Australian Journal of Psychology, Vol. 61, No. 2, June 2009, pp. 107–116.

The Australian Psychological Society Ltd

Internet versus paper-and-pencil survey methods in psychological experiments: Equivalence testing of participant responses to health-related messages

IONI LEWIS[1], BARRY WATSON[1], & KATHERINE MARIE WHITE[2]

[1]*Centre for Accident Research and Road Safety – Queensland, and* [2]*Psychology and Counselling, Queensland University of Technology, Brisbane, Queensland, Australia*

Abstract

Despite experiments being increasingly conducted over the Internet, few studies have tested whether such experiments yield data equivalent to traditional methods' data. In the current study, data obtained via a traditional sampling method of undergraduate psychology students completing a paper-and-pencil survey ($N = 107$) were compared with data obtained from an Internet-administered survey to a sample of self-selected Internet-users ($N = 94$). The data examined were from a previous study that had examined the persuasiveness of health-related messages. To the extent that Internet data would be based on a sample at least as representative as data derived from a traditional student sample, it was expected that the two methodologies would yield equivalent data. Using formal tests of equivalence on persuasion outcomes, hypotheses of equivalence were generally supported. Additionally, the Internet sample was more diverse demographically than the student sample, identifying Internet samples as a valid alternative for future experimental research.

Worldwide, an estimated 15 million people access the Internet each day and reports indicate that every 3 months this rate increases by a further 25% (Rhodes, Bowie, & Hergenrather, 2003). In Australia, a similar pattern has been evidenced, with recent estimates indicating that households with Internet access have almost quadrupled between 1998 (16%) and 2005–06 (60%) (Australian Bureau of Statistics [ABS], 2005–06). With increasing Internet use has come greater use of the Internet as a modern medium for conducting psychological research (Buchanan & Smith, 1999; Gosling, Vazire, Srivastava, & John, 2004).

Internet-based research: Advantages and issues relating to data quality

There are a number of advantages of conducting Internet-based research such as the ability to acquire large and diverse samples; greater time efficiency; the reduced costs and fixed costs (i.e., the costs of conducting an Internet survey remain the same irrespective of the number of respondents); the reductions in data entry errors; the capacity to incorporate visual and auditory stimuli; heightened anonymity and confidentiality, which is particularly advantageous for surveys addressing sensitive issues; and greater convenience for respondents in terms of the time and place of participation (Birnbaum, 2004; Carlbring et al., 2005; Iragüen & de Dios Ortúzar, 2004; Pasveer & Ellard, 1998; Perkins & Yuan, 2001; Rhodes et al., 2003). Accompanying the increased reliance on Internet-administered surveys, however, is the obligation for researchers to demonstrate the reliability, validity, and overall quality of the data obtained via the Internet.

Many studies have sought to establish the worth of Internet surveys by comparing the data they obtain with data obtained from more traditional methods such as paper-and-pencil surveys completed in person or over the phone. These comparative studies have aimed to demonstrate that the different methodological approaches produce equivalent data.

Researchers have undertaken various approaches in their attempts to demonstrate the equivalence of the two approaches (Meyerson & Tryon, 2003).

Correspondence: Ms I. Lewis, Centre for Accident Research and Road Safety – Queensland, QUT Carseldine Campus, Beams Road, Carseldine, QLD 4034, Australia. E-mail: i.lewis@qut.edu.au

ISSN 0004-9530 print/ISSN 1742-9536 online
Published by Taylor & Francis
DOI: 10.1080/00049530802105865

These approaches include *t* tests comparing the means and standard deviations of items/scales (e.g., Whittier, Seely, & St. Lawrence, 2004), psychometric analyses via comparisons of Cronbach alphas and factor structures of scales (e.g., Pasveer & Ellard, 1998); formal tests of equivalence to compare means (or proportions) of specific items (e.g., Epstein, Klinkenberg, Wiley, & McKinley, 2001); and a comparison of item completion rates, response time, and item completion errors for the two methods (e.g., Pealer, Weiler, Pigg, Miller, & Dorman, 2001). Generally, the results of these studies suggest that Internet-based surveys produce data that are at least as reliable, valid, and of equal quality as data obtained via more traditional survey methodologies. Consequently, Internet surveys and more traditional paper-and-pencil surveys have been reported as producing equivalent data. But there are gaps in this literature as well as definitional inconsistency with the term ''equivalence'' that need to be noted and that highlight the need for further research.

In relation to the gaps in understanding, two key omissions are evident. First, these comparison studies have largely been based upon non-experimental research designs (Musch & Reips, 2000). Experimental designs that feature more than one level of an independent variable (and/or more than one independent variable) have been increasingly utilised in Internet research since the late 1990s (Musch & Reips, 2000; for an example, see Kypri & Gallagher, 2003). Despite their increasing usage, few published studies are available that provide a comparison of the data obtained in Internet experimental studies with data obtained using a traditional research methodology such as a paper-and-pencil questionnaire (Musch & Reips, 2000). Second, where experimental designs have been utilised on the Internet, the research is often likely to have a cognitive psychological focus as opposed to other research areas such as social psychology (Musch & Reips, 2000; for an example, see Eichstaedt, 2002). An implication of this cognitive research focus is that the comparison studies are more likely based on a comparison of an experiment conducted on a computer in a laboratory setting versus the same experiment conducted on a computer over the Internet. This methodology differs from comparisons in which the data obtained via a paper-and-pencil-administered survey or questionnaire are compared with the data obtained by a computer-administered version of the same survey or questionnaire completed via the Internet (Musch & Reips, 2000). These limitations notwithstanding, the available evidence regarding psychological experiments on the Internet suggests that such experiments yield equivalent data to more traditional approaches (Musch & Reips, 2000; for an example see Eichstaedt, 2002).

