What causes the difference in mean structures between online and paper-and-pencil administrations?: Examination using the Self-report Depression Scale

Naoya Todo, PhD 1 and Yusuke Umegaki, PhD, 2

1 *Faculty of Human Sciences, University of Tsukuba*

2 *Faculty of Human Life and Environment, Nara Women’s University*

Corresponding Author:

Naoya Todo, Assistant Professor, Faculty of Human Sciences, University of Tsukuba, 1-1-1 Tennodai, Tsukuba, Ibaraki, 305-8577, Japan.

Email: [ntodo@human.tsukuba.ac.jp](mailto:ntodo@human.tsukuba.ac.jp)

Tel: +81 29-853-6712

What causes the difference in mean structures between online and paper-and-pencil administrations?: Examination using the Self-report Depression Scale

**ABSTRACT**

Due to the many advantages of online surveys, many researchers are taking advantage of this survey method. Although for many psychological instruments, previous studies have shown that online and paper-and-pencil administration formats have equivalent results, other studies have shown that some online surveys result in different score distributions from those observed through paper-and-pencil administration. In this study, we conducted surveys using Zung’s self-report depression scale (SDS) to Japanese undergraduates both online and through paper-and-pencil and examined whether there are differences between different administration formats only in the scale’s mean structures and, if so, why the difference occurs. Analysis results showed that there was the difference only in mean structures. Results also implied that the online administration format lowers item thresholds; this decrease would cause the difference between the two formats’ mean structures. Finally, we think about the future directions of this research; to examine whether similar results would be seen in other scales, other countries, and other generations.

Key Words: Online, Paper-and-pencil, SDS, Item threshold

What causes the difference in mean structures between online and paper-and-pencil administrations?: Examination using the Self-report Depression Scale

**Introduction**

Today, online surveys have become more and more popular in social sciences (e.g., Miura & Kobayashi, 2016). There are many advantages with the online administration of questionnaires. First, online surveys can be taken long-distance. This gives researchers opportunities to access individuals who cannot participate in paper-and-pencil surveys. Online surveys, for example, have provided samples which were more diverse in terms of gender, socioeconomic status, geographic region, and age than paper-and-pencil surveys, and which are more representative of studied populations (Gosling, Vazire, Srivastava, & John, 2004). Second, in online surveys, we can also prevent missing values by requiring all items to be completed before the survey is submitted. In addition, online surveys provide us data which can be directly used in statistical programs directly. This reduces costs and time associated with studies and is relatively inexpensive. Furthermore, online surveys enable researchers to provide immediate feedback regarding the study results to participants, which may increase participants’ motivation to complete the study. Because of these advantages, many psychological instruments have been used in an online format (Buchanan & Smith, 1999b; Carlbring et al., 2007; Schmidt, 1997).

To administer psychological instrument online when it was originally developed for paper-and-pencil use, we must assess the equivalence of results from the two administration formats rather than assume that they will be the same (e.g., Buchanan et al., 2005; Hirose, 2000). For example, scale scores of the two administration formats need to have similar means, variances, and distributions, and give similar percentile ranks of respondents.

Many studies have tackled equivalence confirmation between different administration formats. These validation studies have shown that although there were some cases that implied differences between paper-and-pencil and online administration formats in terms of measuring constructs (e.g., Buchanan et al., 2005), equivalence was confirmed for many psychological instruments (e.g., Buchanan & Smith, 1999a; Carlbring et al., 2007; Gosling et al. 2004; Pasveer & Ellard, 1998; Spek, Nyklicek, Cuijpers, & Pop, 2008; Thorén, Andersson, & Lunner, 2012). For example, Buchanan and Smith (1999a), Carlbring et al. (2007), and Pasveer and Ellard (1998) showed that alpha coefficients of online and paper-and-pencil administrations were similar in SMS-R (Gangestad & Snyder, 1985), BDI II (Beck & Steer, 1996), and STQ (Pasveer & Ellard, 1998) and so on. In addition, Buchanan and Smith (1999a) and Pasveer and Ellard (1998) reported that factor structures of scales were nearly identical between the two administration formats in SMS-R and STQ.

