Electronic Versus Paper Surveys:
Analysis of Potential Psychometric Biases

Scott McCoy
College of William & Mary
Scott.Mccoy@business.wm.edu

Peter V. Marks, Jr.
US Army Medical Department
Peter.Marks@CEN.AMEDD.ARMY.MIL

Christopher L. Carr
University of Pittsburgh
clcarr@katz.pitt.edu

Victor Mbarika
Louisiana State University
vmbarika@lsu.edu

Abstract

The Internet and World Wide Web are increasingly being used for survey distribution and administration in both academic and practitioner research. Little systematic research exists on the efficacy of these survey administration media or their potential psychometric effects. This paper reports on a study of the potential biasing effects of online versus paper surveys. We consider issues specifically related to the information systems research context and introduce psychometric issues of general interest to those considering testing for response stability between online and paper survey administration. The study provides an assessment of the psychometric differences between paper-and-pencil and online survey administration for the well-known Technology Acceptance Model (TAM) instrument. The results indicate that biasing effects can occur which significantly reduce the stability of an instrument across administration methods. We sound a cautionary note on the practice of placing any existing paper-and-pencil survey instrument on the web without consideration of the biasing effects of web-based survey administration.

1. Introduction

This research focuses on the existence of psychometric biases in the use of online surveys. These biases could result from environmental, technological, psychological, or other factors inherent in or exacerbated by online survey administration. The central research question formally stated is: What potentially biasing psychometric effects, if any, result from the use of web-based versus paper surveys?

2. Theoretical Foundations

Are the psychometric properties of the scores derived from paper and online surveys containing the same items identical? Can the online survey delivery/response technique be utilized by researchers without concern for possible differences in the resulting scores vis-à-vis the paper versions of the same instruments? Little academic research exists specifically relating to the use of online media (Internet/World Wide Web/computerized survey administration in general) for the collection of survey data. Work in this area has not clarified the issue to the extent necessary to allow researchers to utilize the World Wide Web as a survey data collection instrument with the same confidence they have in paper forms.

One study in which an Internet-mediated questionnaire was directly compared with a pencil-and-paper version is that of Smith & Leigh [25], using a questionnaire designed to assess the nature and frequency of sexual fantasies experienced by respondents [11]. Smith & Leigh compared responses of the two groups to 6 of the 28 questions, and found no significant differences in answers to the individual items. Smith & Leigh's study is an example of a self-report survey adapted for Internet administration, and as such it is representative of one of the main types of questionnaire-based studies currently appearing on the WWW.

The development of Internet (WWW)-based surveys is an extension of the rise of psychometric survey administration using stand-alone computers. Honaker & Fowler [15] give an account of the history and development of
computerized surveying and note that most computerized surveys are simply translations of traditional paper-and-pencil surveys, with questions being presented on screen and responded to via the keyboard (Meier [22] notes several prominent examples).

With the translation of surveys from paper to computer formats, the question arises of equivalence between forms [24]. Bartram & Bayliss [2] review research suggesting that computerized versions of traditional surveys are generally equivalent to their paper-and-pencil antecedents, given that the surveys are (a) not timed and (b) require some form of multiple or forced-choice response to textual items, a conclusion endorsed by Cohen et al. [5]. However, these authors also discuss instances in which computerized and traditional versions have not been found to be equivalent. Certain factors can combine to create differences.

2.1. Psychological Factors

Several studies [18; 19] have found increased levels of self-disclosure in computerized survey administration. Others [32] have found that for some measures different patterns of responses are seen using traditional and computerized surveys. It is possible that in some cases, levels of computer knowledge, computer anxiety, or specific computer-related attitudes might affect participants' responses. For instance, a survey might give different results for people who are not confident in the use of computers, especially if the construct being measured was in some way related to computer anxiety or computer acceptance. While the size of the influence these and other factors may exert is unclear, both Meier [22] and Cohen et al. [5] agree that equivalence of computerized and traditional versions of surveys cannot simply be assumed but must be demonstrated for each survey. The psychometric situation is perhaps best characterized by Cronbach [6]: “It seems that the conventional and computer versions of a survey do usually measure the same variables, but difficulty or reliability can easily change. Whether the computer version is ‘the same survey’ must be questioned with each instrument in turn psychologically” (p. 48). This recommendation applies especially to surveys administered via the World Wide Web.

