Scientific Research Publishing

Open Journal of Social Sciences, 2017, 5, 69-86
http://www.scirp.org/journal/jss
ISSN Online: 2327-5960
ISSN Print: 2327-5952

Does Internet-Based Survey Have More Stable and Unbiased Results than Paper-and-Pencil Survey?

Pei-Wen Liao[1], Jun-Yi Hsieh[2]

[1]Department of Human Resource Management and Development, Hsiuping University of Science & Technology, Taiwan
[2]Department of Social and Public Affairs, University of Taipei, Taiwan
Email: pearlliao@hust.edu.tw, strategic60@hotmail.com

How to cite this paper: Liao, P.-W. and Hsieh, J.-Y. (2017) Does Internet-Based Survey Have More Stable and Unbiased Results than Paper-and-Pencil Survey? *Open Journal of Social Sciences*, **5**, 69-86.
http://dx.doi.org/10.4236/jss.2017.51006

Received: December 7, 2016
Accepted: January 10, 2017
Published: January 13, 2017

Abstract

This study designed to compare responses from an internet-based survey to those from a paper-and-pencil survey in terms of measurement reliability, validity, and equivalence using homogeneous demographic profiles of the undergraduates studying in Taiwan. Several similarities and differences were found between two types of survey in this study. For examples, contents of the survey items (*i.e.*, internet-related vs. behavior-related) and survey environments significantly influence the distribution of responses. The normal distribution, internal consistency, in addition to construct, and convergent validity for individual construct are quite similar. However, the homological validity evidence was demonstrated through structural equation modeling across two survey modes. Implications and future research are also discussed.

Keywords

Internet-Based Survey, Paper-and-Pencil Survey, Measurement Reliability, Validity, Equivalence

1. Introduction

Internet data collection is becoming increasingly popular in all research fields studying human perceptions, behaviors and opinions. A key methodological assumption underlying use of the same survey instrument is that data from internet-based and paper-and-pencil modes have been reported as producing equivalent results and that these have been compared meaningfully [1]. As well as, it is expected that the internet-based survey will yield equivalent mean ratings as the paper-and-pencil survey on the same research measures. However, measurement validity, reliability, and equivalence of internet gathered data must be established,

in comparison to the usual paper-and-pencil survey, before an inferential analysis can be done [2]. The lack of evidence of measurement equivalence may weaken or bias conclusions because the findings may be highly doubtful. In this case indicated observed mean differences on relevant constructs across survey methods might result from measurement artifacts related to the measurement instrument used rather than from true differences across methods [3].

The previous findings from comparative research on two survey methods were inconsistent [4]. Some researches intend to understand their similarity and difference. However, they did not control sample heterogeneity which may lack the base of comparison [5]. Although other research found that the contents of survey items related to survey methods may influence the results [6]. The subject of survey items was ignored by their research. If the subject of the survey item is "the internet", the respondents may tend to have higher scores in the survey environment of the internet than the paper-and-pencil. As well, results of previous studies suggested that internet-based surveys produce data that is at least as reliable, valid, and of equal quality as data obtained more than via paper-and-pencil survey method [7]. Nevertheless, many of these studies had methodological limitations such as using non-equivalent comparison groups and inappropriate data analytic strategies [8]. Therefore, results may not fully explain reliability and validity of the theoretical framework. As such, more research is needed a meaningful substantive understanding of the results for two survey methods of data collection should be explored [9]. Consequently, the current study seeks to determine measurement reliability, validity, and equivalence of data derived from the two survey methods with identical items, an additional objective of the study is to discuss the homological validity of equivalence demonstrated empirically in the same analytical framework.

The current research does not focus on hypothesis testing of the theoretical framework. Rather this research aims to yield a better understanding and a more comprehensive and intriguing picture of ways, which contents of survey items and survey methods produce similar measurement reliability, validity, and equivalence, holding on the same theoretical framework and the homogeneous samples. Therefore based on the previous research, face validity and content validity of the survey items, the internet-related constructs (internet learning, internet behavior, and internet social relationship) [10] and behavior-related constructs (relationship with family, relationship with friend, relationship with classmate, and relationship with student-school) [11] were selected for analysis in this research. This research selected the undergraduates with the similar backgrounds studying in Taipei City as our research samples.

This research is expected to advance knowledge in this area in several ways. First, it presents a brief description of the literature reviews between two survey methods for internet-based and paper-and-pencil survey. Second, results were presented by their measurement reliability (e.g., internal consistency), validity (e.g., construct validity, convergent validity), and equivalence (e.g., homological validity) for the responses elicited using paper-and-pencil and internet-based

modes. Thirdly, the empirical results are then presented and compared in our proposed analytical framework, along with descriptive statistics, Cronbach's alpha, item-total correlation, t-test, and multiple-group structural equation modeling (SEM), establishing measurement equivalence between two survey methods. Finally, limitations and conclusions are presented and implications are made for further research.

2. Literature Review

2.1. The Advantages and Disadvantages of Internet-Based and Paper-and-Pencil Survey

In recent years, an increasing number of researchers have started to use the internet to collect data [12]. Internet-based surveying is believed to have several advantages over the more conventional paper-and-pencil [1] [13] [14]. Internet-based surveys are less costly, less time, design flexibility, lead to faster survey responses, allow for greater flexibility in survey design, and offer a wider variety of response formats [15]. In addition, internet-based surveys have a wider geographical reach than telephone and mail surveys [8], are less sensitive to order of question effects due to the ease of randomizing questions [9].

Internet-based surveys have some disadvantages, example, limited sampling and respondent availability and no interviewer that lack of a trained interviewer to clarify and probe can possibly lead to less reliable data. Other research on internet surveying suggests that self-administered computerized surveys yield more honest results as social desirability does not factor into the respondent's answers [16]. As well, it has ability to acquire large and diverse samples, 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 [2] [17] [18]. However, internet surveying has drawbacks, including higher non-response rates [19], potential technological problems, decreases in item reliability, the possibility of multiple submissions, and insufficient coverage of all occupational groups represented in the organization being studied [9]. Some of the problems can be effectively ruled out, for examples the possibility of multiple submissions can be controlled effectively. The problem of insufficient coverage of all occupational groups is likely highly dependent on the organizational context. More fundamentally, it is an issue of variable response propensities among different groups, which is a problem with any survey method.

