips, U. -D. (2000). The web experiment method: Advantages, disadvantages and solutions. In M. Birnbaum (Ed.), Psychological experiments on the Internet (pp. 89-117). Orlando, FL: Academic Press.

Reips, U. -D. (2002). Standards for internet-based experimenting. Experimental Psychology, 49, 243-256.

Ryan, R. S., & Schooler, J. W. (1998). Whom do words hurt? Individual differences in susceptibility to verbal overshadowing. Applied Cognitive Psychology, 12, S105-S125.

Schmidt, F. & Hunter, J. (2002). Are there benefits from NHST? American Psychologist, 57, 65-56.

Schooler, J. W., & Engstler-Schooler, T. Y. (1990). Verbal overshadowing of visual memories: Some things are better left unsaid. Cognitive Psychology, 17, 36-71.

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of the studies? Psychological Bulletin, 105, 309-316.

Westerman, D. L., & Larsen, J. D. (1997). Verbal-overshadowing effect: Evidence for a processing shift. American Journal of Psychology, 110, 417-428.

Notes

1. 20 participants did not state their sex.

2. 25 outliers were omitted from the analysis.

3. This value refers to the number of people who began the internet study.

4. The one-tailed prediction is based upon the results of a previous pilot study which found markedly different dropout rates between conditions.

5. It is possible, but unlikely, that the effect size for the internet study may be medium or large and hence having sufficient power to detect an interaction between the experimental variable and the study type, in this case, may not pose a problem.

Acknowledgments

Andrew Brand was funded by EPSRC Studentship Award No. 00303803.
The authors thank Kimberly Finger and Kathy Pezdek for the slides of the target face and six-face lineup. And they would also like to thank Todd Bailey and Jacky Boivin for their comments on previous drafts of this manuscript.

Table 1: Means and Standard Deviations for Accuracy, Confidence Ratings and Response Latencies as a Function of Study and Verbalization.

	Lab		Internet	
	No Verbal	Verbal	No Verbal	Verbal
N	73	73	147	126
Accuracy	0.66	0.52	0.72	0.68
	(0.48)	(0.50)	(0.45)	(0.47)
Confidence Ratings	5.67	4.89	6.56	5.88
	(1.78)	(2.05)	(2.36)	(2.28)
Response Latencies	11330	17340	8977	12508
	(8275)	(10423)	(5395)	(6557)

Note. Standard deviations are in parentheses. The possible range of confidence ratings was from 1 to 9, where 1 denoted low confidence and 9 denoted high confidence. The response latencies were measured in milliseconds.