CyberPsychology & Behavior
Volume 10, Number 4, 2007

DOI: 10.1089/cpb.2007.9998

Comparison of Internet-Based and Paper-Based Questionnaires in Taiwan Using Multisample Invariance Approach

SEN-CHI YU, Ph.D.[1] and MIN-NING YU, Ph.D.[2]

ABSTRACT

This study examines whether the Internet-based questionnaire is psychometrically equivalent to the paper-based questionnaire. A random sample of 2,400 teachers in Taiwan was divided into experimental and control groups. The experimental group was invited to complete the electronic form of the Chinese version of Center for Epidemiologic Studies Depression Scale (CES-D) placed on the Internet, whereas the control group was invited to complete the paper-based CES-D, which they received by mail. The multisample invariance approach, derived from structural equation modeling (SEM), was applied to analyze the collected data. The analytical results show that the two groups have equivalent factor structures in the CES-D. That is, the items in CES-D function equivalently in the two groups. Then the equality of latent mean test was performed. The latent means of "depressed mood," "positive affect," and "interpersonal problems" in CES-D are not significantly different between these two groups. However, the difference in the "somatic symptoms" latent means between these two groups is statistically significant at $\alpha = 0.01$. But the Cohen's d statistics indicates that such differences in latent means do not apparently lead to a meaningful effect size in practice. Both CES-D questionnaires exhibit equal validity, reliability, and factor structures and exhibit a little difference in latent means. Therefore, the Internet-based questionnaire represents a promising alternative to the paper-based questionnaire.

INTRODUCTION

The Internet is an effective medium for posting, exchanging, and collecting data in psychology-related research.[1] Given the accelerated growth of the Internet, some studies contend that Internet surveys (Web surveys, online surveys, Internet-administrated questionnaires, or online questionnaires) will soon replace the conventional methods.[2] Advantages of using the Internet are methodological and economical. First, the Internet can acquire a large, heterogeneous sample that generates powerful statistical inferences. Access to a large population also provides enhanced external validity. Second, since the Internet can offer respondents anonymity, recruiting a specialized sample (people with rare characteristics) is possible. Third, time and cost savings are associated with eliminating printing and mailing of an instrument, having data returned in an electronic format. Furthermore, the In-

[1] Center for Teacher Education, Huafan University, Taiwan.
[2] Department of Education, National Chengchi University, Taiwan.

ternet provides access to the tool around the clock, that is, without time or space limitations.[1,3,4]

Given that the Internet is a powerful and efficient tool for conducting psychological experiments or surveys, invariance of psychometric properties of Internet-based and paper-based questionnaires must be investigated when using the Internet as a tool for collecting data. To determine whether Internet-based questionnaires are as psychometrically invariant as traditional mailed questionnaires, a random sample of 2,400 teachers in Taiwan was recruited. This sample population was chosen for several reasons.

First, Taiwan is an appropriate environment for Internet study because Web penetration is high, thereby reducing the influence of coverage error. More than 42% of Taiwan's population are online; in excess of 80% are broadband users.[5] Taiwan was ranked the third highest worldwide for broadband coverage and twelfth in Internet coverage. Moreover, subway stations and most schools are equipped with free wireless Internet.

Second, teachers are a qualified sample for an Internet study because computer and Internet literacy are required skills for Taiwanese teachers. Moreover, the completeness of the population frame is essential for random sampling to obtain reliable and generalizable results. A complete population frame of elementary and secondary schools is available at the Ministry of Education website (http://www.edu.tw), and lists of faculty at these schools are available on school websites.

Third, this study utilized probability-based sampling that generated good external validity. Although certain studies recruited participants via the Web,[1] bias in self-selected samples is inevitable. Voluntary samples of Web users should not be considered as random samples of any particular populations. Birnbaum[3] indicated that self-selected samples are on average older than college students and have greater mean level of education and greater variance in age and education than do college students. Couper[2] also reported that the "Internet population," compared with general populations, has a higher income and education. Moreover, online samples overrepresented males, college graduates, and young users.

Structural equation modeling (SEM) methodology has been utilized to determine whether Internet-based and paper-based samples exhibit invariant psychometric properties. Structural equation modeling, also called the latent variable model, has become the preeminent statistical technique in the social sciences.[6] Most theories in the social and behavioral sciences are formulated in terms of hypothetical constructs or latent variables. Latent variables are not directly measurable and are generally poorly defined; consequently, measurement errors are inevitable. However, traditional statistical techniques, which assume zero measurement error and do not account for the errors of measurements, attenuate the true association between variables.[7] Conversely, the SEM describes how well observed indicators serve as an instrument for latent variables and takes measurement errors into consideration.