On the other hand, some research shows that online administration formats may decrease embarrassment, feelings of being judged, or shyness in responding scales, and encourage participants’ honest responses (Gosling et al. 2004; Whitehead, 2007). Joinson (1999) also reported that online administration formats lower participants’ social anxiety and social desirability than paper-and pencil administration. Along with these studies, Buchanan (2002) suggests that there is a difference in some scale scores levels between online and paper-and-pencil administration. Previous studies have also reported that some online administered scale scores tended to have different levels from those administered by paper-and-pencil on average (Davis, 1999; Schulenberg & Yutrzenka, 1999). Based on these studies, we believe that there are differences between online and paper-and-pencil administration formats in a scale’s mean structures, but not covariance structures. The aim of this study was to understand why this difference occurs, and what mechanism influences the difference.

In this study, we conducted surveys using the same scale (SDS) on both paper-and-pencil and online formats. Then, using the collected data, we examined whether there were differences between administration formats in the scale’s mean structures and there were no differences in the scale’s covariance structures. Furthermore, if there were only differences in mean structures, we examined why the difference occurred.

**Methods**

**Measures**

In this study, we used Zung’s self-report depression scale (SDS; Zung, 1965; Japanese translation: Fukuda & Kobayashi, 1973) for data collection. The SDS measures the severity of depressive symptoms at the time of testing (Fukuda & Kobayashi, 1973; Zung, 1965). The SDS is a 4-point 20-item Likert-type scale ranging from 1 (a little of the time) to 4 (most of the time). The items reflect common symptoms of depression, e.g., “I feel down-hearted and blue” and “morning is when I feel the best.” Of the 20 items, 10 are negatively worded.

**Participants**

Undergraduates were chosen as study participants. Depression at this age is associated with significant psychological or social outcomes (e.g., Angst, 1996; Ettner, Frank, & Kessler, 1997; Kessler et al., 2005; Kessler, Foster, Saunders, & Stang, 1995; Kessler, Walters, & Forthofer, 1998; Weitzman, 2004), and many studies focus on measuring depression during this time period. For the disciplines such as clinical psychology, it is therefore useful to investigate the difference of the psychometric properties of SDS scores between online and paper-and-pencil administration.

**Paper-and-pencil administration**.

Between April and December 2011, paper-based questionnaires were distributed to 471 Japanese students enrolled in undergraduate psychology classes in 2 universities (men: 274, women: 188, unknown: 9). The mean age of participants was 20.28 years (*SD* = 1.22).

**Online administration.**

In December 2014 and February 2015, online questionnaire surveys were conducted with 1,000 (men: 500, women: 500) and 824 (men: 412, women: 412) Japanese university students, respectively. Both participant groups were recruited from the general Japanese undergraduate population by an Internet panel survey provider Macromill, Inc. The mean age of participants was 20.78 years (*SD* = 1.69) and 20.82 years (*SD* = 1.70) respectively.

**Procedure**

**Paper-and-pencil administration.**

Participants completed the SDS questionnaire as part of a study designed to investigate the relations between depression and the intention to seek help. The questionnaire was distributed to undergraduates either right before or after their scheduled lecture sessions. Students were asked to read the information sheet and then complete the survey immediately if they agreed to participate in the study. The top page of the questionnaire contained the consent paragraph which noted that the process of participants proceeding to the survey and completing it constituted consent. Participants gave demographic information concerning age, gender, and grade, and then read and answered the SDS as well as other questionnaires. After completion, the questionnaire was immediately collected. The study was approved by the ethics committee of the second author’s affiliated university.

**Online administration**.

For online administration, participants completed the SDS questionnaire as part of a separate study designed to compare multiple different depression scales. Participants were asked to read the information page first, and then answer the online questionnaire immediately if they agreed to participate in the study. The top page of the questionnaire contained the consent paragraph which noted that completing the survey constituted consent. Participants also gave demographic information such as age and gender. This study was approved by the ethics committee of the second author’s affiliated university.