Defining ''equivalence''

As noted previously, various approaches have been utilised in attempts to determine data equivalence. The number of different approaches highlights that definitional ambiguity has surrounded the term ''equivalence'' (Schulenberg & Yutrzenka, 1999). Schulenberg and Yutrzenka cite the definition of equivalence provided by the American Psychological Association within its Guidelines for Computer-Based Tests and Interpretations (American Psychological Association, 1986). According to this definition, one aspect of determining equivalence between computerised tests and the paper-and-pencil versions is, ''if the means, dispersions, and shapes of the score distributions are *approximately the same*'' (italics added). Given this definition, it becomes evident that the absence of a statistical difference found by null hypothesis statistical testing (NHST) does not indicate that two means are the same or ''equivalent''; rather, it suggests that insufficient evidence was found to reject the null hypothesis (Tryon, 2001).

Despite there being substantial literature attesting to the fact that the absence of a significant statistical difference does not indicate statistical equivalence (Anderson & Hauck, 1983; Cook & Campbell, 1979; Tryon, 2001), some studies have examined the equivalence of different data collection strategies by using such methods (e.g., Horswill & Coster, 2001; Perkins & Yuan, 2001; Whittier et al., 2004). Indeed, concluding statistical equivalence on the basis of the absence of a significant difference has been identified as one of the most common misuses of NHST methods (Tryon, 2001). In NHST, the null hypothesis tested is that there is no significant difference between group means (Rogers, Howard, & Vessey, 1993). This hypothesis is different from the research hypothesis tested by formal tests of equivalence. The latter research hypothesis tests whether the means of two groups are equivalent or, more specifically, whether two group means are sufficiently near to each other to be considered equivalent (Cribbie, Gruman, Arpin-Cribbie, 2004; Rogers et al., 1993). In formal tests of equivalence, a researcher seeks to reject the null hypothesis that there is a difference and accept the alternative hypothesis that the two means are equivalent (Rogers et al., 1993). Moreover, the approaches are not mutually exclusive; the results obtained via tests of equivalence using non-equivalence null hypothesis testing methods can often contradict the results obtained via traditional NHST approaches (Cribbie et al., 2004; Rogers et al., 1993).

Internet-based research: Some challenges and concerns

Internet-based samples and the data they derive are not without challenges and concerns. Among some of the most commonly cited issues are sample representativeness and the loss of control over the testing conditions (for a review of the advantages and disadvantages of Internet research, see Birnbaum, 2004). Concerns surrounding the representativeness of Internet samples are often discussed together with the concept of the ''digital divide''. This divide refers to the disparities in Internet access based on socio-demographic dimensions that exist between Internet users and non-users (Rhodes et al., 2003). Early research suggested that the digital divide favoured greater Internet use by younger, more educated, higher income, white male subjects but, with the dramatic increases in Internet use around the world, this issue is becoming less relevant (Rhodes et al., 2003). Recent Australian evidence suggests that the sociodemographic profiles of Internet users are broadening, with gender gaps largely disappearing, and age disparities lessening, but some other disparities in use remain based on education, income, geographical region, disability, and indigenous status (ABS 2004–05; Rhodes et al., 2003; Willis & Tranter, 2006).

Despite the disparities, evidence suggests that Internet samples are at least as representative as other traditionally used samples such as university student samples. In a recent meta-analytical study exploring the relative diversity of a typical self-selected Internet sample with more traditional (predominantly student) samples, Gosling et al. (2004) found that the Internet sample was found to be more diverse than the traditional samples with respect to gender, age, geographic region, education and socioeconomic status. Similarly, comparative studies based on non-random assignment of participants to the Internet and paper-and-pencil conditions (with university students assigned to the latter condition and self-selected Internet users assigned to the former) have also provided evidence of Internet samples being at least as representative as traditional student samples (e.g., Pasveer & Ellard, 1998; Whittier et al., 2004) and more diverse for some variables (e.g., gender and age; Smith & Leigh, 1997). Moreover, comparisons of the outcome data derived from these comparative studies utilising non-random assignment found that traditional student samples and typical self-selected Internet sample produce similar responses (see Pasveer & Ellard, 1998; Smith & Leigh, 1997; Whittier et al., 2004) albeit without the use of formal tests of equivalence. Overall, these studies have concluded that, while Internet samples are not representative of the general population, they are as diverse as student samples, if not more so.

In relation to the loss of control that experimenters have over testing conditions in Internet-based research, this loss of control applies to a range of context-related factors from whether other people are present while a participant is completing a survey through to hardware and software variations across respondents (Skitka & Sargis, 2006). For the latter factor at least, adequate piloting can ensure that the survey runs effectively on a range of computer systems prior to the survey being released on-line (Birnbaum, 2004). But, despite piloting, in studies where stimuli such as audio or video files are included differences in equipment will mean that the stimuli received may vary between participants, and this possibility should be acknowledged by researchers (Birnbaum, 2004; Smith & Leigh, 1997).