The multisample invariance approach, derived from SEM, was utilized to investigate whether the Internet-based sample and paper-based sample exhibit invariant psychometric properties. The multisample invariance approach determines whether items in an instrument operate equivalently across different populations. When parameter values of a measurement model differ across groups, a risk of serious errors exists.[8] For example, say that a measure of "depressive mood" (x) for males is related to the latent construct "depression" (ξ) such that $x = \xi + \epsilon$, whereas for females, the measure equation is $x = 0.6\ \xi + \epsilon$. Even when males and females have the same nonzero mean for depression, the means for depressive mood would typically indicate that the average level of depression for females is lower than that for males. Therefore, it is essential that measurement model invariance ensures that Internet-based instruments are equivalent alternatives to paper-based instruments.

Several studies, via exploratory factor analysis (EFA), demonstrated that Internet-based and paper-based instruments exhibited "similar" levels of reliability, numbers of factor, and factor loadings (e.g., Riva et al.,[1] Buchanan & Smith[9]). However, these findings do not guarantee the invariance of psychological properties of Internet-based and paper-based instruments. Since EFA is a sample-dependent technique, various factor structures can obtain when an instrument is administrated to different populations. Moreover, no criterion exists for comparing differences in the factor analysis parameters based on different groups.

Conversely, in SEM, group difference for any individual or sets of parameters can be tested by specifying cross-group equality constraints.[6–8] Additionally, the fit between data and the model can be evaluated by the fit indices of SEM. Byrne[10] concluded that multisample invariance can answer the following questions:

- Is the measurement model group-invariant?
- Is the structural model group-invariant?
- Are certain paths in specific causal structure invariant across populations?

- Are the latent means of particular constructs in a model invariant across populations?
- Does the factorial structure of a measurement replicate across independent samples for the same population; that is, does it attain the cross-validation?

Two studies were conducted to evaluate the invariance across Internet-based and paper-based groups. In Study 1, the invariance of the factor structure, instead of traditional EFA technique, was examined. In Study 2, the invariance of the mean structure, instead of the traditional *t*-test technique, was analyzed.

SAMPLE

The target population in this study was teachers at elementary schools and junior and senior high schools in Taiwan. The sampling unit was a school. The population frame was available at the Ministry of Education website. Based on the proportion of teachers at school levels to the total population of teachers, a total sample of 300 schools consisting of 38 senior high schools, 57 junior high schools, and 205 elementary schools were sampled. Eight teachers from each sampled school were randomly sampled and randomly assigned to an Internet-based group ($n = 4$) or a paper-based group ($n = 4$).

INSTRUMENT

The instrument utilized in this study was the Mandarin-Chinese version of the Center for Epidemiologic Studies Depression Scale (CES-D). The CES-D includes 20 items that reflect affective, attitudinal, and somatic elements symptomatic of depression. The CES-D is one of the most commonly applied brief depression instruments and has been translated into many languages, including Russian, Spanish, French, Japanese, Italian, American Indian languages, Cantonese, and Mandarin.[11] The CES-D was translated into a Mandarin-Chinese version by the authors of this study.

PROCEDURE

A hard copy of the questionnaire was mailed to the members of the paper-based group, and a letter was mailed to the members of the Internet-based group inviting them to access the Internet to complete the questionnaire that was hosted on the National Chengchi University Web server (http://e-testing.nccu.edu.tw). To control for multiple submissions from the same respondents, the Internet-based group was asked to key in a password provided on their invitation letter before completing the questionnaire.

To improve the response rate, nonrespondents in the Internet-based group were reminded by follow-up postcards to complete the questionnaire, whereas nonrespondents in the paper-based group were sent a follow-up letter and a replacement hard copy of the questionnaire. The reminders were mailed two weeks after the first delivery of the questionnaire and only to those individuals who had not yet responded. Table 1 presents the descriptive statistics of the CES-D total score.