For both administrations, no identifying information was collected in order to keep the data anonymous.

**Analysis**

To examine whether there are differences between online and paper-and-pencil formats in the mean structures of the SDS and, if so, why the difference occurs, we completed the following analyses. First, we calculated descriptive statistics, including the means and standard deviations of SDS total scores in the two administration samples. Then, based on the cut-off scores provided in a previous study (Kawakami & Koizumi, 1986), we checked frequency distributions of depression in the two samples. Next, we calculated an effect size, Hedges’ *g*, and variance-covariance matrices of the item scores in the two samples. Further, we made scree plots and calculated alpha coefficients. In addition, we performed explanatory factor analyses (EFA) and multiple-group confirmatory factor analysis (MGCFA). In MGCFA, we assumed that items of each scale were on ordinal scales, and we estimated the parameters by weighted least squares estimation with robust standard errors and a mean- and variance-adjusted test statistic. Within MGCFA, we assumed that the two groups had equal loadings, uniqueness, and factor means & variances (only item thresholds were different between two groups) based on the analyses results at that point. After fitting the model, we assessed model fit indices. Finally, we applied *t* test to estimated threshold differences and calculated Cohen’s *d* of the estimated thresholds.

Statistical analyses were conducted using R 3.5.3 (R Core Team, 2019) and R packages; compute.es (Del Re, 2013), psych (Revelle, 2018), and lavaan (Rosseel, 2012).

**Results**

**Excluding participants with missing values**

We excluded the data of 7 participants’ with missing values from the analysis. This reduced the sample size to 2288.

**Descriptive statistics**

We calculated descriptive statistics of the SDS total scores. The mean±SD for the total scores were 42.36±7.52 in paper-and-pencil administration and 44.58±8.49 in online administration.

**Frequency distribution of depression**

Based on the cut-off scores presented in a previous study using Japanese samples (Kawakami & Koizumi, 1986), we examined frequency distributions of depression severity. The results showed that 175 (38%) of the participants were non-depressed (39), 172 (37%) slightly (40-47), 99 (21%) mildly (48-55), and 18 (4%) severely (56≤) on the paper-and-pencil administration. On the other hand, among the participants who completed on the online format, 515 (28%) were non-depressed (39), 588 (32%) slightly (40-47), 563 (31%) mildly (48-55), and 158 (9%) severely (56≤).

**Hedges’ *g* of the total scores**

We calculated the Hedges’ *g* in order to estimate the standardized mean difference between the total scores (the mean of online sample – the mean of paper-and-pencil sample). Calculated Hedges’ *g* was .27, implying that there was small effect between paper-and-pencil and online administration (Cohen, 1969).

**Variance-covariance matrices of the item scores**

We calculated variance-covariance matrices of item scores in the two samples. The results showed that both matrices had almost the same structure; the mean difference of the two matrixes’ components was -0.04.

**Scree plots of the correlation matrices**

To examine the dimensionality, we made scree plots from the sample correlation matrices of the paper-and-pencil and online administration data (Figure 1). The scree plots showed that the two data sets had similar scree plots, and that these scree plots implied that there were two factors behind the both data.