In summary, of the issues affecting Internet-based studies, although some are unique to the Internet (e.g., loss of control and the threat to internal validity) others apply also to traditional methodologies. For instance, concerns surrounding the representativeness of convenience samples of students have been long-standing (Sears, 1986) and self-selection occurs in most traditional samples (Madge & O'Connor, 2004). Although Internet methods have limitations, other approaches are not without their limitations. Consequently, Internet-based surveys may be considered at least as acceptable as other survey methods (Harrison & Christie, 2004).

Current study

Even as recently as 2003, it was suggested that ''further comparisons across a range of procedures will help clarify the validity of Internet research in other domains'' (Hewson, 2003; p. 292). Consistent with this suggestion, the aim of the current study was to address limitations in the extant literature relating to the validity of Internet-based research. Specifically, the current study will compare the effectiveness of Internet and paper-and-pencil methods for experimental research in an applied social psychological research context. The study will examine the data derived from an earlier experimental study that had examined the effectiveness of persuasive messages in the context of an important health area: that of road safety (Lewis, Watson, & White, in press). It is important to note that the data utilised in the current study are by way of example only; such data were selected because they were derived from a social psychological experiment that consisted of two independent variables each with two levels and was completed by samples of participants responding to either a paper-and-pencil or an Internet version of the same survey. Furthermore, participants responding to the paper-and-pencil survey were university undergraduate students (i.e., a more traditional

sampling methodology), while those completing the Internet survey represented a sample of self-selecting Internet users. In order to facilitate understanding of the current results, it should be noted that the original experiment from which the data were derived was a 2 (appeal type: positive/humorous, negative/fear-evoking) × 2 (response efficacy: low, high) mixed-group design with appeal type as a between-groups variable and response efficacy as a repeated-measures variable. Additionally, in relation to outcome measures, the advertisements were compared in terms of their persuasive impact on individuals' reported attitudes and intentions.

The main aim of this study was to determine whether an applied social psychological experiment administered to an Internet sample yields equivalent data to that of a more traditional, university student-based sample. This aim is underpinned by the notion that, until the equivalence of a particular administration format is demonstrated empirically, its validity remains unknown (Schulenberg & Yutrzenka, 1999). Currently, the validity of Internet survey methods for experimental social psychological research remains largely unknown. Given that the current study seeks to determine statistical equivalence of data derived from the two sampling methodologies, an additional objective of the study was to illustrate the suitability and usability of equivalence testing in psychological research. The second main aim of this study was to provide empirical comparisons of demographical characteristics of the two samples.

Research hypotheses

It was expected that, based on previous empirical evidence, the Internet sample would be more diverse than the traditional student sample in relation to age and gender. Thus, it was expected that significant differences would be found between the two samples of drivers.

It was expected that, based on previous empirical comparisons of Internet and paper-and-pencil methods, participants in the two conditions should enter and exit the study with equivalent mean scores. More specifically, it was hypothesised that the Internet survey would yield equivalent mean ratings as the paper-and-pencil survey on pre-attitudinal measures, as well as on the outcome measures of persuasion (post-exposure attitudes and intentions). Moreover, it was expected that, when performing comparisons between Internet and paper-and-pencil conditions based on the cells within the original 2 × 2 experimental data (i.e., Positive/High Response Efficacy, Positive/Low Response Efficacy, Negative/High Response Efficacy, and Negative/Low Response Efficacy), the mean post-exposure attitudinal and intentional scores would be equivalent.

Methods

Participants

The study sample consisted of 201 participants (71 men, 130 women). All participants were holders of a current Australian drivers' or motorcyclists' licence. Almost half of the participants ($n = 94$, 46.8%) completed the Internet-based version (this number represents surveys that were completed or that contained only minimal missing data), while 107 participants (53.2%) completed the paper-and-pencil version of the same survey.

The link to the Internet survey was placed on the authors' research centre's homepage. To promote the existence of the Internet survey, the survey and its location were advertised extensively through radio and print media. In addition, an email calling for participants was forwarded to staff at a multifaceted organisation involved in many aspects of motoring (e.g., insurance and travel). This organisation also provided a link to the Internet survey on their homepage, thus increasing the likelihood that drivers would find the study while visiting a driving-related website.

The majority of participants completing the paper-and-pencil version of the survey were undergraduate students studying a first year psychology unit at a major Australian university (74.8%). The remaining participants were second year psychology students ($n = 27$; 25.2%). Thus, participants included in the paper-and-pencil version were considered typical of more traditional undergraduate student samples. The first year students were recruited via a flyer on a university noticeboard while the second year students responded to a request by the researchers made at the end of a lecture. Of all the participants, the first year psychology undergraduate students were the only participants who received an incentive (i.e., partial course credit) for participating. All participants in the paper-and-pencil condition completed the survey in groups.

Measures and procedure

The survey from the original study was divided into two sections: measures assessed prior to exposure to the advertisements and measures assessed following exposure to each advertisement. Prior to exposure, participants were asked to provide demographic information, information about their drinking and driving histories, and attitudes towards drinking and driving. Following exposure to each advertisement (i.e., individuals viewed a low and high response efficacy advertisement in either the positive or negative appeal condition) participants were assessed on their attitudes towards drink driving as well as their intentions to drive after drinking.