MULTISAMPLE INVARIANCE TECHNIQUE

In analyzing a single sample, there is seldom any interest in mean value of latent variables and intercept terms in the equation. All latent variables are measured in deviations from their means.[7] The linear structural relations of SEM can be defined as

$$\eta = \mathrm{B}\eta + \Gamma\xi + \xi \tag{1}$$

$$y = \Lambda_y \eta + \epsilon \tag{2}$$

$$x = \Lambda_x \xi + \delta \tag{3}$$

where η and ξ are latent variables, x and y are observed variables, Λ_x and Λ_y are factor loadings, ϵ and δ are error variables, and B and Γ are structural parameters.

TABLE 1. DESCRIPTIVE STATISTICS OF THE CES-D TOTAL SCORE

Group	*Sample size*	*Mean*	*S.D.*	*Skewedness*	*Kurtosis*
Internet-based group	541	11.03	7.87	1.62	4.10
Paper-based group	630	12.14	8.02	1.31	2.34

Consider a set of G populations. The model for group g is defined by the eight parameter matrices:

$$\Lambda_x^g, \Lambda_y^g, \Theta_\delta^g, \Theta_\epsilon^g, \mathrm{B}^g, \Gamma^g, \Phi^g, \Psi^g$$

where the superscript g refers to the gth group, and $g = 1, 2, \ldots G$. Φ^g and Ψ^g are variance/covariance matrices. Multisample analysis was applied to test whether these eight parameter matrices are equal for different groups.

Typically, SEM focuses on analyzing covariance structures; thus, only parameters representing regression coefficients, variance, and covariance were of interest. Analyzing covariance structure implicitly assumes that all observed variables have a zero mean and are measured as deviations from their means. Therefore, intercept terms associated with regression equations are irrelevant to analysis. However, in mean structure, observed means take on nonzero values; therefore, the intercept terms must be considered.[10] Within the covariance structure, the regression equation can be expressed as

$$\mathrm{Y} = \mathrm{Bx} + \epsilon \tag{4}$$

Conversely, within the mean structure, the regression equation can be expressed as

$$\mathrm{Y} = \alpha + \beta\mathrm{x} + \epsilon \tag{5}$$

Therefore, the regression equation in SEM of mean structure can be modified from Eqs. 1–3 and expressed as

$$\eta = \alpha + \mathrm{B}\eta + \Gamma\xi + \xi \tag{6}$$

$$y = \tau_y + \Lambda_y\,\eta + \epsilon \tag{7}$$

$$x = \tau_x + \Lambda_x\,\xi + \delta \tag{8}$$

where α, τ_y, τ_x are intercept terms. So, the expected values of latent exogenous and endogenous variable can be expressed as[10]

$$\mathrm{E}(\xi) = k \tag{9}$$

RESULTS

To compare the latent mean between two groups, the total sample confirmatory factor analysis (CFA), factor structure invariance, and mean structure invariance must be conducted in turn. Analytical results of these three analyses are shown as follows.

1. Total-sample CFA

The model for the second-order CFA, based on Radloff's study,[12] assumed that CES-D exhibited a four-factor structure composed of "depressed mood," "positive affect," "interpersonal problems,"

Table 2. Summary of Fit Indices

Model	χ^2	*df*	$\Delta\chi^2$	Δ *df*	*RMSEA*	*CFI*	*GFI*
			Total-Sample CFA				
Second-order CFA	777.11	166			0.056	0.98	0.94
			Testing factor Structures Invariance				
Model 1: H_{form} (baseline model)	940.00	328			0.056	0.98	0.92
Model 2: H_{λ_x}	949.56	348	9.56	20	0.054	0.98	0.92
Model 3: $H_{\lambda_x\phi}$	965.68	353	13.63	5	0.055	0.98	0.92
Model 4: $H_{\lambda_x\phi\Theta_\delta}$	982.02	374	17.98	20	0.053	0.98	0.92
			Testing Mean Structures Invariance				
Model 5: $H_{\lambda_x\phi\Theta_\delta}$	982.02	374			0.053	0.98	0.92
Model 6: $H_{\lambda_x\phi\Theta_\delta\mathrm{K}}$	1023.31	389			0.053	0.98	0.92

