Internet Interventions 5 (2016) 1–4

Contents lists available at ScienceDirect

Internet Interventions

journal homepage: www.elsevier.com/locate/invent

Interformat reliability of the patient health questionnaire: Validation of the computerized version of the PHQ-9

Doris Erbe [a,*,1], Hans-Christoph Eichert [a], Christian Rietz [a], David Ebert [b]

[a] University of Cologne, Germany
[b] University of Erlangen, Germany

ARTICLE INFO

Article history:
Received 19 January 2016
Received in revised form 3 June 2016
Accepted 21 June 2016
Available online 27 June 2016

Keywords:
PHQ-9
Interformat reliability
Depression
Psychometric
Questionnaire
Assessment
Computer
Internet

ABSTRACT

Background: Computerized versions of well-established measurements such as the PHQ-9 are widely used, but data on the comparability of psychometric properties are scarce.
Objective: Our objective was to compare the interformat reliability of the paper-and-pen version with a computerized version of the PHQ-9 in a clinical sample.
Methods: 130 participants with mental health disorders were recruited during psychotherapeutic treatment in a mental health clinic. In a crossover design, they all completed the PHQ-9 in both the computerized and paper-and-pen versions in randomized order.
Results: The internal consistency was comparable for the computer ($\alpha = 0.88$) and paper versions ($\alpha = 0.89$), and highly significant correlations were found between the formats ($r = 0.92$). PHQ-9 total scores were not significantly different between the paper and the computer delivered versions. There was a significant interaction effect between format and order of administration for the PHQ-9, indicating that the first administration delivered slightly higher scores.
Limitations: In order to reduce the required effort for the participants, we did not ask them to fill out anything but the PHQ-9 once in paper and once in computer version.
Conclusions: Our findings suggest that the PHQ-9 can be transferred to computerized use without affecting psychometric properties in a clinically meaningful way.

1. Introduction

Web-based administrations of questionnaires have various advantages over paper-and-pen assessments, such as an increased chance of avoiding missing data, the automated calculation of scores, and being able to save time and eliminate the risk of calculation errors (Andersson et al., 2008). It has been argued that Internet-based questionnaires could help increase the use of self-report measures in clinical practice because they might facilitate administration (Holländare et al., 2010). Also, computer-based screenings seem to be highly accepted by clients (Campbell et al., 2015; Weber et al., 2003).

In addition, Internet-delivered treatments of common mental disorders have become more and more common within the last couple decades. A large number of studies suggests that they result in clinically meaningful changes with effect sizes as large as those found for face-to-face-therapy, both for adults (Barak et al., 2008; Griffiths et al., 2010; Hedman and Lindefors, 2012; Riper et al., 2014; Saddichha et al., 2014) and for children & adolescents (Ebert et al., 2015). The increasing popularity of Internet-based interventions in the field of mental health is necessarily accompanied by online assessments of psychopathological symptomatology (Austin et al., 2006).

However, when transferring paper-and-pen psychometric questionnaires into electronic forms of administration, it is important to realize that this may influence their outcome (Buchanan, 2003). For instance, there have been findings that computer administrations may produce a certain level of disinhibition concerning topics such as alcohol consumption and risky sexual behaviors (Booth-Kewley et al., 2007). Also issues such as computer anxiety and familiarity with the medium need to be taken into account (Schulenberg and Yutrzenka, 2004). Although the vast majority of studies examining interformat reliability of depression measures suggest that paper-and-pen versions may be transferred to digital formats without losing diagnostic properties (Alfonsson et al., 2014), there are also a few results suggesting that higher values are obtained on the Becks Depression Inventory if it is administered over the Internet or on a computer (Carlbring et al., 2007; George et al., 1992).