Statistical analyses

Sample comparisons. The categorical data relating to the samples' demographic characteristics were analysed using chi-square tests for independence. Post hoc analyses were conducted for all significant chi-square tests using an adjusted standardised residual statistic ($\hat{e}$) (Haberman, 1978).

Equivalence testing of persuasion outcomes. Of the equivalence tests that are available, the Schuirmann (1987) two one-sided tests procedure was selected (Cribbie et al., 2004; Rogers et al., 1993; Seaman & Serlin, 1998). This approach is widely used and offers advantages such as a bounded Type 1 error rate and good power (Dixon & Pechmann, 2005). Two steps are needed to perform a test of equivalence: first, determining what constitutes equivalence; and second, performing two simultaneous one-sided hypothesis tests (Rogers et al., 1993). In determining equivalence, an a priori decision must be made regarding the minimum difference between the means of two groups that would be important enough to make the groups non-equivalent: any difference smaller than delta would be considered meaningless within the context of a particular experiment (Cribbie et al., 2004; Rogers et al., 1993). Thus, two means would be considered equivalent if they differed by less than delta in both a negative (δ_1) and positive (δ_2) direction (Rogers et al., 1993). As noted previously, the argument for greater use of equivalence testing for psychological research has been proffered only recently. Consequently, a standard equivalence criterion for use in such research has not yet been established (Epstein et al., 2001). After reviewing equivalence criteria utilised in available studies (e.g., Cribbie et al., 2004; Epstein et al., 2001; Rogers et al., 1993; Streiner, 2003), the decision was made to utilise the equivalence criterion of $\pm 20\%$ of the mean outcome scores derived in the paper-and-pencil condition. The paper-and-pencil condition was the condition on which the criterion was based because it represents the traditional approach with which the more modern Internet approach is being compared.

The second step in equivalence testing related to the need to perform two simultaneous one-sided tests to establish equivalence. The null hypothesis relates to the non-equivalence of the group means and may be expressed as two composite hypotheses: the upper and lower null hypothesis. The upper and lower hypothesis can be expressed as follows, respectively (Cribbie et al., 2004, p. 3; Seaman & Serlin, 1998):

$$H_{o1} : \mu_1 - \mu_2 \geq \delta_2; H_{o2} : \mu_1 - \mu_2 \leq \delta_1$$

Rejection of the upper hypothesis implies that $\mu_1 - \mu_2 < \delta_2$ and rejection of the lower hypothesis implies that $\mu_1 - \mu_2 > \delta_1$. The logic underpinning the test is that rejection of both hypotheses implies that $\mu_1 - \mu_2$ falls within δ_1 to δ_2, rendering the difference between the means less than the minimum difference of importance (determined a priori) and the means equivalent (Cribbie et al., 2004, p. 3; Rogers et al., 1993; Seaman & Serlin, 1998). Thus, to establish equivalence both one-sided null hypotheses must be rejected. But in determining equivalence, only one test is required; the test relating to the shorter distance between the observed difference (i.e., $\mu_1 - \mu_2$) and either δ_1 or δ_2. The one-sided test with the shorter distance will be associated with the smaller test statistic and the larger p and will be the least likely to be rejected. In instances where the test with the larger p is rejected, the other one-sided test, which will necessarily evidence a smaller p, will not need to be performed because it also will always be rejected. But in instances where the test with the largest p is not rejected, the second test still will not need to be conducted because both tests must be rejected for equivalence to be determined. For the error rate of the equivalence test it follows that, although two sides are being tested, the error rate depends on one side only (i.e., the side with the largest difference) and the critical value chosen needs to be set at alpha for each side of the test (Rogers et al., 1993, p. 554).

In the current study, equivalence tests were performed on a number of pre- and post-exposure variables measured in the Internet and paper-and-pencil versions of the survey for both the study's full sample ($N = 201$) as well as for the cells of the original 2×2 experimental design. Specifically, the variables examined were attitudes towards drink driving (assessed both before and after exposure) and behavioural intentions (assessed after exposure).

Results

Demographical characteristics

The comparisons of the Internet and student samples' demographics are shown in Table I. As can be seen, there was a significant difference between the Internet and paper-and-pencil versions of the survey in terms of gender. There were significantly more female subjects (and fewer male subjects) in the paper-and-pencil condition than the Internet condition. Of the two conditions, the rate of men to women was more equally distributed in the Internet version than in the paper-and-pencil version, for which women outnumbered men at a rate of approximately 4:1. Additionally, an age-related difference was found between the two conditions.

Table I. Sociodemographic characteristics of participants by survey version

Variable	Survey version		
	Internet (%)	Paper and pencil (%)	Significance level[a]
Gender	$n = 94$	$n = 107$	χ^2 ($df = 1$) = 19.15, $p < .001$
Male	51.1	21.5	$\hat{e} = 4.4$, $p < .001$
Female	48.9	78.5	
Age (years)	$n = 94$	$n = 107$	χ^2 ($df = 6$) = 39.40, $p < .001$
<18	0.0	11.2	$\hat{e} = 3.3$, $p < .001$
18–24	12.8	38.3	$\hat{e} = 4.1$, $p < .001$
25–34	25.5	18.7	
35–44	25.5	16.8	
45–54	20.2	13.1	
55–64	11.7	1.9	$\hat{e} = 2.8$, $p = .002$
≥65	4.3	0.0	$\hat{e} = 2.2$, $p = .01$

Note. [a]Significant adjusted standardised residuals ($\hat{e}$) are shown.