Mode effects, also a form bias, present another challenge to survey research. According to [13], using mixed modes (such as both Internet and paper versions) within a single administration can be problematic in that different modes may impact responses. For example, the result from the two modes cannot be equated because of differences in who responds (*i.e.*, nonresponse bias) and how they respond (*i.e.*, response bias) [1].

2.2. Measurement Reliability, Validity, and Equivalence

Measurement reliability is an assessment of the degree of consistency between multiple measurements variable. A more commonly used measure of reliability is internal consistency, which applies to the consistency among the variables in a summated scale. The summated scale should be analyzed for reliability to ensure its appropriateness before proceeding to an assessment of its validity. Measurement validity is the extent to which a scale or set of measures accurately represents the concept of interest. In addition to face or content validity, the three most widely accepted forms of validity are construct, convergent, and homological validity. Convergent validity assesses the degree to which two measures of the same concept are correlated. The estimates of convergent validity indicate that different measures of theoretically similar or overlapping constructs should be strongly interrelated. For example, show the convergent validity of test work skills, the scores on the test can be correlated with scores on other tests that are also designed to measure basic work ability. High correlations between the test scores would be evidence of a convergent validity [2]. Finally, homological validity refers to the degree that the summated scale makes accurate predictions of other concepts on a theoretically based model. The researcher must identify theoretically supported relationships from prior research or accepted principles and then assess whether the scale has corresponding relationships [20].

The number of different approaches highlights the definition of equivalence provided by the American Psychological Association within its Guidelines for Computer-Based Tests and Interpretations [21]. According to this definition, one aspect of determining equivalence between internet-based and paper-and-pencil versions is, "if the means, dispersions, and shapes of the score distributions are approximately the same". This definition is similar to that a test or a subscale is said to have measurement equivalence across groups or populations if respondents with identical scores on the underlying construct have the same expected raw score or true score at the item level, the subscale total score level, or both [22]. If a measure provides equivalent structural relations across two survey methods, it is natural to ask whether a measure's scores have comparable relations with other important variables [23]. Without measurement equivalence, it is difficult to interpret observed mean score differences meaningfully [24].

Previous findings were mixed with regard to levels of measurement equivalence in two survey methods. Donovan, acknowledged that the whiter-collar university employees who received the computerized instrument were distinctly different from the blue-collar employees who completed the paper-and-pencil instrument, and therefore, sample characteristics were confounded with modes of assessment [25]. In light of this limitation, they recommended that researchers obtain multiple samples from the same organization to permit an assessment of measurement equivalence. The researchers have noted this concern and similarly argued that measurement equivalence of internet-based versions of paper-and-pencil surveys cannot be assumed but must be empirically demonstrated

[26] [27] [28]. Other research found that different survey methods often produce different answers to the same questions [29] [30]. A lack of evidence of measure invariance weakens conclusions, because findings are open to alternative interpretations [31] [32] [33]. The importance of measurement equivalence is underscored when it is realized that if a theoretical framework vary across internet-based and paper-and pencil applications, as noted above, there would be no scientific basis for drawing meaningful inferences or establishing the generalizability of competing theories. Moreover, it is important to examine whether survey method affects structural relations among theoretical constructs comprising a homological validity.

3. Research Method

The related research is still needed to fully understand how different methods for presenting research stimuli influence survey responses [5]. As a result, failure to reject the null hypothesis is not the same as proving "no difference" or equivalence in theoretical framework [34] [35]. That is, observed mean score differences may reflect the true mean difference between the individuals as well as a difference in the relationship between the observed score that is not identical across individuals [34]. Therefore, establishing measurement reliability, validity, and equivalence enables us to answer a series of important key points that the current study addressed concerns [28]: 1) Differences in response styles across survey methods such as observed cross method differences in mean item scores; 2) Homogeneous respondents in different survey methods calibrate the intervals on the measurement scale used in similar ways; 3) The survey items asked on internet-based and paper-and-pencil versions of the same survey instrument yield measurement equivalence.

3.1. Analytical Framework and Hypotheses

Accordingly, the research hypothesis, *H*1: *There is no difference for measurement reliability and validity between internet-based and paper-and-pencil modes using the same survey instrument.* To address H1, the focus will be on the "difference for individual construct" between internet-based and paper-and-pencil modes. Further evidence is needed to answer the following questions. Do the styles of the question items systematically influence their responses distributions? Do the two modes of survey implementation have similar impacts on the variables of interest such as yielding the similar measurement reliability and validity? The second research hypothesis refers to the *H*2: *There is no difference on ratings of the analytical framework from respondents collected via internet-based mode, would be equivalent to ratings collected through a traditional paper-and-pencil site.*

To test the hypotheses, the non-recursive model which includes the reciprocal causal relations of internet behavior and internet social relationships, and their antecedents of internet-related (e.g., internet behavior, internet social relationship, internet learning) and behavior-related (e.g., relationship with family, rela-

tionship with friend, relationship with classmate, relationship with student-school) measures are defined. The rest of this paper moves beyond these theoretical and speculative arguments and will test the empirical evidence related to the expected differences or similarities in measurement reliability, validity, and equivalence for internet-based survey in comparison to paper-and pencil survey.

3.2. Data Collection

The internet-based survey and paper-and-pencil survey were conducted during the spring. The data sampling came from a study of undergraduates studying in one General University and one Technology College in Taipei City. The brief demographic overview of the respondents separately by internet-based survey and paper-and-pencil survey is provided in Table 1 (e.g., gender, age ranges, and education).

The internet-based survey samples, we emailed the internet-based survey and questionnaire webpage to the 1,000 undergraduates from the student. 341 college and university undergraduates completed surveys. The internet-based sample consisted of 42.52% females. The majority (66.57%) were between the ages of 21 and 24, and more than half (79.77%) had studied at a general university.