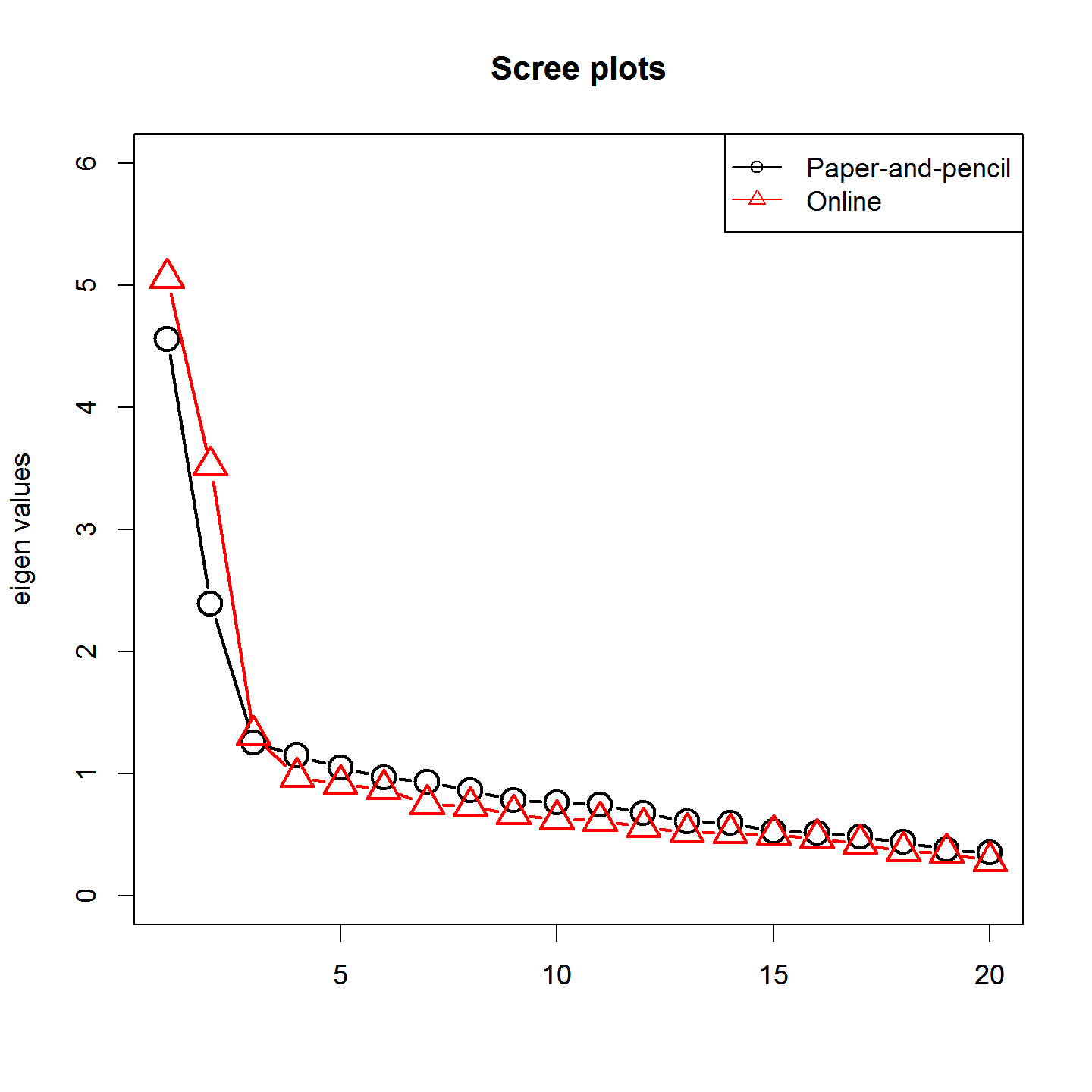


Figure 1. Scree plots of the paper-and-pencil and online administration samples

**Alpha coefficients of the sum scores**

We calculated the alpha coefficients of the sum scores in the two samples. The value in the paper-and-pencil administration was .777, while the online format’s value was .819. From these values, we can state that the reliabilities of the two sum scores were similarly high.

**Exploratory factor analysis**

To assess the equality of measured constructs between paper-and-pencil and online administration formats, we conducted exploratory factor analyses to the two samples. The scree plots implied that, regardless of administration formats, there were two factors behind SDS items; this is similar to the results of Umegaki and Todo (2017). SDS was, however, theoretically developed to measure a single factor, depression; a meta-analysis supported this (Shafer, 2006). Therefore, within this study, we conducted EFAs hypothesizing one-factor structure to the samples.

Conducting one-factor EFAs, we calculated the correlation of factor loadings between paper-and-pencil and online administration formats. The calculated correlation was a very high positive value (.901); this implied that these two administration formats measured the same construct, depression.