The post hoc tests showed that the Internet sample had significantly fewer participants aged ≤ 24 years but significantly more participants aged ≥ 55 years than the student condition.

Equivalence testing of outcome variables

To demonstrate the potential utility of formal tests of equivalence, prior to reporting the results of the formal tests of equivalence, results are provided from analyses based on $2 \times 2 \times 2$ analysis of variance in which the survey version (Internet or paper and pencil) was added as a third independent variable. In instances where a researcher may not be aware of equivalence testing, it could be anticipated that this analysis (i.e., $2 \times 2 \times 2$ ANOVA) would be the most likely approach selected given the research design and hypotheses posed. The NHST analyses results reported are based upon immediate post-exposure attitudes and intentions only (as opposed to pre-exposure measures).

Analyses based on NHST techniques. For both attitudes and intentions, no significant main effects of survey version, attitude, $F(1,195) = 0.68$, $p = .411$, $\eta_p^2 = .003$; intention, $F(1,196) = 1.78$, $p = .183$, $\eta_p^2 = .009$, or two-way effects involving survey version and appeal type, attitude, $F(1,195) = 0.77$, $p = .380$, $\eta_p^2 = .004$; intention, $F(1,196) = 2.55$, $p = .112$, $\eta_p^2 = .013$) or three-way effects involving survey version, appeal type, and response efficacy, attitude, $\Lambda = .99$, $F(1,195) = 1.94$, $p = .166$, $\eta_p^2 = .010$; intention, $\Lambda = .99$, $F(1,196) = 0.28$, $p = .601$, $\eta_p^2 = .001$, were found. This finding indicates that the version of survey did not differentially influence the immediate post-exposure results obtained.

Analyses based on formal tests of equivalence. Mean attitudinal and intentional scores were computed for the study's full sample ($N = 201$) and the results are shown in Table II. Table II shows that a significant result (using the larger p) was found for all three variables, indicating that the mean scores for pre-exposure attitudes, and post-exposure attitudes and intentions were equivalent for the paper-and-pencil and Internet conditions. Similarly, when examining the confidence intervals, for all three variables the upper and lower confidence intervals fall within the equivalence interval that was established a priori. The results of the equivalence tests conducted between the Internet and paper-and-pencil conditions according to the cells of the original study's 2×2 experimental design are reported in Table III.

As shown in Table III, for post-exposure attitudes towards drink driving, the p are significant for each cell of the experimental design (i.e., positive/high response efficacy, positive/low response efficacy, negative/high response efficacy, and negative/low response efficacy), indicating that the mean attitudinal scores were equivalent for the two conditions. Additionally, for all cells, the specified upper and lower confidence intervals fall within each relevant equivalence interval that was established a priori.

For post-exposure intentions, the results show that the p scores were significant for both the negative/high response efficacy and negative/low response efficacy advertisements, indicating that the Internet and paper-and-pencil versions had equivalent mean intention scores for the negative advertisements. In contrast, however, the results show that the p were not significant for the positive/high response efficacy or the positive/low response efficacy advertisement, indicating that the two conditions' mean intentional scores were not equivalent for the positive condition. Moreover, the upper confidence interval associated with each comparison exceeds the relevant equivalence interval.

Discussion

One aim of the current study was to determine whether the data obtained from an Internet-based survey would yield equivalent data to a more traditional paper-and-pencil version in an applied psychological research context. Specifically, the study explored whether the responses from a sample of participants recruited via the Internet who completed an Internet survey were equivalent to the responses derived from a sample of university undergraduate student participants who completed a paper-and-pencil version. Generally, the results were consistent with expectations, with the majority of mean comparisons between the two conditions

Table II. Mean differences in ratings of the complete sample ($N = 201$) by survey condition

	Paper-and-pencil ($n = 107$)		Internet ($n = 94$)		Difference					90%CI	
Variable	*M*	*SD*	*M*	*SD*	*M*	*SE*	Equivalence criterion[a]	*z*	*p*[b]	Lower	Upper
Pre-exposure attitude	4.64	1.48	4.95	1.49	−0.31	0.31	±0.93	1.98	.020*	−0.82	0.20
Post-exposure attitude	4.92	1.46	5.18	1.48	−0.26	0.31	±0.98	2.36	.009*	−0.74	0.28
Post-exposure intention	5.63	1.56	5.40	1.73	0.23	0.38	±1.13	−2.36	.009*	−0.40	0.86

Notes. CI = confidence interval.
[a]Equivalence criterion equals ±20% of paper-and-pencil group mean. [b]The largest *p* derived from the smallest difference between the ±EC and $\text{Mean}_1 - \text{Mean}_2$ is shown (Rogers et al., 1993).
*Significant result whereby the means of the two groups are statistically equivalent.

found to be equivalent. The only exception was in the positive appeal condition of the study, in which equivalent mean post-exposure intentions scores were not found. This result was found for both the positive, low response efficacy advertisement as well as the positive, high response efficacy advertisement and, thus, is indicative of an issue specific to the positive condition overall. Inspection of the means indicates that the mean intentional score was lower in the Internet condition than the paper-and-pencil condition for both advertisements.

