Journal of Affective Disorders Reports 5 (2021) 100168

Contents lists available at ScienceDirect

Journal of Affective Disorders Reports

journal homepage: www.sciencedirect.com/journal/journal-of-affective-disorders-reports

Brief Report

Impact of measurement frequency on self-reported depressive symptoms: An experimental study in a clinical setting

Nicole Geschwind [a,*], Martijn van Teffelen [a,b], Elin Hammarberg [a], Arnoud Arntz [c], Marcus J. H. Huibers [d], Fritz Renner [e]

[a] Department of Clinical Psychological Science, Maastricht University, Netherlands
[b] Met GGZ, Maastricht, Netherlands
[c] Department of Clinical Psychology, University of Amsterdam, Netherlands
[d] Department of Clinical Psychology, VU University, Amsterdam, Netherlands
[e] Institute of Psychology, Freiburg University, Germany

A B S T R A C T

Background: Previous research suggests a relationship between measurement frequency of selfreported depressive symptoms and change in depressive symptom scores for the Beck Depression Inventory II (BDI-II). The goal of the current study was to investigate the differential effects of weekly and monthly completion of the BDI-II and Quick Inventory of Depressive Symptomatology self-report (QIDS-SR).

Methods: Seventy individuals diagnosed with major depressive disorder (MDD) waiting for treatment were randomly assigned to either completing BDI-II weekly, BDI-II monthly, QIDS-SR weekly, or QIDS-SR monthly for a duration of nine weeks. After nine weeks participants also completed the Zung depression scale once. Mixed multilevel regression modelling and Bayesian Statistical Analysis were used to test the relationship between the measurement frequency and depression scores, and to compare scores of the repeatedly completed instruments with the instrument completed only in week nine.

Results: Measurement frequency was not related to BDI-II, QIDS-SR or Zung scores. However, depression scores declined in the weekly and monthly QIDS-SR (but not BDI-II) conditions, while Bayesian analyses indicated moderate support for equal depression scores on the Zung SDS.

Limitations: Lack of a clinician-rated depression scale at week nine in addition to the self-report measure.

Conclusions: In contrast to previous studies in non-clinical samples, our findings suggest that measurement frequency does not have an impact on scores of the BDI-II. Implications for clinical studies monitoring depressive symptom scores with self-report scales are discussed.

Major Depressive Disorder (MDD) is a common mood disorder and a major contributor to the overall disease burden worldwide (Ferrari et al., 2013). To detect a change in depressive symptoms, effective symptom monitoring during treatment is of key importance. Monitoring requires repeated measurements, which may produce a form of measurement error called retest effects. This study aimed to test potential measurement error in two commonly used self-report measurements for depressive symptoms: the Beck Depression Inventory-II (BDI-II; Beck et al., 1996) and the Quick Inventory of Depressive Symptomatology self-report (QIDS-SR; Rush et al., 2003).

Despite its reported high psychometric qualities, the BDI has been suspected of showing retest effects (Wang and Gorenstein, 2013). One study randomly assigned 237 college students into one of three conditions with different completion frequency: completing the BDI-II bimonthly, monthly, or weekly (Longwell and Truax 2005). In week five, the participants in the weekly condition had significantly lower BDI-II scores compared to the other conditions. This difference was not found on an instrument measuring depressive symptoms that was only completed once, the Zung SDS. Similar studies have also shown a decrease in depressive symptoms following repeated measurements within a short time interval (Hatzenbuehler et al., 1983; Renner et al., 2016; Richter et al., 1997; Wolfner et al., 1998). If repeated completion of the BDI-II results in retest effects, effect sizes reported in earlier treatment studies may have been structurally overestimated.

Another relevant question is whether retest effects are specific to the BDI-II or generalize to other self-report measures of depression. Moreover, previous studies (Longwell and Truax 2005) relied on non-clinical samples. The present study, therefore, sought to investigate potential retest effects associated with repeated completion of both the BDI-II and QIDS-SR; another frequently used self-report measurement of depression, in a sample of individuals diagnosed with MDD waiting for treatment. We expected a completion frequency by time interaction, where the standardized depression score of the repeatedly completed instrument will be lower in the weekly compared to the monthly completion

* Corresponding author.
E-mail addresses: Nicole.geschwind@maastrichtuniversity.nl (N. Geschwind), fritz.renner@psychologie.uni-freiburg.de (F. Renner).

https://doi.org/10.1016/j.jadr.2021.100168
Received 26 November 2020; Received in revised form 5 May 2021; Accepted 25 May 2021
Available online 6 June 2021