**Multiple-Group Confirmatory Factor Analysis**

We performed MGCFA on the data from paper-and-pencil and online administration samples. Fit indices were CFI = .741; RMSEA = .135; 𝜒2 (359) = 7858.068, *p* < .001. The low CFI, slightly high RMSEA and significant chi-square statistic were not surprising, given the evidence for two-factor structures of the SDS (e.g., Figure 1) and the large number of participants included in the present study. On the other hand, the fit indices of this restricted MGCFA model were better than those of the MGCFA model with no restriction (CFI = .722; RMSEA = .144). The test statistic, which indicates discrepancy between fitted model and data, of the unrestricted MGCFA model was much higher (𝜒2 (340) = 8384.013) than that of the restricted model. Therefore, from these results, it was likely that the restricted model fit the data.

**Comparison of threshold parameter estimates**

Finally, utilizing estimates of the restricted MGCFA model, we calculated threshold differences (an estimated threshold parameter of the paper-and-pencil sample – that of online sample), and applied one sample (one-sided) *t* test to the results. The outcomes showed that the mean of thresholds differences is likely to be not equal to 0 (*t* = 3.18, *df* = 59, *p* < 0.01). In addition, we calculated Cohen’s *d* of estimated threshold parameters (the mean of paper-and-pencil sample – the mean of online sample), which was .13. This result implied that administration formats slightly affect item thresholds.

**Discussion**

**Summary and discussion of the results**

In this study, to assess whether different administration formats do not affect scales’ covariance structures but do affect mean structures and to understand the factor which causes the difference in mean structures, we conducted surveys using a self-report depression scale (SDS) to Japanese undergraduates using paper-and-pencil and online administration formats.

Within this study, the mean of SDS total scores in the online administration sample was higher than that of the paper-and-pencil administration. The ratio of Japanese undergraduates with mild or severe depression in the online administration was also higher than that of the paper-and-pencil administration. This was the expected result. In addition, the estimates of standardized mean difference (Hedges’s *g*) showed small effect between the administration formats, which implied that the same tendency would also be seen in the population (Japanese undergraduates). Although total score variance in the online samples was bigger than that of paper-and-pencil samples, this result was consistent with previous studies (Buchanan & Smith, 1999b; Pasveer & Ellard, 1998). We can regard this result as evidence that this study’s results were valid.

On the other hand, the variance-covariance matrices of item scores, scree plots, alpha coefficients, and the results of one-factor EFA were almost the same between administration formats. These results support our hypothesis that administration formats do not change measured constructs, and we can suppose that regardless of administration formats, SDS accurately measures depression.

By MGCFA, we confirmed that the model with different thresholds between administration formats had a better fit to the data than the unrestricted model. Analyses of the threshold estimates (one sample *t* test and Cohen’s *d*) revealed that the estimates of the online administration format tended to be slightly smaller than those of paper-and-pencil administration. These results suggest that online administration lowers item thresholds slightly and that this decrease in threshold values causes the difference in mean structures between administration formats.

**Limitations and future directions for research**

Before the information gained during this study can be utilized in the future, there are some points that should be examined.

First, we confirmed that there was a difference in mean structures of SDS between paper-and-pencil and online administrations. However, it is not certain whether the all scales show similar trends between different administration formats. It may be that this trend is seen only in the scales which measure subtle and/or extreme constructs such as depression, divorce, sexual behavior, and so on. Therefore, we need to examine whether this trend would be seen in other scales.

Second, Japanese people tend more towards collectivism than Western people. Because the participants in this study were Japanese undergraduates, they may have felt embarrassed, that they were being judged, or have felt shy when responding to SDS. This may have heightened their social anxiety and social desirability when completing the paper-and-pencil formatted SDS, which may be one of the reasons for the differences in this study. Thus, before the results from this study can be generalized, we need to examine whether similar results would be seen in other countries.

Third, because the participants of the present study were adolescents, they may be in the midst of forming their own identity. If this is correct, they may be sensitive to evaluations from others, and which may have also been a factor in the differences in this study. In future studies, we need to conduct similar surveys with adults to examine whether similar results would be seen.