The International Test Commission has developed a set of guidelines concerning equivalence between paper-and-pencil versions to be

* Corresponding author at: Psychologie und Psychotherapie in Heilpädagogik und Rehabilitation, Universität Köln, Humanwissenschaftliche Fakultät, Klosterstraße 79b, D - 50931 Köln, Germany.
E-mail address: Doris.Erbe@uni-koeln.de (D. Erbe).
[1] https://www.hf.uni-koeln.de/37325.

http://dx.doi.org/10.1016/j.invent.2016.06.006
2214-7829/

ensured by psychometric properties such as comparable reliabilities of both versions, correlations at the expected level from the reliability estimates, and comparable means and standard deviations (The International Test Commission, 2006). Many established and well-evaluated paper-and-pen assessment instruments for mental health have already been validated in electronic formats by studies in the last decade (e.g. Austin et al., 2006; Holländare et al., 2010; Vallejo et al., 2008).

A recent review on the interformat reliability of computer administered psychiatric measures identified 33 studies exploring 40 different symptom scales (Alfonsson et al., 2014). The authors claim that while some instruments (e.g. the BDI-II) have been investigated multiple times, a number of prominent instruments—such as the PHQ-9—have not been examined sufficiently in order to be able to recommend their computerized versions. Thus, additional high quality studies must be done.

The PHQ-9 is a widely used measure to assess depression severity within various settings (Manea et al., 2015). It has been translated into many languages and performs well in various cultures (Gilbody et al., 2007). Several studies have evaluated electronic administrations of the PHQ-9 such as a on a touch screen computer (Fann et al., 2009) or as a smartphone app (BinDhim et al., 2014) without exploring interformat reliability. The first study comparing paper-and-pen and electronic versions was published in 2013 and described a crossover design showing comparable psychometric properties on the PHQ-9 and six other measures completed on paper, computer, and iPhone (Bush et al., 2013). However, the small and nonclinical sample of 45 army soldiers calls for a replication of these results. The first study examining a larger sample exploring interformat reliability of the PHQ-9 was published recently (Spangenberg et al., 2015). However, in the Spangenberg study, only elderly primary care patients took part, with only 4.3% of them suffering from clinically relevant depression (Spangenberg et al., 2015). This study aims to evaluate the interformat reliability between paper and computer versions of the PHQ-9 in a clinical sample.

2. Methods

2.1. Participants and procedures

Participants were treated for a mental disorder in an inpatient routine mental health clinic in Germany and recruited between February and March of 2012. Psychotherapeutic treatment of participants was based on Cognitive Behavioral Therapy. Patients received one or two sessions of individual therapy and an average of four double sessions of group therapy per week. Interventions were supplemented with elements such as sports therapy, physiotherapy, art therapy, and medical treatment. Patients who recently started treatment or were about to finish their treatment and leave the clinic at the end of their stay were invited to participate in the study. 143 patients (79%) agreed to take part and provided full written informed consent.

In order to control for order effects, participants were randomized on which version, the paper or the computerized questionnaire, to complete first. Randomization was performed on an individual level. In order to reduce potential recall effects, the time between filling out the two versions was set at 24 h.

Ethical approval was provided by the ethical committee of the University of Marburg and the hospital review board.

2.2. Measure

The PHQ-9 is a widely used instrument and consists of nine items matching the Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition (DSM-IV) criteria for major depression. The criteria for Major Depressive Disorder have been minimally changed in the DSM-5, the most important change being that bereavement is no longer an exclusion criterion. PHQ-9 scores are not affected by this change since the questionnaire does not include an item on bereavement (Volker et al., 2015). Subjects are asked to rate each of the items on a scale of 0 to 3 on the basis of how much a symptom has bothered them over the last 2 weeks with 0 ("not at all") to 3 ("nearly every day"). PHQ-9 scores of 5, 10, 15, and 20 indicate mild, moderate, moderately severe, and severe depression severity. The PHQ-9 has excellent reliability and its construct validity has been proven by strong correlations between PHQ-9 scores and disability days, functional status, and symptom-related difficulty (Kroenke et al., 2001).