The paper-and-pencil survey samples, we firstly distributed about 1,000 paper-and-pencil survey instruments to the undergraduates from the student. The respondents with paper-and pencil survey were given a four-page questionnaire in class to complete. Because we should insure sample equivalence across internet and paper-and-pencil survey, we follow up to distribute the questionnaires to other students until two samples are equivalent. We got the final 341 samples which consist of 43.99% females. The majority (68.33%) were 21 to 24 years old, and more than half (75.66%) had studied at a general university.

Table 1. Sample characteristics of the respondents between internet-based and paper-and-pencil survey sample equivalence test.

Demographics	Internet-based		Paper-and -Pencil		Sample Equivalence Test	
	n	%	n	%	χ^2	P
Gender						
Female	145	42.52%	150	43.99%	0.15	0.70
Male	196	57.48%	191	56.01%		
Age						
18 to 20 years old	87	25.51%	80	23.46%		
21 to 24 years old	227	66.57%	233	68.33%	0.39	0.82
More than 25 years	27	7.92%	28	8.21%		
Affiliated Education						
General University	272	79.77%	258	75.66%	1.66	0.20
Technology College	69	20.23%	83	24.34%		

Although the two samples varied somewhat with regard to demographic characteristics, such differences were anticipated because of the university and college structure and the geographical dispersion of students. The responding rates may be low; however, the demographics of two-sample have no significant difference in gender ($\chi^2 = 0.15$, $p = 0.70$), age ($\chi^2 = 0.39$, $p = 0.82$), and affiliated education ($\chi^2 = 1.66$, $p = 0.20$) in terms of chi-square test shown in **Table 1**. Controlling such demographic homogeneity would systematically rule out some measurement errors, which overcome the potential sampling bias existing in the previous research that we discussed above.

3.3. Variables Measurement

All of the study variables in this study were assessed on a 7-point Likert scale (where 1 = strongly disagree and 7 = strongly agree). We chose constructs that are closely related to the survey mode this research is examining (e.g. internet-related and behavior-related measures). There are three construct total 15 items refer the Internet-related Measures, including Internet Learning, Internet Social Relationship, and Internet Behavior. Internet learning is defined in terms of what person learns about or learning from the internet, seven items are measured with this concept [36]. Internet social relationship refers most generally to a relationship between two or more people created by way of the internet; four items are measured with this concept [10]. Internet behavior refers to the type of material found on the internet, the frequency of internet use; do you mean email, social networking, research, four items are used to measure this concept [10].

There are four construct total 15 items refer the Behavior-related Measures, including Relationship with family, Relationship with Friend, Relationship with Classmate, and Relationship with Student-school [11]. Relationship with family is defined as a close association between two or more people with kinship or adopted behavior total four items developed. Relationship with friend is constructed as mutually cooperative and supportive behavior between two or more people; this measure includes four survey items used. Relationship with classmate refers to the connection, friends and acquaintances because of attending the same kindergarten, primary school, high school, college, and military service. This factor is measured with four items. Relationship with student-school is conceptualized as involvements with the place where he enrolls and studies at present. This factor is measured with three items.

3.4. Analytical Methods

This part consist of the estimated measures with means, standard deviations, normal distribution listed separately in internet-based (n = 341) and paper-and-pencil (n = 341) shown in **Table 2**, also presented the independent samples t-test was used to compare the responses' differences between two survey modes. The Cronbach's alpha and itemtest correlation shown in **Table 3** for present internal consistency. Item-total correlations are an assessment of convergent construct

Table 2. Mean, standard deviation, and T-values for each survey item of individual factor between internet-based and paper-and-pencil surveys.

Factor/Survey Items	Internet-Based		Paper-and-Pencil		T-Value
	Mean	SD	Mean	SD	
Internet Learning					
1. Internet helps to search the needed resource	5.82	0.94	6.07	0.95	−3.49***
2. Internet helps to collect certain information	5.85	0.78	5.96	0.99	−1.67
3. Internet provides contents, which satisfy your needs	5.55	0.93	5.60	1.05	−0.73
4. Internet provides the well contents	5.35	1.07	5.45	1.19	−1.09
5. Internet provides the latest contents	5.80	1.00	5.81	1.09	−0.11
6. It is easy to understand for the contents of the Internet	5.37	1.06	5.38	1.12	−0.18
7. Internet can promote my learning efficiency	5.27	1.25	5.33	1.21	−0.65
Internet Social Relationship					
8. I like to cheat with friends by Internet	5.43	1.24	5.50	1.36	−0.68
9. I like to exchange information with friends by Internet	5.55	0.93	5.58	1.17	−0.40
10. I enable to understand my friends' thinking by Internet	5.13	1.19	5.09	1.31	0.40
11. I freely express my thinking to friends by Internet	5.17	1.34	5.00	1.44	1.63
Internet Behavior					
12. Using the Internet is required to get information	5.65	0.91	5.80	1.04	−1.96*
13. Posting an article onto the BBS	4.72	1.41	3.75	1.57	8.43**
14. Learning about computer terminology	4.80	1.42	4.11	1.51	6.14***
15. My favorite recreation is the use of internet	5.31	1.19	5.10	1.39	2.19*
Relationship with Family					
16. I like spending time at home with my family	5.18	0.07	5.10	1.22	0.16
17. My family and I enjoy being together	5.14	1.16	5.14	1.23	−0.03
18. My family helps me with solving my problems	4.97	1.27	4.99	1.40	−0.23
19. My family likes to hear my thinking	4.83	1.32	4.89	1.37	−0.60
Relationship with Friend					
20. My friends are kind to me	5.52	0.98	5.74	1.01	−3.05**
21. I enjoy being with my friends	5.51	0.93	5.62	1.1	−1.50
22. My friends are excellent people	5.35	0.99	5.49	1.08	−1.77
23. My friends like to hear my thinking	5.27	0.95	5.37	1.07	−1.21
Relationship with Classmate					
24. My classmates are kind to me	5.17	1.11	5.19	1.17	−0.20
25. I enjoy being with my classmates	5.18	1.05	5.10	1.14	0.98
26. My classmates are excellent people	5.03	1.13	5.02	1.21	0.03
27. My classmates like to hear my thinking	4.99	1.05	4.88	1.15	1.32
Relationship with Student-School					
28. I like to be in school	4.86	1.27	4.44	1.48	3.97***
29. I learn many things at school	5.06	1.26	4.56	1.43	4.85***
30. I like my teachers	5.28	1.1	4.79	1.42	5.05***

Note: Wang (2007); Yilmaz and Türküm (2008).