Finally, this time, we used data of different cohorts to examine research questions. Although we confirmed the administration formats’ effect on scales’ mean structures and got the implication of the cause, this cohort difference might have caused this study’s results. Therefore, in the future, we have to do the same research with the different data sets to examine whether similar results would be replicated.

**Conflicts of Interest**

The authors declare that there is no conflict of interest.

**Funding**

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

**References**

1. Angst, J. (1996). Comorbidity of mood disorders: A longitudinal prospective study. *British Journal of Psychiatry*, *168*(S30), 31–37. <https://doi.org/10.1192/S0007125000298383>
2. Buchanan, T. (2002). Online assessment: Desirable or dangerous? *Professional Psychology: Research and Practice*, *33*(2)*,* 148–154. <https://psycnet.apa.org/doi/10.1037/0735-7028.33.2.148>
3. Buchanan, T., Ali, T., Heffernan, T. M., Ling, J., Parrott, A. C., Rodgers, J., & Scholey, A. B. (2005). Nonequivalence of on-line and paper-and-pencil psychological tests: The case of the prospective memory questionnaire. *Behavior Research Methods*, *37*, 148–154. https://doi-org.ezproxy.tulips.tsukuba.ac.jp/10.3758/BF03206409
4. Buchanan, T., & Smith, J. L. (1999a). Research on the Internet: Validation of a World-Wide Web mediated personality scale. *Behavior Research Methods, Instruments, & Computers*, *31*(4), 565–571. https://doi.org/10.3758/BF03200736
5. Buchanan, T., & Smith, J. L. (1999b). Using the Internet for psychological research: Personality testing on the World Wide Web. *British Journal of Psychology*, *90*(1), 125–144. https://doi.org/10.1348/000712699161189
6. Carlbring, P., Brunt, S., Bohman, S., Austin, D., Richards, J., Öst, L.-G., & Andersson, G. (2007). Internet vs. paper and pencil administration of questionnaires commonly used in panic/agoraphobia research. *Computers in Human Behavior*, *23*(3), 1421–1434. https://doi.org/10.1016/j.chb.2005.05.002
7. Cohen, J. (1969). *Statistical power analysis for the behavioral sciences*. Academic Press.
8. Davis, R. N. (1999). Web-based administration of a personality questionnaire: Comparison with traditional methods. *Behavior Research Methods, Instruments, & Computers*, *31*, 572–577. https://doi.org/10.3758/BF03200737
9. Del Re, A. C. (2013). compute.es: Compute Effect Sizes. R package version 0.2-2. <http://cran.r-project.org/web/packages/compute.es>
10. Ettner, S. L., Frank, R. G., & Kessler, R. C. (1997). The impact of psychiatric disorders on labor market outcomes. *Industrial and Labor Relations Review*, *51*(1), 64–81. https://doi.org/10.1177%2F001979399705100105
11. Fukuda, K., & Kobayashi, S. (1973). A study on a self-rating depression scale. *Psychiatria et Neurologia Japonica*, *75*(10), 673–679. (In Japanese.)
12. Gangestad, S., & Snyder, M. (1985). ‘‘To carve nature at its joints’’: On the existence of discrete classes in personality. *Psychological Review*, *92*(3), 317–349. <https://psycnet.apa.org/doi/10.1037/0033-295X.92.3.317>
13. Gosling, S. D., Vazire, S., Srivastava, S., & John, O. P. (2004). Should we trust web-based studies?: A comparative analysis of six preconceptions about internet questionnaires. *American Psychologist*, *59*(2), 93–104. https://psycnet.apa.org/doi/10.1037/0003-066X.59.2.93
14. Hirose, E. I. (2000). Computerization of psychological testing: Review of recent studies. *Japanese Journal of Educational Psychology*, *48*(2), 235–246. (in Japanese)
15. Joinson, A. (1999). Social desirability, anonymity, and internet-based questionnaires. *Behavior Research Methods, Instruments, & Computers*, *31*, 433–438. <https://doi.org/10.3758/BF03200723>