The computer program that was used to administer the PHQ-9 was a format used in routine mental health care at the cooperating clinic. Participants filled out this version on a standard Personal Computer belonging to the clinic, one of each standing in each group therapy room. They were scheduled for filling it out at a certain time when the room was not used so they could do it without anybody else being in the room. One question was presented per each page in the same order for every patient. The program made it impossible to submit answers without answering all items in order to avoid missing values. Respondents could not backtrack (and thus not change answers after submitting).

2.3. Statistical analyses

Statistical analysis was conducted using SPSS version 23. Cronbach alpha coefficients were used to estimate internal consistency, and the correlations between Internet and paper-based questionnaires were calculated with Pearson correlations. Significance testing of differences in questionnaire administration format (paper/computer) and order (paper first/computer first) was done with a 2×2 Analysis of Variance (ANOVA). A significance level of 0.05 (two-sided) was used for all analyses. Effect sizes (Cohen's d) were calculated by dividing the difference between scores by the pooled standard deviation (Lenhard and Lenhard, 2015). We also evaluated whether the one-factor structure found before for the original PHQ-9 (Cameron et al., 2008) holds across both the online- and the paper/pen assessment of the PHQ-9. Measurement Invariance was tested by conducting two independent confirmatory factor analyses with uncorrelated residuals and the chi-square statistic and the chi-square/df value were used as an indicator for the fit of the model. In general, a chi-square/df ratio of approximately 2:1 or 3:1 is considered an acceptable fit to the data (Carmines and McIver, 1981) and measurement invariance was concluded when both independent models showed at least an acceptable fit to the data.

3. Results

Out of 143 patients who provided informed consent, 130 filled out questionnaires on both formats of administration. The mean age of participants was 43.46 years (SD = 12.56, Range = 18–71), 66 (50.8%) were women, and 64 were men (49.2%). Out of the 130 patients that filled out both questionnaires, 127 were diagnosed with depression (n = 76 with ICD-10 F 32, n = 51 with ICD-10 F 33), and n = 3 did not have a diagnosis of depression but a different diagnosis (tinnitus, adjustment disorder, or agoraphobia with panic disorder). Nine of the participants diagnosed with depression had an additional diagnosis of dysthymia (ICD-10 F 34.1). 47 participants had one mental disorder diagnosis, n = 51 had two, n = 24 had three, n = 5 had four, and n = 2 had five mental health disorder diagnoses. Mean number of mental health disorders per participant were n = 2. Apart from depression/mood disorders (ICD-10 F32-F33, n = 127, 98% of participants), the most frequent other diagnoses were somatoform disorders (ICD-10 F 45, n = 40, 30.8% of participants), anxiety disorders (ICD-10 F 40–41, n = 24, 18.5% of participants), and disorders of adult personality and behavior (ICD-10 F60–F69, n = 21, 16.1% of participants).

Out of 130 returned paper and pen questionnaires, 128 did not have any missings. The computerized version did not allow missings, hence

Table 1
Means (SD), main effects, and interaction effects.

	Order group	Computer	Paper	Main effects		Interaction
		Mean (SD)	Mean (SD)	Format F (p)	Order of administration F (p)	F (p)
PHQ-9	Paper first	11.22 (5.94)	11.62 (6.42)			
	Computer first	11.15 (6.49)	10.64 (6.61)			
	Total	11.19 (6.20)	11.13 (6.51)	0.06 (p = 0.805)	0.22 (p = 0.638)	4.06 (p = 0.046)

there were no missing data. Only complete questionnaires were included into statistical analysis.

3.1. Internal consistency

The questionnaires' internal consistencies (Cronbach's alpha) was $\alpha = 0.88$ for the computerized version and $\alpha = 0.89$ for the paper version. This satisfying result was supported by a confirmatory factor analysis testing the one factor solution (computer version: chi-square = 103.908, df = 27, $p < 0.001$; paper version: chi-square = 61.479, df = 27, $p < 0.001$).

3.2. Correlational analyses

Pearson correlation for PHQ-9 was $r = 0.92$ between paper and computer versions. The correlation was highly significant ($p < 0.001$).