Table 3. Mean standard deviation, skewness, and kurtosis for each study factor between internet-based and paper-and-pencil survey.

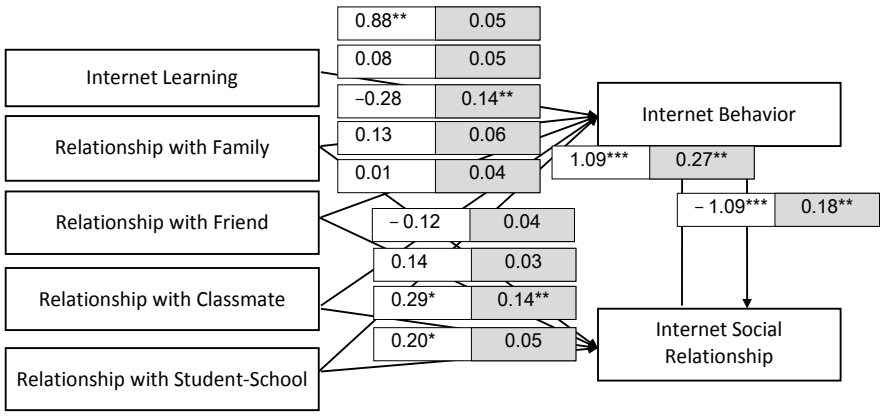
Measure	Internet-Based Survey				Paper-and-Pencil Survey				T
	Mean	SD	Skew	Kurt	Mean	SD	Skew	Kurt	
Internet Learning	5.57	0.74	−0.60	4.15	5.66	0.87	−0.69	4.00	−1.39
Internet Social Relationship	5.32	0.89	−0.50	3.73	5.29	1.08	−0.62	3.70	0.37
Internet Behavior	5.12	0.93	−0.51	3.92	4.69	0.99	−0.10	3.45	5.88***
Relationship with Family	5.01	1.08	−0.59	3.10	5.03	1.11	−0.23	2.79	−0.22
Relationship with Friend	5.41	0.81	−0.56	3.81	5.56	0.94	−0.65	4.13	−2.18*
Relationship with Classmate	5.09	0.95	−0.50	3.46	5.05	1.06	−2.30	3.09	0.58
Relationship with Student-School	5.07	1.03	−0.72	3.87	4.60	1.23	−0.47	3.58	5.42***

T = T-values, ***$p < 0.001$, **$p < 0.01$, *$p < 0.05$.

validity at the item level analysis. An item-total correlation test is performed to check if any item in the set of tests is inconsistent with the averaged behavior of the others, and thus can be discarded.

We tested the analytical framework in Figure 1 with multiple-group (SEM using AMOS 7.0. Since there are computational limitations for a structural equation analysis involving too many indicators, consistent with other researchers we used item parcels for each construct to reduce the number of indicators (e.g., [37] [38]). Specifically, we combined items with test scores by averaging them until we yielded the aggregated items composing individual construct.

Multiple-Group SEM allows to contrast the inequality or equality of two survey methods which to test measurement equivalence [20] [39]. The advantage of this method is that a wide variety of hypotheses about group differences and similarities can be tested. For example, this method is useful for testing the tenability of a series of increasingly restrictive models, using goodness-of-fit tests. To determine the degree of measurement equivalence, the fit constrained model can be compared to that of the unrestricted model without the equality constraints with the chi-square difference statistic [39]. The methods followed for the measurement equivalence analysis is detailed [31]. Their procedure is a set of hierarchical tests, where each subsequent test becomes increasingly restrictive. The first unconstrained model tests whether similar path structures are unequal across survey methods. The second structural weights model tests whether regression weights are equal across survey methods. The third structural covariance model tests whether the variances of the observed variables and the covariance between the observed variables are equal across survey methods. The fourth structural residuals model tests whether equal error variances are tenable are equal across survey methods. Then we compare the "unstandardized" coefficients across two survey methods samples [39]. We also evaluated model fit using the normed chi-square (NC) (χ^2/df), the goodness-of-fit index (GFI), the comparative fit index (CFI), the tucker-lewis index (TLI), and the root-mean square error of approximation (RMSEA) [40].



PS: The Internet-Based survey standardized effect size of gray bottom

Figure 1. Standardized structural effects for the analytical framework.

4. Findings

4.1. Responses Distribution and Mean Differences

Descriptive statistics and t-test are reported in Table 2, which show there was statistically significant difference for the comparable factors between the internet-based and paper-and-pencil samples with t-test. For example, the mean of all three survey terms measuring "relationship with student-school" was significantly larger in the internet-based survey than in the paper-and-pencil survey.

One possible reason is that the survey context influences the responses. The contents of survey items interacting with the survey environment mark a significant difference of the results on survey modes. The survey environment for the paper-and-pencil survey is "school," and the subject of the survey items measuring "relationship with student-school" is also "school." This confound condition may cause the key variables to shift perceptions from the respondents. This is also the case with some items of internet behavior factor-the mean of internet-based survey were significantly larger than that of paper-and-pencil survey, such as the items of "Posting an article onto the Bulletin Board System (BBS)", "Learning about computer terminology," and "One of my favorite pastimes is to go to the internet for recreation."

The subject of the survey items for "internet behavior" is "internet," this may lead to the respondents' high perceptions of internet-related questions in internet-based than those in paper-and-pencil survey. Further, in the paper-and-pencil survey, the mean of "my friends are kind to me" is larger than that in internet-based survey. One possible reason is that the paper-and-pencil respondents who together set up in grouping class answer the questionnaire, but the internet-based ones who independently use the computer in own room respond the survey. However, this is not a general rule; for example, in this research the mean of "internet helps to search the needed resource" is significantly smaller in internet-based than it is in the paper-and-pencil survey, one item of internet learning factor. It is the same with one internet-behavior item "using the Internet is required to get information". Although there may have some anomalies, the survey environment interacting with the contents of survey items, to a certain degree, control the perceptive responses.