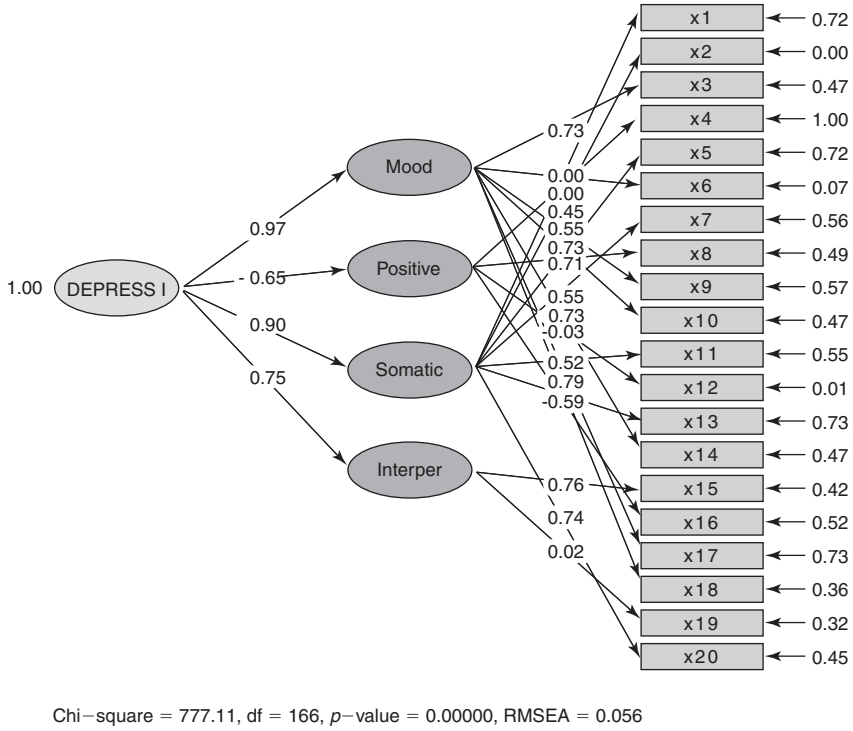

FIG. 1. CFA of total sample.

and "somatic symptoms". Table 2 and Figure 1 present the analytical results for CFA.

To evaluate data-model fit, this study first examined overall model fit—that is, whether the model is theoretical meaningful. This study examined whether signs of paths are in opposite directions, whether all estimated parameters are statistically significant, and whether the errors are close to 1. The overall model fit was acceptable, except that item 4 ("I am as good as others") exhibited a nonsignificant factor loading.

Next, formal statistical fit indices, comprising chi-square, goodness of fit index (GFI), comparative fit index (CFI), and root-mean square error of approximation (RMSEA), were utilized in this study. GFI statistics evaluate the closeness between unrestricted sample covariance matrix Σ, and the restricted (model-implied) covariance matrix $\Sigma(\theta)$. The null hypothesis is equivalent to the hypothesis that $\Sigma - \Sigma(\theta) = 0$.

The chi-square statistic equaled $(N-1)F_{\min}$ (sample size minus 1, multiplied by the minimum-fit function), and in large samples, the chi-square statistic is distributed as a central chi-square with degree of freedom (*df*) equal to $(p)(p+1)/2 - t$, where p is the number of observed variables and t is the number of parameters to be estimated.[8,10] The model chi-square equaled 736.94 ($p < 0.01$), indicating a bad fit. However, some problems exist when relying solely on chi-square as a fit statistics.

First, chi-square is sensitive to sample size and model complexity. Large sample size, which is critical to obtaining a precise parameter estimate and the tenability of an asymptotic distribution, results in a high value of chi-square and model rejection. Furthermore, the chi-square approximation assumes a multinormal distribution. However, nonnormal observed variables occur in practice. In light of the problems associated with chi-square, using it as a measure of badness-of-fit may be more appropriate than as test statistics.[6]

Other researchers have addressed chi-square limitations by developing other fit indices. The RMSEA, a parsimony-adjusted index, estimates the amount of error of approximation per model *df* and takes sample size into account. RMSEA values below 0.05 indicate a good fit, values of 0.05–0.08 suggest a moderate fit, and values above 0.10 indicate a poor fit.[6,13] The RMSEA of this model equals 0.055, indicating a moderate fit. Furthermore, the CFI equals 0.98 and GFI equals 0.94, indicating the model fits well.

Given the model fits well prior to any model modification, we keep the original CES-D factor structures without any model trimming to retain the original factor structures and scoring.