3.3. Mean differences

Means and standard deviations for both measures and formats are presented in Table 1. There was no statistical difference in the mean score on the PHQ-9 between the paper-and-pen and the computerized versions. There was also no significant main effect for administration format nor for administration order.

A significant interaction effect between format and order of administration was found. The second administration had a significantly lower result compared to the first administration. The effect size for the difference was small (Cohen's d = 0.07*).

4. Discussion

The aim of this study was to examine the interformat reliability between the paper-and-pen and the computerized versions of the PHQ-9. Results indicated a high internal consistency concerning Cronbach's alpha in both formats. The correlation between computer and paper format was high, suggesting high reliability. There was no significant difference between mean scores of the computer and paper versions, but a significant (format) × (order of administration) interaction was found representing the decrease of PHQ-9 scores from first to second administration.

Results of the present study are in line with the studies that tested the interformat reliability of the PHQ-9 in non-clinical samples of army soldiers (Bush et al., 2013) and older adults (Spangenberg et al., 2015) which also found high internal consistency of the computer version, high correlations with the paper version, and no significant difference between formats. Spangenberg et al. (2015) found a significant (format) × (order of administration) interaction just as was found in our study, yet in our study, the effect was not higher in any of the order conditions.

For the BDI-II, a similar interaction effect has been found (Holländare et al., 2010). A possible explanation for this finding is that administering a depression questionnaire in any format may have a small effect on patients' self-assessment of depressive symptoms, such that in the second administration they consider their symptomatology as less severe. An alternative explanation—at least for our study—may be that since participants were being treated in a mental health clinic and receiving intense psychotherapeutic inpatient treatment, the decrease in PHQ-9 scores from first to second administration, which was administered 24 h later, might be actually caused by clinical improvement of depression due to the treatment.

However, in our study, the interaction effect did not seem to be an effect caused by administration format.

This study has the following limitations. First, 13 patients (9%) did not return the paper-and-pen version. Thus, we cannot rule out a potential bias of the results due to missing data. Second, we did not measure computer anxiety, computer knowledge, or preference of format. Thus, we cannot make any conclusions whether the interformat reliability of the PHQ-9 varies as a function of these variables. Third, in the present study we only focused on the interformat reliability and our design did not consider other tests of psychometric properties such as the test-retest reliability (as in Bush et al., 2013). Fourth, the inpatient clinical setting – while yielding the advantage of offering a comparably large clinical sample - may also reduce the external validity of our study. Although patients filled out the computer version in a room by themselves and were allowed to fill out the paper version wherever they wanted, the setting was in a certain way controlled and substantially differs from the setting in studies where disinhibition effects have been found (such as Booth-Kewley et al., 2007). If participants complete an internet questionnaire at home, on the participant's own device and without knowing the researchers in person, the perceived anonymity might be higher and the setting more naturalistic than in our study. As such, possibly our results may have limited validity for settings outside a clinic.

5. Conclusion

Our findings suggest that the PHQ-9 can be transferred to computerized use without a change in psychometric properties.

Acknowledgements

Our special thanks go to the Schön Klinik Bad Arolsen for making this study possible. The (mostly technical and practical) support of the Schön Klinik Bad Arolsen did not influence the design, implementation, or results. We claim no conflict of interests.

References

Alfonsson, S., Maathz, P., Hursti, T., 2014. Interformat reliability of digital psychiatric self-report questionnaires: a systematic review. J. Med. Internet Res. 16 (12), e268. http://dx.doi.org/10.2196/jmir.3395.

Andersson, G., Ritterband, L.M., Carlbring, P., 2008. Primer for the assessment, diagnosis and delivery of Internet interventions for (mainly) panic disorder. Lessons learned from our research groups. Clin. Psychol. 12 (1), 1–8. http://dx.doi.org/10.1080/13284200802069027.

Austin, D.W., Carlbring, P., Richards, J.C., 2006. Internet administration of three commonly used questionnaires in panic research: equivalence to paper administration in Australian and Swedish samples of people with panic disorder. Int. J. Test. 6 (1), 25–39. http://dx.doi.org/10.1207/s15327574ijt0601.