4.2. Measurement Reliability and Validity

Cronbach's alpha in Table 3 the internal consistency reliabilities, a commonly-accepted rule of thumb is that $\alpha \geq 0.7$ indicates acceptable reliability, $\alpha \geq 0.8$ indicates good reliability, and $\alpha \geq 0.9$ indicates excellent reliability [41]. The reliabilities of all the study factors in the two survey methods are acceptable and nearly identical (*i.e.*, varies from 0.70 to 0.93). Then we tested the convergent properties of the items, by assessing the item-total correlation and comparing it to the correlation of the item to its own subscale. The advocated average item-total correlations of 0.30 or better as exemplary [42]. If a correlation value less than 0.2 or 0.3 indicates that the corresponding item does not correlate very well

with the scale overall, thus, it may be dropped [43]. The items with an item-total correlation was below 0.88 and above 0.30, the Cronbach α between 0.7 and 0.93, indicating that the evidences support convergent validity for this research and the survey items can represent that construct. We next use the item parcels as indicators of a conceptually defined construct. As can be seen in Table 3, all of the study constructs follow normal distributions with no serious skewness (Skew) and kurtosis (Kurt) across two survey modes. This indicates that there is no problematic summarizing scale for this survey instrument, which may yield unbiased statistical results for these factors. However, further evidence is needed to be confirmed as the following statistical procedure.

4.3. Measurement Equivalence

In terms of measurement equivalence test as mentioned above, if assuming unconstrained model to be correct, chi-square difference test in Table 4 suggests that the structural weights, structural covariance, and structural residuals models are significantly worse than the unconstrained model. The later model provides a better fit than the first three models because of its smaller χ^2 values in statistical significance. For example, the change of 171.25 in χ^2 from unconstrained model to structural residuals model exceeds the critical value of chi-square (41.34) with 28 degrees of freedom, the difference in the χ^2 is statistically significant ($P < 0.05$). For examples, unconstrained model and constrained model are hierarchical so that chi-square difference test can be applied to assess their relative fit. Unconstrained model: $\chi^2 = 146.03$, df = 6. Structural residuals model: $\chi^2 = 317.28$, df = 34. Chi-square difference: χ^2 diff = 171.25, df = 28, $p < 0.05$. Chi-square difference test suggests that the structural residuals model is significantly worse than the unconstrained model.

That is, the unconstrained model fits well more than the structural residuals model. Taken together, these findings lead to rejection of the hypothesis that the

Table 4. Comparison of model fit and structural relations indices for the unconstrained and constrained models between internet-based and paper-and-pencil survey modes.

Model	Unconstrained Model	Structural Weights	Structural Covariance	Structural Residuals	Threshold Value
Overall Model Fit	$Chi^2 = 146.03$ (df = 6)	$Chi^2 = 254.27$ (df = 17)	$Chi^2 = 309.72$ (df = 32)	$Chi^2 = 317.28$ (df = 34)	
Chi-square Test	assuming unconstrained model to be correct	108.240 (df = 11)	163.689 (df = 26)	171.251 (df = 28)	$P \leq 0.05$
GFI	0.95	0.91	0.89	0.89	≥0.90
IFI	0.91	0.84	0.81	0.81	≥0.90
CFI	0.90	0.84	0.81	0.81	≥0.95
NFI	0.90	0.83	0.79	0.79	≥0.90
RMSEA	0.07	0.14	0.11	0.11	≤0.08
RMR	0.08	0.10	0.12	0.12	≤0.08

analytical frameworks for internet-based and paper-and-pencil modes have the same structural weights, structural covariance, and structural residuals. Therefore, the following section will explain the results of the unconstrained model with the better model efficiency than other three models. The model fit indices shown in Table 4 suggest that the sample data have an acceptable fit to the SEM in the unconstrained model (χ^2 = 146.03, df = 6; GFI = 0.95; IFI = 0.91; CFI = 0.90; NFI = 0.90; RMSEA = 0.07; RMR = 0.08). All goodness-of-fit indices point to consistent model validation. Then we conclude that the unconstrained model represents a reasonable approximation to the data across internet-based and paper-and pencil survey.

Comparison of structural effects in the unconstrained model shown in Table 5 and Figure 1. We see that the unstandardized path coefficient, internet learning to internet behavior for internet-based survey 0.07 ($p \geq 0.05$) versus for paper-and-pencil survey is 1.10 ($p < 0.001$), which verifies much difference in results of their statistical significance.

In addition, the results with the path of internet social relationship to internet behavior in statistical significance between internet-based survey (unstandardized coefficient = 0.21, $p < 0.001$) and paper-and-pencil (unstandardized coefficient = 0.99, $p < 0.001$). As is the similar with the path effects of relationship with classmate to internet social relationship in internet-based survey (unstandardized coefficient = 0.16, $p < 0.01$) and paper-and-pencil survey (unstandardized coefficient = 0.29, $p < 0.05$). The effects of the above two paths are larger in the paper-and-pencil survey than those in internet-based survey. The non-recursive path coefficients, internet social relationship to internet behavior verify the difference in statistical significance. However, their path directions present to be negative across two survey modes.

5. Discussion and Conclusion

5.1. Discussion

The primary purpose of this study was to examine whether or not the internet-based and paper-and-pencil survey modes yield similar measurement reliability, validity, and equivalent results, given the same analytical framework, the same survey instrument, and the homogeneous samples constant. We first compared the groups of different survey modes with respect to demographic variables, including gender, age, and affiliated education. As expected, significant differences were found between the two samples.

We employ the given internet-related and behavior-related measures, which had been estimated by other published studies, to test the hypotheses. The findings indicate that two survey methods share the similar results for individual factor: (a) internal consistency reliabilities; (b) construct validity; (c) convergent validity. The research evidences supports H1. Previous findings also had empirical supports that two survey methods did not have marked differences in normal distributions, reliability and validity for the measures of transformation leadership [28], and self-esteem scale [44]. However, some research found differences

Table 5. Standardized and unstandardized structural effects for the analytical framework in unconstrained model.