2. Study 1: Testing for invariant factor structures

Testing multisample invariance involves a series of increasingly restrictive hypotheses. Among all procedures, testing the equality of factor structures is the first step. Prior to testing for invariance across groups, the baseline models (i.e., the least restricted models) must be established for each group separately. In baseline models, groups have the same form without restricting any nonfixed parameters to have the same values across models.

Table 2 presents the results of testing the invariance hypothesis. The baseline model (model 1) shows a good fit with RMSEA of 0.056 and GFI and CFI greater than 0.90.

Given the baseline model fit, the next step is to assess the model (model 2) in which factor loadings are constrained to be equal in both groups. Since the equality scaling is prior to the measurement error variance or the equality of covariance, model 2 ($^{H_{\lambda_x}}$) precedes the last two hypothesis ($^{H_{\lambda_x\Phi}}$ and $^{H_{\lambda_x\Phi\Theta_\delta}}$).[8] Model 2 also exhibited a good fit for RMSEA, GFI, and CFI (Table 2). The difference in chi-square equals $9.56 < \chi^2_{.01}(df = 20) = 37.57$. Based on the nonsignificant difference in χ^2 values, this study concluded that the hypothesis of an invariant pattern of factor loadings held.

Next, this study tested the invariance of factor loadings and measurement error variances. Model 3 also had a good fit for RMSEA, GFI, and CFI (Table 2). The difference in chi-square equals $13.63 < \chi^2_{.01}(df = 5) = 15.09$. Based on the nonsignificant difference in χ^2 values, we conclude that the hypothesis of an invariant pattern of factor loadings and measurement error variances held.

Finally, the last step in this hierarchy is model 4, in which all three parameter matrices are simultaneously tested for equality. Since the chi-square between model 3 and model 4 was not significant ($\Delta\chi^2 = 17.98 < \chi^2_{.01}(df = 20) = 37.57$), the hypothesis of an invariant pattern of factor loadings, measurement error variances, and factor variances were not rejected at $a = 0.01$.

3. Study 2: Testing for invariant latent mean structures

Given the invariance of factor structures across the Internet-based and paper-based groups, the differences in latent means across groups were estimated. To achieve this goal, a dummy variable that linear structural relations (LISREL) terms constant was incorporated into the model. Next, the factor intercepts kappa (κ) (as shown in Eq. 9) for one group was fixed to zero; this group then operated as the reference group against which latent means for other groups are compared.[10] In the specifications of LISREL syntax, the KA (kappa) matrix was specified as free for the Internet-based group and as fixed at zero for the paper-based group. Table 3 presents the analytical results. For the four latent factors in CES-D, the latent means of "depressed mood," "positive affect," and "interpersonal problems" are not significantly different between these two groups. However, the difference between the "somatic symptoms" latent means of these two groups is statistically significant at $a = 0.01$. Since the magnitude of difference (0.19) is minor concerning the CES-D total scores ranging from 0 to –60, effect size was utilized to evaluate such difference.

Effect-size measurements represent the relative magnitude of the experimental treatment, whereas the statistical tests of significance indicate the likelihood that experimental results differ from chance expectations. Concerning the availability of parameters in this study, Cohen's d statistics was utilized. Cohen's d is expressed as[14]

$$d = t\sqrt{\frac{(n_t + n_c)\,(n_t + n_c)}{(n_t n_c)\,(n_t + n_c - 2)}} \tag{10}$$

where d is the Cohen's d effect size, t is the t statistics, and n is the sample size. The subscript t represents the treatment condition, and c denotes the comparison condition. In this study, $n_t = 541$ and $n_c = 630$, $t = -2.93$; therefore, $d = 0.172$ was ob-

Table 3. Testing for Latent Mean Structures

	Depressed mood	*Positive affect*	*Somatic symptoms*	*Interpersonal problems*
Differences of kappa	−0.09	0.09	−0.19	−0.15
Standard deviation	0.06	0.07	0.07	0.07
t value	−1.38	1.35	−2.93	−2.17
Decision ($\alpha = 0.01$)	Accepted	Accepted	Rejected	Accepted

tained. According to Cohen,[15] effect sizes of 0.20 are small, 0.50 are medium, and 0.80 are large. However, Cohen's *d* statistic was only 0.172 in this study, indicating a small effect size. That is, the minor difference in "somatic symptoms," 0.19 point, albeit statistically significant, does not result in meaningful effect size (clinical significance) in practice.