Among the advantages of computerized surveys (non web-based) is the fact that they offer greater control over surveying conditions and thus greater objectivity than situations in which a human researcher administers the instrument [2]. In the self-administered, Web-based survey paradigm, the researcher may have little or no control over the conditions under which a survey is completed. Respondents completing the survey in, for example, a noisy computer lab will experience a different set of environmental stimuli or distractions to somebody completing the survey on a computer in their home or workplace. Certain non-stable attributes of the individual (such as mood state or fatigue), possibly exacerbated by the online survey situation, may also increase unexplained variance in responses, although these may also be present in paper survey administrations. Such factors are not only beyond the control of researchers, they are beyond their awareness as well. Factors possibly contributing to adverse psychological effects include the anonymity of participants and the psychological distance between themselves and the researcher. While evidence exists that anonymity might increase honesty and self-disclosure, some authors [16] place value on establishing rapport between survey takers and administrators.

2.2. Information Technology Factors

A Web-based survey is likely to be completed by people using different browser software packages, each differently configured, running on different computer platforms with different displays [21]. The presentation of the survey may thus be different for every participant. Bartram & Bayliss [2] suggest that there may be reasons to question the equivalence of surveys administered via different hardware and software platforms. Also dependent on hardware, software and network configuration is the speed (time lag) with which documents may be accessed. Slow connections could lead to time-delays, presenting a frustration factor not present with paper surveys.

2.3. Assessing Test-Retest Reliability and Psychometric Theory

Churchill [4] suggests that scale reliability is not a uni-dimensional concept and consists of two different components: stability and equivalency. The former represents temporal stability of a measure at two different points in time, while the latter is more focused on the internal consistency or internal homogeneity of the set of items forming the scale. This study will
focus on both measures of internal consistency and stability by using a test-retest method as a way to examine the stability of a commonly used model in IS research. The technology acceptance model (TAM) was used across instrument administration methods (paper vs. web-based). This model has been widely used to investigate technology acceptance decisions.

The rationale for using a test-retest approach is that TAM involves the assessment of beliefs, attitudes and intentions that are assumed to be stable over a short period of time [12]. Adequate sample size should be able to even out the true changes in these psychological variables. Nunnally [23] suggests that, if an instrument is indeed stable, the two administrations should have a resulting correlation of .80 or higher. Nunnally [23] asserts that “if a scale does not even correlate with itself (in this case, the TAM scales with two forms of administration in this study) when administered on two occasions, it is hopeless to seek other evidence of reliability and hopeless to employ the test in correlation studies.”

Three potential problems must be addressed when applying the test-retest methodology: recall, time, and reactivity [23]. A recall problem may arise when the retest interval is too short. Carry-over effects may dominate the response pattern. Subjects may recall their first responses at the second administration of the measure. Similarly, a time problem may arise if the subjects are tested within too long an interval. True changes in the subjects concerned, rather than inconsistencies in the instrument, lead to differing test-retest results. Lastly, a problem with reactivity can occur when subjects are administered the instrument multiple times since they become sensitized to the instrument and 'learn' to answer as they perceive they are expected to respond.

Nunnally [23] recommends that a period of two weeks to one month should elapse between test and retest administrations in order to minimize the memory effect. Kline [17] suggests, however, that the two tests should be separated by at least a three-month gap and the samples should comprise at least 100 subjects from a large and representative sample of the population for whom the test is intended. In measuring end-user satisfaction and involvement, the intervals between test and retest administrations used by Torkzadeh & Doll [28; 29] were two hours and two weeks, respectively, with a sample size of 41 subjects, while Hendrickson et al. [14] only obtained 22 subjects for the test-retest administrations at two-week intervals over two years. Hendrickson et al. [14] used a three-day interval to measure the stability of two of the TAM scales (perceived usefulness and perceived ease of use) based on a sample of 51 subjects for a spreadsheet package and 72 subjects for a database management package.