Path	Internet-Based Model			Paper-and-Pencil Model		
	Stand.	Unstand.	Variances	Stand.	Unstand.	Variances
Internet Behavior ← Internet Learning						
	0.05	0.07 (0.04)	0.55	0.88	1.10*** (0.21)	0.75
Internet Behavior ← Internet Social Relationship						
	0.27	0.21*** (0.03)	1.00	1.09	0.99*** (0.14)	0.98
Internet Behavior ← Relationship with Family						
	0.05	0.05 (0.04)		0.08	0.07 (0.054)	
Internet Behavior ← Relationship with Friend						
	0.14	0.16** (0.06)		−0.28	−0.29** (0.10)	
Internet Behavior ← Relationship with Classmate						
	0.06	0.06 (0.05)		0.13	0.121 (0.07)	
Internet Behavior← Relationship with student-School						
	0.04	0.04 (0.05)		0.01	0.01 (0.05)	
Internet Social Relationship ← Internet Behavior						
	0.18	0.21*** (0.03)	0.65	−1.09	−1.19*** (0.368)	2.30
Internet Social Relationship ← Relationship with Family						
	0.04	0.07 (0.04)	1.17	−0.12	−0.11 (0.08)	1.23
Internet Social Relationship ← Relationship with Friend						
	0.03	0.05 (0.04)	0.66	0.14	0.16 (0.13)	0.87
Internet Social Relationship ← Relationship with Classmate						
	0.14	0.16** (0.06)	0.90	0.29	0.29* (0.13)	1.2
Internet Social Relationship ← Relationship with student						
	0.05	0.06 (0.05)	1.06	0.22	0.20* (0.08)	1.50

Note: Values in parentheses are standard errors; ***p < 0.001, **p < 0.01, *p < 0.05.

between proctored internet-based and paper-and-pencil survey in selection contexts; nearly all such differences favor the internet-based survey, even though the items were identical [45]. In addition, we found that survey contexts and contents of survey items (*i.e.*, internet-related, behavior-related) will significantly influence the response effects. This contrast with [45] finding that biodata

measure appears to show the most equivalence across test contexts and testing formats, but one of the weaknesses is that they did not control for the heterogeneous backgrounds of the respondents. As indicated above, this research addresses that the response propensity interacts with the mode of survey administration. However, the mode variance seems not to influence effects in validity and reliability that this research found. That is, the scales of the survey instrument stay stable and accurate in this research.

If the same analytical framework was analyzed by the data with different survey methods, the results were expected to yield the equivalent results. If not, the analytical research will fall into lack of theoretical validity. Most of the conducted tests show that the surveys are largely similar in measurement reliability and validity. Other research results indicated that internet-based and paper-and-pencil data collection methods produced equivalent mean ratings of physical and sexual attractiveness for pictures of male and female targets [8]. However, the measurement equivalence is not present in this research, the relationship between the observed internet-related and behavior-related variables remains variant across two survey modes. That is, we did not get more convincing evidences to support H_2. One possible reason is that some potential synergistic factors influence the statistical results. The synergistic results rather than individual key impacted factors provide reliable and valid answers. The previous comparative research which places undue emphasis on the comparison of individual variables may lead to unstable or spurious findings. By most accounts the data was quite similar, so it appears that there is a relationship between survey items and survey methods that may be addressed by those conducting multiple method surveys. The researchers may consider censuring their questions so as to not engage this bias. Otherwise, it should consider this mode bias into research limitation.

In sum, the findings provide sufficient evidence to conclude that the measurement scale used to score the observed variables (*i.e.* the indicators of constructs) is identical across survey methods, enabling us to draw meaningful comparisons of measurement reliability and validity across survey methods [46]. However, the results did not indicate that the respondents use a similar frame of-reference when completing the items of the survey. That is, this makes it impossible to draw meaningful comparisons of homological validity across methods (*i.e.* comparisons dealing with cause-effect relationships).

5.2. Conclusion and Recommendations

This study addresses an important topic, whether surveys conducted in different ways yield equivalent results from undergraduates. Specifically, with respect to internet surveys, this topic has been examined by many scholars and has yet to yield some evidences. However, no research examines whether it offers the possibility of improving statistical results. In this study, we sought to improve upon prior studies comparing measurement reliability, validity, and measurement equivalence for well-developed survey items collected via internet-based and

paper-and-pencil modes, controlling for homogeneous samples. In particular, our findings open a methodological door for survey researchers wishing to assess internet and behavior-related measures with internet-based and paper-and-pencil surveys.

This research found that the two surveys are not perfectly equivalent; despite many similarities there remain key differences. One of the most interesting findings is that there are differences in how questions are answered when they are related to the survey mode, for example questions about internet use were answered differently depending on mode, perhaps due to the priming provided by the survey itself. Our results suggest that two survey methods may tell a different story. It implies we cannot verify homological validity of the analytical frameworks, even if the internal consistency reliabilities, construct validity, and convergent validity for observed variables were supported. It is therefore important to take caution when using their findings to explain theoretical frameworks or feedback to practices need more accumulating evidence or control more factors that impact the statistical results. If not, we may not provide coherent answers to the research questions and hypotheses.

There is still much for survey researchers to learn concerning how to design, interpret, and account for responses obtained using multiple survey methods. Our study demonstrates that multiple-group SEM was a useful alternative for testing measurement equivalence across research samples and multiple survey methods. For examples, even if the tests of measurement reliability and validity for individual construct were quite similar in two survey modes, their measurement equivalence for the analytical framework did not appear in two survey modes. The result shows that when a measurement equivalence, both survey modes were to use of the internet-based and paper-and-pencil based. In this research we believe that the use of the internet-based survey methodology would save more time and cost with which researchers explore the contents of internet. In addition, the measurement equivalence is based on internet content, and researchers should make the internet-based survey.