16. Kawakami, N., & Koizumi, A. (1986). Validity of self-assessing depression scale in industry. *Sangyo Igaku* (*Japanese Journal of Industrial Health*), *28*(5), 360–361. (in Japanese) <https://doi.org/10.1539/joh1959.28.360>
17. Kessler, R. C., Berglund, P., Demler, O., Jin, R., Merikangas, K. R., & Walters, E. E. (2005). Lifetime prevalence and age-of-onset distributions of DSM–IV disorders in the National Comorbidity Survey Replication. *Archives of General Psychiatry*, *62*(6), 593–602. doi:10.1001/archpsyc.62.6.593
18. Kessler, R. C., Foster, C. L., Saunders, W. B., & Stang, P. E. (1995). Social consequences of psychiatric disorders, I: Educational attainment. *The American Journal of Psychiatry*, *152*(7), 1026–1032. https://psycnet.apa.org/doi/10.1176/ajp.152.7.1026
19. Kessler, R. C., Walters, E. E., & Forthofer, M. S. (1998). The social consequences of psychiatric disorders, III: Probability of marital stability. *The American Journal of Psychiatry*, *155*(8), 1092–1096. https://psycnet.apa.org/doi/10.1176/ajp.155.8.1092
20. Miura, A., & Kobayashi, T. (2016). To say, or not to say “Good-bye, Mr/Ms online survey panels.” *Journal of Media, Information and Communication*, *1*, 27–42.
21. Pasveer, K. A., & Ellard, J. H. (1998). The making of a personality inventory: Help from the WWW. *Behavior Research Methods, Instruments, & Computers*, *30*, 309–313. https://doi.org/10.3758/BF03200659
22. R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>.
23. Revelle, W. (2018). psych: Procedures for Personality and Psychological Research. Northwestern University, Evanston, Illinois, USA. https://CRAN.R-project.org/package=psych Version = 1.8.12.
24. Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. doi:10.18637/jss.v048.i02
25. Schmidt, W. C. (1997). World-Wide Web survey research: Benefits, potential problems, and solutions. *Behavior research methods, Instruments, & Computers*, *29*, 274–279. <https://doi.org/10.3758/BF03204826>
26. Schulenberg, S. E., & Yutrzenka, B. A. (1999). The equivalence of computerized and paper-and-pencil psychological instruments: Implications for measures of negative affect. *Behavior Research Methods, Instruments, & Computers*, *31*, 315–321. https://doi.org/10.3758/BF03207726
27. Shafer, A. B. (2006). Meta-analysis of the factor structures of four depression questionnaires: Beck, CES-D, Hamilton, and Zung. *Journal of Clinical Psychology*, *62*(1), 123–146. https://doi.org/10.1002/jclp.20213
28. Spek, V., Nyklíček, I., Cuijpers, P., & Pop, V. (2008). Internet administration of the Edinburgh Depression Scale. *Journal of Affective Disorders*, *106*(3), 301–305. https://doi.org/10.1016/j.jad.2007.07.003
29. Thorén, E. S., Andersson, G., & Lunner, T. (2012). The use of research questionnaires with hearing impaired adults: online vs. paper-and-pencil administration. *BMC ear, nose, and throat disorders*, *12*, 12. https://doi.org/10.1186/1472-6815-12-12
30. Umegaki, Y., & Todo, N. (2017). Psychometric properties of the Japanese CES-D, SDS, and PHQ-9 depression scales in university students. *Psychological Assessment*, *29*(3), 354–359. doi:10.1037/pas0000351
31. Weitzman, E. R. (2004). Poor mental health, depression, and associations with alcohol consumption, harm, and abuse in a national sample of young adults in college. *Journal of Nervous and Mental Disease, 192*(4), 269–277. doi:10.1097/01.nmd.0000120885.17362.94
32. Whitehead, L. C. (2007). Methodological and ethical issues in Internet-mediated research in the field of health: An integrated review of the literature. *Social Science & Medicine*, *65*(4), 782–791. https://doi.org/10.1016/j.socscimed.2007.03.005
33. Zung, W. W. K. (1965). A self-rating depression scale. *Archives of General Psychiatry*, *12*(1), 63–70. doi:10.1001/archpsyc.1965.01720310065008