Barak, A., Hen, L., Boniel-Nissim, M., Shapira, N., 2008. A comprehensive review and a meta-analysis of the effectiveness of internet-based psychotherapeutic interventions. J. Technol. Hum. Serv. 26 (2), 109–160. http://dx.doi.org/10.1080/15228830802094429.

BinDhim, N.F., Shaman, A.M., Trevena, L., Basyouni, M.H., Pont, L.G., Alhawassi, T.M., 2014. Depression screening via a smartphone app: cross-country user characteristics and feasibility. J. Am. Med. Inform. Assoc. 29–34 http://dx.doi.org/10.1136/amiajnl-2014-002840.

Booth-Kewley, S., Larson, G.E., Miyoshi, D.K., 2007. Social desirability effects on computerized and paper-and-pencil questionnaires. Comput. Hum. Behav. 23 (1), 463–477. http://dx.doi.org/10.1016/j.chb.2004.10.020.

Buchanan, T., 2003. Internet-based questionnaire assessment: appropriate use in clinical contexts. Cogn. Behav. Ther. 32 (3), 100–109. http://dx.doi.org/10.1080/16506070310000957.

Bush, N.E., Skopp, N., Smolenski, D., Crumpton, R., Fairall, J., 2013. Behavioral screening measures delivered with a smartphone app: psychometric properties and user preference. J. Nerv. Ment. Dis. 201 (11), 991–995. http://dx.doi.org/10.1097/NMD.0000000000000039.

Cameron, I.M., Crawford, J.R., Lawton, K., Reid, I.C., 2008. Psychometric comparison of PHQ-9 and HADS for measuring depression severity in primary care. Br. J. Gen. Pract. 58, 32–37. http://dx.doi.org/10.3399/bjgp08X263794.

Campbell, N., Ali, F., Finlay, A.Y., Salek, S.S., 2015. Equivalence of electronic and paper-based patient-reported outcome measures. Qual. Life Res. 24 (8), 1949–1961. http://dx.doi.org/10.1007/s11136-015-0937-3.

Carlbring, P., Brunt, S., Bohman, S., Austin, D., Richards, J., Andersson, G., 2007. Computers in Internet vs. paper and pencil administration of questionnaires commonly used in panic/agoraphobia research. Comput. Hum. Behav. 23 (3), 1421–1434. http://dx.doi.org/10.1016/j.chb.2005.05.002.

Carmines, E.G., McIver, J.P., 1981. Analyzing models with unobserved variables. In: Bohrnstedt, G.W., Borgatta, E.F. (Eds.), Social Measurement: Current Issues. Sage, Beverly Hills.

Ebert, D.D., Zarski, A.-C., Christensen, H., Stikkelbroek, Y., Cuijpers, P., Berking, M., Riper, H., 2015. Internet and computer-based cognitive behavioral therapy for anxiety and depression in youth: a meta-analysis of randomized controlled outcome trials. PLoS One 10 (3), e0119895. http://dx.doi.org/10.1371/journal.pone.0119895 (72).

Fann, J.R., Berry, D.L., Wolpin, S., Austin-Seymour, M., Bush, N., Halpenny, B., ... McCorkle, R., 2009. Depression screening using the patient health questionnaire-9 administered on a touch screen computer. Psycho-Oncology 18 (1), 14–22. http://dx.doi.org/10.1002/pon.1368.

George, C.E., Lankford, J.S., Wilson, S.E., 1992. The effects of computerized versus paper-and-pencil administration on measures of negative affect. Comput. Hum. Behav. 8 (2–3), 203–209.

Gilbody, S., Richards, D., Brealey, S., Hewitt, C., 2007. Screening for depression in medical settings with the patient health questionnaire (PHQ): a diagnostic meta-analysis. J. Gen. Intern. Med. 22 (11), 1596–1602. http://dx.doi.org/10.1007/s11606-007-0333-y.