An implication of this research focus is that the researchers under the study must consider the methodological problems associated with the survey methods if they plan to survey the college students. The research potential exists, but the methodological issues discussed in this research are cautious and, if not addressed, they may affect the accuracy of study findings. Therefore, the current study represents one step forward in just such a research agenda.

Acknowledgements

This research is partially supported by the Ministry of Science and Technology, Taiwan, R.O.C. under Grant No. MOS 105-2410-H-164-002-.

References

[1] Sax, L.J., Gilmartin, S.K. and Bryant, A.N. (2003) Assessing Response Rates and Nonresponse Bias in Web and Paper Surveys. *Research in Higher Education*, **44**, 409-432. https://doi.org/10.1023/A:1024232915870

[2] Campos, J.A.D.B., Zucoloto, M.L., Bonafé, F.S.S., Jordani, P.C. and Maroco, J. (2011) Reliability and Validity of Self-Reported Burnout in College Students: A Cross Randomized Comparison of Paper-and-Pencil vs. Online Administration. *Computers in Human Behavior*, **27**, 1875-1883. https://doi.org/10.1016/j.chb.2011.04.011

[3] Beuckelaer, A.D., Lievens, F. and Swinnen, G. (2007) Measurement Equivalence in the Conduct of a Global Organizational Survey across Countries in Six Cultural Regions. *Journal of Occupational and Organizational Psychology*, **80**, 575-600. https://doi.org/10.1348/096317907X173421

[4] Smither, J.W., Walker, A.G. and Michael, K.T. (2004) An Examination of the Equivalence of Web-Based versus Paper and Pencil Upward Feedback Ratings: Rater and Rate Level Analysis. *Educational and Psychology Measurement*, **64**, 40-61. https://doi.org/10.1177/0013164403258429

[5] Stanton, J.M. and Rogelberg, S.G. (2001) Using Internet/Intranet Web Pages to Collect Organizational Research Data. *Organizational Research Methods*, **4**, 200-217. https://doi.org/10.1177/109442810143002

[6] Alwin, D.F. (2007) Margins of Error: A Study of Reliability in Survey Measurement. John Wiley and Sons, Inc., Hoboken, NJ. https://doi.org/10.1002/9780470146316

[7] Meyerson, P. and Tryon, W.W. (2003) Validating Internet Research: A Test of the Psychometric Equivalence of Internet and In-Person Samples. *Behavior Research Methods, Instruments, and Computers*, **35**, 614-620. https://doi.org/10.3758/BF03195541

[8] Epstein, J., Klinkenberg, W.D., Wiley, D. and McKinley, L. (2001) Insuring Sample Equivalence across Internet and Paper-and-Pencil Assessments. *Computers in Human Behavior*, **17**, 339-346. https://doi.org/10.1016/S0747-5632(01)00002-4

[9] Beuckelaer, A.D. and Lievens, F. (2009) Measurement Equivalence of Paper-and-Pencil and Internet Organizational Surveys: A Large Scale Examination in 16 Countries. *Applied Psychology*, **58**, 336-361. https://doi.org/10.1111/j.1464-0597.2008.00350.x

[10] Wang, Y.S. (2007) Development and Validation of a Mobile Computer Anxiety Scale. *British Journal of Educational Technology*, **38**, 990-1009. https://doi.org/10.1111/j.1467-8535.2006.00687.x

[11] Yilmaz, V. and Türküm, A.S. (2008) Factors Affecting Hopelessness Levels of Turkish Pre-Teenagers Attending Primary School: A Structural Equation Model. *Social Behavior and Personality*, **36**, 19-26. https://doi.org/10.2224/sbp.2008.36.1.19

[12] Thompson, L.F., Surface, E.A., Martin, D.L. and Sanders, M.G. (2003) From Paper to Pixels: Moving Personnel Surveys to the Web. *Personnel Psychology*, **56**, 197-227. https://doi.org/10.1111/j.1744-6570.2003.tb00149.x

[13] Dillman, D.A. (2000) Mail and Internet Surveys: The Tailored Design Method. 2nd Edition, John Wiley and Sons, New York.

[14] Yun, G.W. and Trumbo, C.W. (2000) Comparative Response to a Survey Executed by Post, Email and Web Form. *Journal of Computer-Mediated Communication*, **6**, 1-26.

[15] Simsek, Z. and Veiga, J.F. (2001) A Primer on Internet Organizational Surveys. *Organizational Research Methods*, **4**, 218-235. https://doi.org/10.1177/109442810143003

[16] Chang, L. and Krosnick, J.A. (2010) Comparing Oral Interviewing with Self-Administered Computerized Questionnaires an Experiment. *Public Opinion Quarterly*, **74**, 154-167. https://doi.org/10.1093/poq/nfp090

[17] Lewis, I.M., Watson, B.C. and White, K.M. (2009) Internet versus Paper-and-Pencil Survey Methods in Psychological Experiments: Equivalence Testing of Participant Responses to Health-Related Messages. *Australian Journal of Psychology*, **61**, 107-116. https://doi.org/10.1080/00049530802105865

[18] Martins, N. (2010) Measurement Model Equivalence in Web- and Paper-Based Surveys. *Southern African Business Review*, **14**, 77-107. http://hdl.handle.net/10500/13576

[19] Manfreda, K.L., Bosnjak, M., Berzelak, J., Haas, I. and Vehovar, V. (2008) Web Surveys versus Other Survey Modes: A Meta-Analysis Comparing Response Rates. *International Journal of Market Research*, **50**, 79-104.

[20] Hair, J.F., Black, W.C., Babin, B.J., Anderson, R.E. and Tatham, R.L. (2010) Multivariate Data Analysis. 7th Edition, Prentice Hall, Upper Saddle River.