This finding may have potential implications for how positive, or, more specifically, humorous road safety appeals are tested. This suggestion is underpinned by the notion that aspects of the respective testing environments of the Internet and paper-and-pencil conditions may have influenced the results. For instance, given that the paper-and-pencil survey was administered in group settings, there was a tendency for the positive advertisements to receive overt emotional responses from participants such as laughter. Consequently, such overt responses may have affected others' responses to the study's items by leading them to presume that others had formed favourable impressions of the advertisement(s). With interest increasing in the potential role that positive emotional approaches may play in health advertising increasing (e.g., Lewis et al., in press), understanding the most valid way of assessing individuals' responses to such appeals becomes most important. Moreover, the notion that the respective testing environments of paper-and-pencil and Internet-based surveys may influence the results may have implications for survey-based, psychological research more broadly, given that many paper-and-pencil surveys are conducted in group settings. A key endeavour for future research may be to identify topics that are suitable for survey testing in groups via paper-and-pencil surveys (i.e., where the impact of others is likely to be minimal), and those topics perhaps better tested in private settings (e.g., as in the case of some Internet-based surveys) so as to minimise the impact of other participants.

Demographic characteristics

The second aim of the current study was to provide empirical comparisons of the age and gender of the two samples. As expected, significant differences were found between the two samples. Generally, the Internet sample appeared to be more diverse than the student sample in relation to age. Additionally, for gender, the Internet sample provided a more equitable representation of male to female than the student sample. While it is interesting to note how the demographic characteristics of the Internet and student samples differed, arguably, the more significant issue is the extent to which the samples are representative of the general population.

Of the two samples, the Internet sample appeared more representative of the Australian population in relation to age and gender. According to the ABS (2006), the median age of Australia's population was 36.6 years. Although the categorical nature of the measure used in the current study prevents determination of the exact median age for each sample, inspection of the data in Table I indicates that the student sample's median would be much lower than the national level, with essentially half of the sample aged ≤ 24 years. In contrast, the Internet sample is much closer to the national level, with the median age situated within the 35–44 years category. Neither sample included the same proportion of people aged ≥ 65 years as within the national population (i.e., 13%; ABS, 2006); but, because the student sample included no persons within that category and the Internet sample included some respondents within this age category (i.e., 4.3%), the latter sample may be regarded as more diverse. In relation

Table III. Mean differences in ratings by appeal type, level of response efficacy, and version of the survey

Variable	Paper-and-Pencil $n = 60$ (Positive) $n = 47$ (Negative)		Internet $n = 42$ (Positive) $n = 52$ (Negative)		Difference				90%CI	
	M	*SD*	*M*	*SD*	*M*	*SE*	Equivalence Criterion[a]	*z*	p^{b}	Lower	Upper
Post-exposure attitude											
Positive, Low Response Efficacy	4.60	1.43	4.89	1.46	−0.29	0.29	±0.92	2.17	.015*	−0.77	0.19
Positive, High Response Efficacy	4.47	1.43	4.88	1.39	−0.41	0.28	±0.89	1.70	.045*	−0.87	0.05
Negative, Low Response Efficacy	5.37	1.45	5.43	1.53	−0.06	0.45	±1.07	2.25	.012*	−0.80	0.80
Negative, High Response Efficacy	5.48	1.48[c]	5.40	1.54	0.08	0.26	±1.10	−3.91	<.001*	−0.35	0.51
Intention											
Positive, Low Response Efficacy	5.37	1.88	4.60	1.85	0.77	0.38	±1.07	−0.81	.209	0.07	1.40
Positive, High Response Efficacy	5.62	1.85	5.05	2.07	0.57	0.39	±1.12	−1.42	.078	−0.07	1.21
Negative, Low Response Efficacy	5.85	1.40	5.96	1.61	−0.11	0.46	±1.17	2.30	.011*	−0.57	0.65
Negative, High Response Efficacy	5.74	1.64[d]	5.79	1.76	−0.05	0.59	±1.15	1.86	.031*	−1.02	0.92

Notes. CI = confidence interval.
[a]Equivalence criterion equals ± 20% of paper-and-pencil group mean. [b]Largest *p* derived from the smallest difference between the ±EC and $Mean_1 - Mean_2$ is shown (Rogers et al., 1993). [c,d]$n = 45$ and 46, respectively.
*Significant result whereby the means of the two groups are statistically equivalent.

to gender, the equitable split of male to female in the Internet sample is more representative of the Australian population, which has been reported as being a ratio of 98.8 male per 100 female individuals (ABS, 2006), than the unequal gender distribution in the student sample.

The current study, in finding evidence of greater diversity in the Internet sample than the student sample in terms of the sociodemographic variables assessed, is consistent with previous research (e.g., Gosling et al., 2004; Smith & Leigh, 1997). While it is acknowledged that participants in the Internet sample were self-selected and cannot be considered a true random or representative sample of the general driving population, the sample of drivers recruited were more diverse and representative of the general population (as based on the comparisons with the ABS data) than those recruited via a more traditional university student sample of drivers. These findings are encouraging and highlight that as Internet use increases and the characteristics of Internet users broaden, the representativeness of Internet samples is likely to continue to improve (Rhodes et al., 2003). It is also important to note, however, that the results highlight not the overall shortcoming of the paper-and-pencil technique per se but the problems associated with sampling from undergraduate psychology classes more generally.