The goal of this study is to assess any differences in responses to two identically worded survey instruments, one paper-and-pencil and one web-based. For the purposes of this study the test-retest reliability assessment procedure, well-known in the psychometrics literature, will be adopted [23]. This procedure involves correlating the scores from one administration of an instrument with those from a subsequent administration of the same instrument to the same respondents. Based on the above discussion, the fact that both the paper surveys and the web-based surveys were self-administered without controls (not-atypical for TAM research), and the relative psychometric stability TAM has exhibited in the past, we offer the following null hypothesis:

Provisional H1: No significant differences will be found between the data collected from the paper and online versions of the TAM survey instrument.

3. Research Methodology

3.1. Instrument

The TAM model has been used in many information systems studies [1; 3; 7; 8; 9; 10; 13; 20; 26; 27; 30; 31] and has proven successful in determining technology acceptance. It is a simple model consisting of two exogenous scaled constructs: Perceived Usefulness (PU), Perceived Ease of Use (PEU) and two endogenous scaled constructs: Attitude (ATT) and Behavioral Intention (BI) eventually predicting use of a technology. A slightly simplified representation of the TAM model is shown in Figure 1.
3.2. Respondents and Procedure

The TAM instrument, assessing the use of an online teaching tool, was first administered in electronic form to upper-division undergraduate students at a large northeastern university. An announcement was made in four sections of an undergraduate MIS course about the research study and a request was made to complete the survey. The URL was made available to all students and they were asked to complete the survey before the next class meeting. This course makes extensive use of the web and students have access to the Internet in several university sponsored computer labs. Of the 200 students requested to participate, 140 accessed the website and completed the survey. This resulted in a usable response rate of 70%.

Three weeks after the administration the same survey was administered in paper form to the same set of students. This time interval was based on Nunnally’s [23] recommendation of between two weeks and one month. Care was taken to ensure that the two survey tools were identical, with the only difference being the medium – paper or web. The paper surveys were handed out in class with the expectation that they be returned at the next class meeting. The second survey administration resulted in 168 responses. This resulted in a usable response rate of 84%. The number of respondents completing both the electronic and paper versions of the survey was 120. Only those respondents that completed both administrations were used in the analysis.

4. Data Analysis and Discussion

Following the lead of Hendrickson et al. [14] in analysis of the TAM scale data, we began by conducting paired t-tests for each item across administrations using SPSS. The results appear in Table 1. Clearly all but three of the raw scores for the TAM scale items exhibit highly significant differences across administration methods. This indicates that the individual responses were universally and significantly lower on the web-based survey than when respondents filled-out paper and pencil surveys. This represents a significant departure from the results reported by Hendrickson et al. [14]. They found only two significant differences in the paired t-tests between the first and second administration of the perceived usefulness and perceived ease of use scale items. Hendrickson et al. [14] did not measure behavioral intention or attitude. Our t-test results indicate a systematic upward bias in the responses from web-based to paper-and-pencil instruments.

We next examined the Pearson correlations for each item across administration methods. As shown in Table 1, no correlations approached the .8 threshold suggested by Nunnally [23]. None of the correlations exceeded the .71 level, a level at which the variance in the scores on one item and one administration method would have a fifty-fifty chance of predicting the magnitude and direction of the variance in the scores on the same item using a different administration.
method. Only four out of sixteen correlations exceeded the .60 level. 75% of the correlations reported by Hendrickson et al. [14] exceeded the .60 level. Over 50% of their correlations exceeded the .71 level. Thus, at the item level, the test-retest results for our cross-administration method study (paper-and-pencil vs. web-based) exhibit a much lower level of reliability than two paper-and-pencil administrations of the TAM scales.