[21] Schulenberg, S.W. and 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. https://doi.org/10.3758/BF03207726

[22] Drasgow, F. and Kanfer, R. (1985) Equivalence of Psychological Measurement in Heterogeneous Populations. *Journal of Applied Psychology*, **70**, 662-680. https://doi.org/10.1037/0021-9010.70.4.662

[23] Drasgow, F. (1984) Scrutinizing Psychological Tests: Measurement Equivalence and Equivalent Relations with External Variables Are the Central Issue. *Psychological Bulletin*, **95**, 134-135. https://doi.org/10.1037/0033-2909.95.1.134

[24] Raju, N.S., Laffitte, L.J. and Byrne, B.M. (2002) Measurement Equivalence: A Comparison of Methods Based on Confirmatory Factor Analysis and Item Response Theory. *Journal of Applied Psychology*, **87**, 517-529. https://doi.org/10.1037/0021-9010.87.3.517

[25] Donovan, M.A., Drasgow, F. and Probst, T.M. (2000) Does Computerizing Paper-and-Pencil Job Attitude Scales Make a Difference? New IRT Analyses Offer Insight. *Journal of Applied Psychology*, **85**, 305-313. https://doi.org/10.1037/0021-9010.85.2.305

[26] Booth-Kewley, S., Edwards, J.E. and Rosenfeld, P. (1992) Impression Management, Social Desirability, and Computer Administration of Attitude Questionnaires: Does the Computer Make the Difference? *Journal of Applied Psychology*, **77**, 562-566.

[27] Cohen, R.J., Swerdlik, M.E. and Phillips, S.M. (1996) Psychological Testing and Assessment. 3th Edition, Mayfield, Mountain View.

[28] Cole, M.S., Bedeian, A.G. and Field, H.S. (2006) The Measurement Equivalence of Web-Based and Paper-and-Pencil Measures of Transformational Leadership: A Multinational Test. *Organizational Research Methods*, **9**, 339-368. https://doi.org/10.1177/1094428106287434

[29] Church, A.H. (2001) Is There a Method to Our Madness? The Impact of Data-Collection Methodology on Organizational Survey Results. *Personnel Psychology*, **54**, 937-969. https://doi.org/10.1111/j.1744-6570.2001.tb00238.x

[30] Dillman, D.A., Phelps, G., Tortora, R., Swift, K., Kohrell, J, Berck, J. and Messer, B.L. (2009) Response Rate and Measurement Differences in Mixed Mode Surveys Using Mail, Telephone, Interactive Voice (IVR) and the Internet. *Social Science Research*, **38**, 1-18. https://doi.org/10.1016/j.ssresearch.2008.03.007

[31] Steenkamp, J.B.E.M. and Baumgartner, H. (1998) Assessing Measurement Invariance in Cross-National Consumer Research. *Journal of Consumer Research*, **25**, 78-107. https://doi.org/10.1086/209528

[32] Vandenberg, R.J. and Lance, C.E. (2000) A Review and Synthesis of the Measurement Invariance Literature: Suggestions, Practices, and Recommendations for Organizational Research. *Organizational Research Methods*, **3**, 4-70. https://doi.org/10.1177/109442810031002

[33] Weijters, B., Schillewaert, N. and Geuens, M. (2008) Assessing Response Styles across Modes of Data Collection. *Journal of the Academy of Marketing Science*, **36**, 409-422. https://doi.org/10.1007/s11747-007-0077-6

[34] Cook, T.D. and Campbell, D.T. (1979) Quasi-Experimentation: Design and Analysis Issues for Field Settings. Houghton Mifflin, Boston.

[35] 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. https://doi.org/10.1016/S0378-7206(03)00028-4

[36] Wang, Y.S. (2003) Assessment of Learner Satisfaction with Asynchronous Electronic Learning Systems. *Information and Management*, **41**, 75-86. https://doi.org/10.1016/S0378-7206(03)00028-4

[37] Hui, C., Law, K.S. and Chen, Z.X. (1999) A Structural Equation Model of the Effects of Negative Affectivity, Leader-Member Exchange, and Perceived Job Mobility on In-Role and Extra-Role Performance: A Chinese Case. *Organizational Behavior and Human Decision Processes*, **77**, 3-21. https://doi.org/10.1006/obhd.1998.2812

[38] Ilies, R., Scott, B.A. and Judge, T.A. (2006) The Interactive Effects of Personal Traits and Experienced States on Intraindividual Patterns of Citizenship Behavior. *Academy of Management Journal*, **49**, 561-575. https://doi.org/10.5465/AMJ.2006.21794672

[39] Kline, R.B. (2005) Principles and Practice of Structural Equation Modeling. 2nd Edition, The Guilford Press, New York.

[40] Hu, L.T. and Bentler, P.M. (1999) Cutoff Criteria for Fit Indexes in Covariance Structure Analysis: Conventional Criteria versus New Alternatives. *Structural Equation Modeling*, **6**, 1-55. https://doi.org/10.1080/10705519909540118

[41] Cronbach, L.J. and Shavelson, R.J. (2004) My Current Thoughts on Coefficient Alpha and Successor Procedures. *Educational and Psychological Measurement*, **64**, 391-418. https://doi.org/10.1177/0013164404266386

[42] Robinson, J.P., Shaver, P.R. and Wrightsman, L.S. (1991) Criteria for Scale Selection and Evaluation. In: Robinson, J.P., Shaver, P.R. and Wrightsman, L.S., Eds., *Measures of Personality and Social Psychological Attitudes*, Academic Press, San Diego, 1-15. https://doi.org/10.1016/B978-0-12-590241-0.50005-8

[43] Field, A. (2005) Discovering Statistics Using SPSS. 2nd Edition, Sage, London.

[44] Vispoel, W.P., Boo, J. and Bleiler, T. (2001) Computerized and Paper-and-Pencil Versions of the Rosenberg Self-Esteem Scale: A Comparison of Psychometric Features and Respondent Preferences. *Educational and Psychological Measurement*, **61**, 461-474. https://doi.org/10.1177/00131640121971329

[45] Ployhart, R.E., Weekley, J.A., Holtz, B.C. and Kemp, C. (2003) Web-Based and Paper-and-Pencil Testing of Applicants in a Proctored Setting: Are Personality, Biodata, and Situational Judgment Tests Comparable? *Personnel Psychology*, **56**, 733-752.

[46] Vandenberg, R.J. (2002) Toward a Further Understanding of and Improvement in Measurement Invariance Methods and Procedures. *Organizational Research Methods*, **5**, 139-158. https://doi.org/10.1177/1094428102005002001