The biases and subsequent problems with generalising from convenient student samples have been long acknowledged (Gosling et al., 2004; Sears, 1986). It has been argued that the popularity of student samples, despite their inherent problems, may be due to the lack of a practical alternative (Gosling et al., 2004). While Internet samples also represent convenience samples of the population, there is a growing body of evidence confirming their diversity relative to student samples. In addition, the many advantages that Internet samples offer may see them become an alternative for psychological research (Gosling et al., 2004).

Of particular note for road safety research is the inclusion of a greater number of men in the Internet sample relative to the student sample. In the context of road safety research, given that men are at a higher risk of being injured or killed on the roads relative to women (Tay, 1999, 2002), there is a need to ensure that this demographic is well-represented within studies that evaluate the effectiveness of particular countermeasures. For certain topics in road safety research the Internet may prove to be an effective means to reach such road users. Furthermore, for researchers examining the effectiveness of advertising in this context, as well as other health topics more generally, the Internet may

become the preferred means to conduct such studies given that radio or television advertisements can be added easily within an Internet survey as stimulus materials.

Equivalence testing

Different results were obtained via the NHST technique utilised and the formal tests of equivalence. As noted previously, the $2 \times 2 \times 2$ ANOVA in which the survey version was entered as a third independent variable was conducted because it was considered the most likely approach that would be utilised without awareness of the existence of formal tests of equivalence. Specifically, although the $2 \times 2 \times 2$ ANOVA results indicated no significant effects (main or interactional effects) of survey version in relation to post-exposure attitudes or intentions, the formal tests of equivalence results indicated that the post-intentional scores in the positive appeal condition were not equivalent between the two survey groups. Finding contradiction between the results obtained via these two types of tests is consistent with previous evidence (Rogers et al., 1993) and highlights the point argued throughout the current paper that NHST techniques are not optimal tests of equivalence.

Conclusion

The current study has addressed some significant omissions in the extant literature relating to the validity of Internet research. Specifically, it has provided an empirical comparison of the sample characteristics and data obtained via Internet and paper-and-pencil approaches for an experimental study addressing health message persuasiveness. Further, the empirical comparison was based upon formal tests of equivalence and, thus, provides a more appropriate and accurate test of equivalence than similar previous empirical comparisons based upon NHST.

Overall, the results suggest that an Internet sample of drivers is more diverse and representative of the general population than a university-student sample of drivers. Additionally, the results indicated that the two samples of participants produce predominantly equivalent data. While it is acknowledged that Internet data are not free from methodological constraints, the results contribute to a growing body of evidence that highlights the feasibility of Internet-based research. During a time when response rates to all sampling methodologies are declining (Birnbaum, 2004; Madge & O'Connor, 2004), the current study's results suggest that Internet surveys may represent a valid, alternative means of accessing participants for psychological research and, in particular, for psychological experimental research that aims to evaluate the effectiveness of health messages. Continued empirical investigation is necessary to gain greater insight into the validity of Internet research for psychological research more broadly (Hewson, 2003). Once validity has been established for data collected on a broad range of psychological research topics and designs, researchers will be able to place greater confidence in the use of the Internet as a means to collect valid data.

References

American Psychological Association. (1986). *Guidelines for computer-based tests and interpretations*. Washington, DC: Author.

Australian Bureau of Statistics. (2004–05). *Household use of information technology*. Canberra: Author.

Australian Bureau of Statistics. (2005–06). *Household use of information technology*. Canberra: Author.

Australian Bureau of Statistics. (2006). *Population by age and sex, Australia*. Canberra: Author.

Anderson, S., & Hauck, W. W. (1983). A new procedure for testing equivalence in comparative bioavailability and other clinical trials. *Communications in Statistics: Theory and Methods, 12*, 2663–2692.

Birnbaum, M. H. (2004). Human research and data collection via the internet. *Annual Review of Psychology, 55*, 803–832.

Buchanan, T., & Smith, J. L. (1999). Research on the internet: Validation of a world-wide mediated personality scale. *Behavior Research Methods, Instruments, and Computers, 31,* 565–571.

Carlbring, P., Brunt, S., Bohman, S., Austin, D., Richards, J., Ost, L., et al. (2005). Internet vs. paper and pencil administration of questionnaires commonly used in panic/agoraphobia research. *Computers in Human Behavior, 23*(3) 1–14.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis for field settings*. Boston: Houghton Mifflin.

Cribbie, R. A., Gruman, J. A., & Arpin-Cribbie, C. A. (2004). Recommendations for applying tests of equivalence. *Journal of Clinical Psychology, 60,* 1–10.

Dixon, P. M., & Pechmann, J. H. K. (2005). A statistical test to show negligible trend. *Ecology, 86,* 1751–1756.

Eichstaedt, J. (2002). Measuring differences in preactivation on the internet: The content category superiority effect. *Experimental Psychology, 49,* 283–291.

Epstein, J., Klinkenberg, W. D., Wiley, D., & McKinley, L. (2001). Insuring sample equivalence across internet and paper-and-pencil assessments. *Computers in Human Behavior, 17*, 339–346.