Table 1: Paired Sample t-tests and Cross-Administration Correlations

<table>
<thead>
<tr>
<th>Item-pair</th>
<th>Paired Differences Means</th>
<th>Sig. (2-tailed)</th>
<th>Significantly Different Online vs. Paper</th>
<th>Corr.</th>
<th>Sig.</th>
<th>Above Nunnally’s .8 Threshold?</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATT1 - ATT1_P</td>
<td>0.375</td>
<td>.000</td>
<td>YES</td>
<td>.619</td>
<td>.000 NO</td>
<td></td>
</tr>
<tr>
<td>ATT2 - ATT2_P</td>
<td>0.358</td>
<td>.003</td>
<td>YES</td>
<td>.518</td>
<td>.000 NO</td>
<td></td>
</tr>
<tr>
<td>ATT3 - ATT3_P</td>
<td>0.350</td>
<td>.006</td>
<td>YES</td>
<td>.555</td>
<td>.000 NO</td>
<td></td>
</tr>
<tr>
<td>ATT4 - ATT4_P</td>
<td>0.350</td>
<td>.008</td>
<td>YES</td>
<td>.583</td>
<td>.000 NO</td>
<td></td>
</tr>
<tr>
<td>B11 - B11_P</td>
<td>0.667</td>
<td>.000</td>
<td>YES</td>
<td>.576</td>
<td>.000 NO</td>
<td></td>
</tr>
<tr>
<td>B12 - B12_P</td>
<td>0.517</td>
<td>.000</td>
<td>YES</td>
<td>.575</td>
<td>.000 NO</td>
<td></td>
</tr>
<tr>
<td>B13 - B13_P</td>
<td>0.550</td>
<td>.000</td>
<td>YES</td>
<td>.590</td>
<td>.000 NO</td>
<td></td>
</tr>
<tr>
<td>B14 - B14_P</td>
<td>0.633</td>
<td>.000</td>
<td>YES</td>
<td>.558</td>
<td>.000 NO</td>
<td></td>
</tr>
<tr>
<td>PU1 - PU1_P</td>
<td>0.433</td>
<td>.002</td>
<td>YES</td>
<td>.567</td>
<td>.000 NO</td>
<td></td>
</tr>
<tr>
<td>PU2 - PU2_P</td>
<td>0.383</td>
<td>.006</td>
<td>YES</td>
<td>.596</td>
<td>.000 NO</td>
<td></td>
</tr>
<tr>
<td>PU3 - PU3_P</td>
<td>0.450</td>
<td>.000</td>
<td>YES</td>
<td>.626</td>
<td>.000 NO</td>
<td></td>
</tr>
<tr>
<td>PU4 - PU4_P</td>
<td>0.508</td>
<td>.000</td>
<td>YES</td>
<td>.699</td>
<td>.000 NO</td>
<td></td>
</tr>
<tr>
<td>PEU1 - PEU1_P</td>
<td>0.283</td>
<td>.029</td>
<td>YES</td>
<td>.545</td>
<td>.000 NO</td>
<td></td>
</tr>
<tr>
<td>PEU2 - PEU2_P</td>
<td>0.233</td>
<td>.062</td>
<td>NO</td>
<td>.498</td>
<td>.000 NO</td>
<td></td>
</tr>
<tr>
<td>PEU3 - PEU3_P</td>
<td>0.033</td>
<td>.766</td>
<td>NO</td>
<td>.659</td>
<td>.000 NO</td>
<td></td>
</tr>
<tr>
<td>PEU4 - PEU4_P</td>
<td>0.175</td>
<td>.235</td>
<td>NO</td>
<td>.463</td>
<td>.000 NO</td>
<td></td>
</tr>
</tbody>
</table>