Griffiths, K.M., Farrer, L., Christensen, H., 2010. The efficacy of internet interventions for depression and anxiety disorders: a review of randomised controlled trials. Med. J. Aust. 192 (11 Suppl.), S4–S11 (http://doi.org/gri10844_fm [pii]).

Hedman, E., Lindefors, N., 2012. Cognitive behavior therapy via the Internet : a systematic review of applications, clinical efficacy and cost – effectiveness. Expert Rev. Pharmacoecon. Outcomes Res. 12 (6), 745–764.

Holländare, F., Andersson, G., Engström, I., 2010. A comparison of psychometric properties between internet and paper versions of two depression instruments (BDI-II and MADRS-S) administered to clinic patients. J. Med. Internet Res. 12 (5), 1–14. http://dx.doi.org/10.2196/jmir.1392.

Kroenke, K., Spitzer, R.L., Williams, J.B.W., 2001. The PHQ-9: validity of a brief depression severity measure. J. Gen. Intern. Med. 16 (9), 606–613. http://dx.doi.org/10.1046/j.1525-1497.2001.016009606.x.

Lenhard, W., Lenhard, A., 2015. Berechnung von Effektstärken. (Retrieved from http://www.psychometrica.de/effektstaerke.html).

Manea, L., Gilbody, S., McMillan, D., 2015. A diagnostic meta-analysis of the patient health questionnaire-9 (PHQ-9) algorithm scoring method as a screen for depression. Gen. Hosp. Psychiatry 37 (1), 67–75. http://dx.doi.org/10.1016/j.genhosppsych.2014.09.009.

Riper, H., Blankers, M., Hadiwijaya, H., Cunningham, J., Clarke, S., Wiers, R., ... Cuijpers, P., 2014. Effectiveness of guided and unguided low-intensity internet interventions for adult alcohol misuse: a meta-analysis. PLoS One 9 (6), e9991. http://dx.doi.org/10.1371/journal.pone.0099912.

Saddichha, S., Al-Desouki, M., Lamia, A., Linden, I.a., Krausz, M., 2014. Online interventions for depression and anxiety – a systematic review. Health Psychol. Behav. Med. 2 (1), 841–881. http://dx.doi.org/10.1080/21642850.2014.945934.

Schulenberg, S.E., Yutrzenka, B.A., 2004. Ethical issues in the use of computerized assessment. Comput. Hum. Behav. 20 (4), 477–490. http://dx.doi.org/10.1016/j.chb.2003.10.006.

Spangenberg, L., Glaesmer, H., Boecker, M., Forkmann, T., 2015. Differences in patient health questionnaire and Aachen depression item bank scores between tablet versus paper-and-pencil administration. Qual. Life Res. 24 (12), 3023–3032. http://dx.doi.org/10.1007/s11136-015-1040-5.

The International Test Commission, 2006. International guidelines on computer-based and internet-delivered testing. Int. J. Test. 6 (2), 143–171. http://dx.doi.org/10.1207/s15327574ijt0602.

Vallejo, M.a., Mañanes, G., Isabel Comeche, M., Díaz, M.I., 2008. Comparison between administration via Internet and paper-and-pencil administration of two clinical instruments: SCL-90-R and GHQ-28. J. Behav. Ther. Exp. Psychiatry 39 (3), 201–208. http://dx.doi.org/10.1016/j.jbtep.2007.04.001.

Volker, D., Zijlstra-Vlasveld, M.C., Brouwers, E.P.M., Homans, W.a., Emons, W.H.M., van der Feltz-Cornelis, C.M., 2015. Validation of the patient health questionnaire-9 for major depressive disorder in the occupational health setting. J. Occup. Rehabil. 1–8 http://dx.doi.org/10.1007/s10926-015-9607-0.

Weber, B., Schneider, B., Fritze, J., Gille, B., Hornung, S., Kühner, T., Maurer, K., 2003. Acceptance of computerized compared to paper-and-pencil assessment in psychiatric inpatients. Comput. Hum. Behav. 19 (1), 81–93. http://dx.doi.org/10.1016/S0747-5632(02)00012-2.