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.

Haberman, S. J. (1978). *Analysis of qualitative data. Vol 1: Introductory topics*. New York: Academic.

Harrison, W., & Christie, R. (2004, November). *Using the internet as a survey medium: Lessons learnt from a mobility survey of young drivers*. Road Safety Research Policing and Education Conference, Perth, Australia.

Hewson, C. (2003). Conducting research on the internet. *Psychologist, 16,* 290–293.

Horswill, M. S., & Coster, M. E. (2001). User-controlled photographic animations, photograph-based questions, and questionnaires: Three instruments for measuring drivers' risk-taking behaviour on the Internet. *Behavior Research Methods, Instruments, and Computers, 33,* 46–58.

Iragüen, P., & de Dios Orútzar, J. (2004). Willingness-to-pay for reducing fatal accident risk in urban areas: An Internet-based Web page stated preference survey. *Accident Analysis and Prevention, 36*, 513–524.

Kypri, K., & Gallagher, S. J. (2003). Incentives to increase participation in an internet survey of alcohol use: A controlled experiment. *Alcohol and Alcoholism, 38,* 437–441.

Lewis, I., Watson, B., & White, K. M. (in press). An examination of message-relevant affect in road safety messages: Should road safety advertisements aim to make us feel good or bad? *Transportation Research Part F: Traffic Psychology and Behaviour.*

Madge, C., & O'Connor, H. (2004). *Exploring the internet as a medium for research: Web based questionnaires and on-line synchronous interviews*. Economic and Social Research Council (ERSC) Research Methods Programme, Working Paper No, 9, University of Manchester. Retrieved 2 February 2007, from http://www.ccsr.ac.uk/methods/publications/documents/WorkingPaper9.pdf

Meyerson, P., & Tryon, W. W. (2003). Validating internet research: A test of psychometric equivalence of internet and in-person samples. *Behavior Research Methods, Instruments, and Computers, 34,* 614–620.

Musch, J., & Reips, U. (2000). A brief history of web experimenting. In M. H.Birnbaum (Ed.), *Psychological experiments on the internet* (pp. 61–87). San Diego, CA: Academic.

Pasveer, K. A., & Ellard, J. H. (1998). The making of a personality inventory: Help from the WWW. *Behavior Research Methods, Instruments, and Computers, 30,* 309–313.

Pealer, L. N., Weiler, R. M., Pigg, R. M, Jr., Miller, D., & Dorman, S. (2001). The feasibility of a web-based surveillance system to collect health risk behaviour data from college students. *Health Education and Behavior, 28,* 547–559.

Perkins, G. H., & Yuan, H. (2001). A comparison of web-based and paper-and-pencil library satisfaction survey results. *College and Research Libraries, 62,* 369–377.

Rhodes, S. D., Bowie, D. A., Hergenrather, K. C. (2003). Collecting behavioural data using the world wide web: Considerations for researchers. *Journal of Epidemiology and Community Health, 51,* 68–73.

Rogers, J. L., Howard, K. I., & Vessey, J. T. (1993). Using significance tests to evaluate equivalence between two experimental groups. *Psychological Bulletin, 113,* 553–565.

Schuirmann, D. J. (1987). A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics, 15,* 657–680.

Schulenberg, S. W., & Yutrzenka, B. A. (1999). The equivalence of computerized and paper-and-pencil psychological instruments: Implications for measures of negative affect. *Behavior Research Methods, Instruments, and Computers, 31,* 315–321.

Seaman, M. A., & Serlin, R. C. (1998). Equivalence confidence intervals for two-group comparisons of means. *Psychological Methods, 3,* 403–411.

Sears, D. O. (1986). College sophomores in the laboratory: Influences of a narrow data base on social psychology's view of human nature. *Journal of Personality and Social Psychology, 51,* 515–530.

Skitka, L. J., & Sargis, E. G. (2006). The internet as psychological laboratory. *Annual Review of Psychology, 57,* 529–555.

Smith, M. A., & Leigh, B. (1997). Virtual subjects: Using the internet as an alternative source of subjects and research environment. *Behavior Research Methods, Instruments, and Computers, 29,* 496–505.

Streiner, D. L. (2003). Unicorns do exist: A tutorial on ''proving'' the null hypothesis. *Canadian Journal of Psychiatry, 48,* 756–761.

Tay, R. (1999). Effectiveness of the anti-drink driving advertising campaign in New Zealand. *Road and Transport Research, 8*(4), 3–15.

Tay, R. (2002). Exploring the effects of a road safety advertising campaign on the perceptions and intentions of the target and non-target audiences to drink and drive. *Traffic Injury Prevention, 3,* 195–200.

Tryon, W. W. (2001). Evaluating statistical difference, equivalence, and indeterminacy using inferential confidence intervals: An integrated alternative method of conducting null hypothesis statistical tests. *Psychological Methods, 6,* 371–386.

Whittier, D. K., Seeley, S., & St. Lawrence, J. S. (2004). A comparison of web- with paper-based surveys of gay and bisexual men who vacationed in a gay resort community. *AIDS Education and Prevention, 16,* 476–485.

Willis, S., & Tranter, B. (2006). Beyond the 'digital divide': Internet diffusion and inequality in Australia. *Journal of Sociology, 42,* 43–59.