At the scale level, the results are similarly poor for our cross-administration method study. The Cronbach’s Alpha, a measure of internal consistency, and the test-retest reliability, a measure of stability, are shown in Table 2, for each TAM scale. While the Alphas are high, indicating homogeneity of response for the TAM scales in one administration, the stability, is low. Certainly, none of the scales exhibit a stability above Nunnally’s .8 threshold. In fact, none of the scales exhibit stability above .71, indicating a probability below chance alone that the scale scores on the web-based version of the TAM instrument could predict the scores on the paper-and-pencil version.
Table 2: Internal Consistency Reliability and Test-Retest Reliability By Scale

<table>
<thead>
<tr>
<th>Construct</th>
<th>Alpha Paper</th>
<th>Alpha Online</th>
<th>Test-Retest Reliability</th>
<th>Above Nunnally’s .8 Threshold?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Attitude</td>
<td>.9448</td>
<td>.9057</td>
<td>.660 (p&lt;.01)</td>
<td>NO</td>
</tr>
<tr>
<td>Behavioral Intention</td>
<td>.9469</td>
<td>.9442</td>
<td>.628 (p&lt;.01)</td>
<td>NO</td>
</tr>
<tr>
<td>Perceived Usefulness</td>
<td>.9535</td>
<td>.9379</td>
<td>.680 (p&lt;.01)</td>
<td>NO</td>
</tr>
<tr>
<td>Perceived Ease of Use</td>
<td>.8702</td>
<td>.9204</td>
<td>.633 (p&lt;.01)</td>
<td>NO</td>
</tr>
</tbody>
</table>

5. Conclusions, Limitations, and Directions for Future Research

5.1. Conclusions

The Internet and World Wide Web are increasingly being used for survey distribution and administration in both academic and practitioner research. However, little research exists on the efficacy of these survey administration media or their potential psychometric effects. This paper outlined possible biasing effects of online surveys and compared an online and paper-based survey within a group of respondents. Clearly our provisional hypothesis of no difference between the web-based and paper-and-pencil form of the TAM instrument was not supported. The results demonstrate, at least for this sample, that the responses to paper and online surveys differ significantly. This is by no means a definitive study; however, it raises a cautionary red flag on both the instability of TAM and the possible instability of all paper-and-pencil surveys when they are migrated to the web. Clearly, more research is needed to assess the potential psychometric biases that exist in online survey research.

Our results echo the suggestions of others, that each online survey must include reliability and validity assessment, even if the survey has been used many times before. The TAM instrument is widely used and accepted. However, based on this research, it is possible that TAM has significant reliability (stability) problems depending upon the method of administration (web-based vs. paper-and-pencil) and this should be examined carefully in future research. Administering a technology acceptance instrument with technology may, by itself, bias responses.

5.2. Limitations

The major limitation in this study is the order of the survey administration. To adequately test the potential biases, the subjects should be split into two groups, one receiving the online version first and one receiving the paper version first. This would allow the researcher to test if receiving one version first biases the second administration. This limitation should be addressed in future research.

5.3. Directions for Future Research

A large stock of well-known and useful survey instruments exists in the information systems research field as well as other business disciplines and the behavioral sciences in general. In the coming years, many researchers will turn to web-based administration of survey instruments. In each case, the validity and reliability of the instruments cannot be assumed, but must be tested and re-proven before valid interpretations can be made from the data they are used to collect. More research is needed in the area of web-based vs. paper and pencil surveys. More valid experimental designs must be developed for future studies to address the limitations outlined in this paper. This includes the development of completely randomized designs and carefully constructed treatments for time and type of administration method. Only through careful studies, testing the validity and reliability of each of the previously paper-and-pencil instruments in the new, web-based administration context can we gain a clear picture of the extent to which the new media
compares to the old. The effects of various context variables of interest, such as response type, e.g. Likert, semantic differential, Guttman, and number of questions/page, scrolling vs. button clicking, font-size, color, images, and others must be examined and their effects on response time, response bias, method variance, and other outcomes of interest measured and understood. Such a research program is necessary, but will take much time and effort to lead to useful knowledge of the constraints and parameters that affect responses to web-based survey instruments.

6. References