CONCLUSION

Analytical results showed that the factor structures remained invariant across the Internet-based and paper-based groups. That is, there exists an invariant pattern of factor loadings, measurement error variances, and factor variances between the two groups. Therefore, we conclude that the CES-D items operate equivalently across these two groups.

Since the factor loadings linking observed and latent variables could be viewed as the validity of a latent construct from the CFA perspective,[8] we conclude that the CES-D items exhibited an equivalent validity across the Internet-based and the paper-based groups given the invariance of factor loadings. Moreover, the proportion of variance, R-square, in an observed variable that is accounted for by a latent variable is an instrument's reliability in the CFA approach.[8] The R-square and square-of-factor loadings also remained invariant based on the factor loadings' invariance. Therefore, we conclude that CES-D exhibited an equivalent reliability across the Internet-based and the paper-based groups. For the latent means of the CES-D scores, the latent means of "depressed mood," "positive affect," and "interpersonal problems" are not significantly different between these two groups. The minor difference in "somatic symptoms," 0.19 point, albeit statistically significant, does not result in meaningful effect size either.

From the invariance of the factor structures, validity, reliability, and little differences of latent means, we conclude that the Internet-based instrument is equivalent to the paper-based version in terms of psychometric properties. Furthermore, analytical results also indicate that the paper-based instrument is cross-validated by the Internet-based version. We conclude that the Internet-based questionnaire is a promising alternative to the paper-based questionnaire.

ACKNOWLEDGMENT

The authors express their thanks for this research grant support by the National Science Counsel of Taiwan under the contract NSC-93-2413-H-004-008 and NSC-95-2413-H-211-001.

REFERENCES

1. Riva, G., Teruzzi, T., & Anolli, L. (2003). The use of the internet in psychological research: comparison of online and offline questionnaires. *CyberPsychology & Behavior* 6:73–80.
2. Couper, M.P. (2000). Web survey: a review of issue and approaches. *Public Opinion Quarterly* 64:464–495.
3. Birnbaum, M.H. (2004). Human research and data collection via the Internet. *Annual Review of Psychology* 55:803–822.
4. Kaplowitz, M.D., Hadlock, T.D., & Levine, R. (2004). A comparison of Web and mail survey response rates. *Public Opinion Quarterly* 68:94–101.
5. Find.org. (2005). Internet users in Taiwan. Available at: http://www.find.org.tw/find/home.aspx?page=many&id=134. Accessed July 29, 2006.
6. Kline, R.B. (2005). *Principles and practice of structural equation modeling*. (2nd ed.). New York: Guilford.
7. Jöreskong, K.G., & Sörborm, D. (1993). *LISERAL 8: Structural equation modeling with the SIMPLIS command language*. Mooresville, IN: Scientific Software.
8. Bollen, K.A. (1989). *Structural equations with latent variables.* New York: Wiley.
9. Buchanan, T., & Smith, J.L. (1999). Using the Internet for psychological research: personality testing on the World Wide Web. *British Journal of Psychology* 90: 125–144.
10. Byrne, B.M. (1998). *Structural equation modeling with LISREL, PRELIS, and SIMPLIS: basic concepts, applications, and programming.* Mahwah, NJ: Erlbaum.
11. Breithaupt, K., & Zumbo, B.D. (2002). Sample invariance of the structural equation model and the item response model: a case study. *Structural Equation Modeling* 9:390–412.
12. Radloff, L.S. (1977). The CES-D scale: a self-report depression scale for research in the general population. *Applied Psychological Measurement* 1:385–401.
13. Browne, M.W., & Cudeck, R. (1993). Alternatives ways of assessing model fit. In Bollen, K.A., & Long, J.S. (eds.) *Testing structural equation models*. Newbury Park, CA: Sage, pp.136–162.
14. Thalheimer, W., & Cook, S. (2002). *How to calculate effect sizes from published research articles: a simplified methodology*. Available at: http://work-learning.com/effect_sizes.htm. Accessed November 31, 2005.
15. Cohen, J. (1992). A power primer. *Psychological Bulletin* 112:155–159.

Address reprint requests to:
Dr. Sen-Chi Yu
Center for Teacher Education
Huafan University
No.1, Huafan Rd., Shiding
Taipei County 223, Taiwan (R.O.C.)

E-mail: rhine@cc.hfu.edu.tw; fuzzyirt@yahoo.com