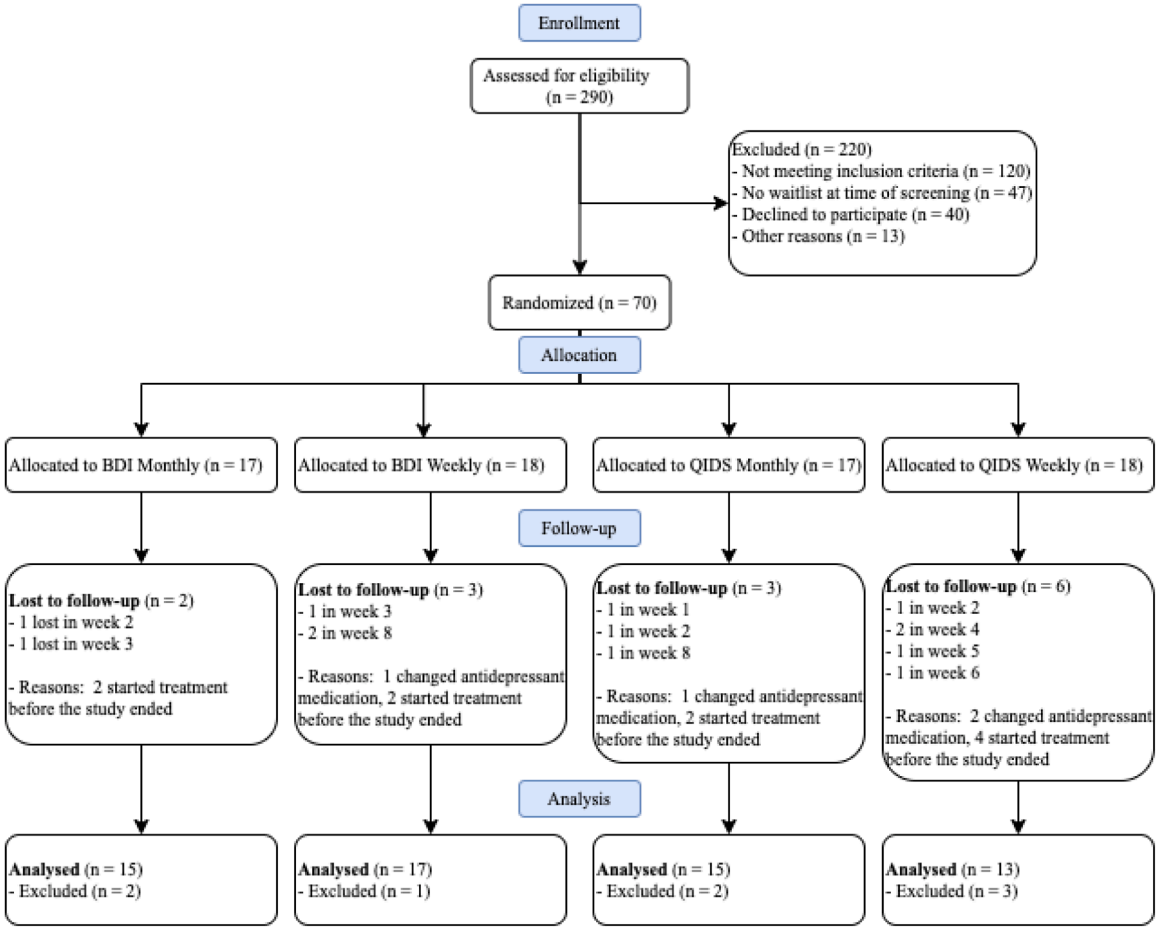

Fig. 1. Participant flowchart.

condition (hypothesis 1). To explore the instrument specificity of the above effect, we also examined the three-way interaction of frequency by instrument (BDI-II, QIDS-SR) by time. We did not expect differences between these two instruments (hypothesis 2). We hypothesized that higher instrument completion frequency was not related to lower depression scores on a self-report depression instrument completed only in week nine (the Zung SDS) (hypothesis 3).

1. Methods

1.1. Participants

Seventy individuals currently diagnosed with MDD in waitlist for treatment were recruited from the mood disorder unit of the center Virenze-RIAGG Maastricht. A priori power calculation indicated that 68 participants were needed to detect an effect of $d = 0.8$ (as found in Longwell and Truax, 2005) with $\alpha = 0.05$ and $\beta = 0.9$. Exclusion criteria were current psychotic disorder, substance dependency, bipolar disorder, or an IQ lower than 80 based on previous health care reports, the clinician's impression, and education level. Moreover, individuals partaking in any treatments were not eligible. For demographic characteristics, see Appendix A.

1.2. Materials

1.2.1. The Structured Clinical Interview for DSM-IV axis I Disorders (SCID)

The SCID is considered the gold-standard for semi-structured assessment of mental disorders as classified in the DSM-IV (American Psychiatric Association, 1994; First et al., 2012). The SCID for DSM-IV axis I disorders was used for the initial screening of inclusion and exclusion criteria, because data was collected before introduction of the DSM-5. However, the core symptom criteria applied to the diagnosis of MDD have not changed from DSM-IV to DSM-5.

1.2.2. Beck's Depression Inventory II (BDI-II)

The BDI-II is a 21-item self-report instrument that assesses depressive symptoms (Beck et al., 1961). In our sample, internal consistency (McDonald's omega) was excellent ($\omega = 0.93$), one-week test-retest reliability was excellent ($r = 0.93$) and one-month test-retest reliability was good ($r = 0.82$).

1.2.3. The Quick Inventory of Depressive Symptomatology Self-Report (QIDS-SR)

The QIDS-SR is a 16-item self-report instrument that assesses depressive symptoms (Rush et al., 2003). In this sample, internal consistency (McDonald's omega) was moderate ($\omega = 0.605$), one-week test-retest reliability was moderate ($r = 0.56$) and one-month test-retest reliability was moderate ($r = 0.54$).

1.2.4. The Zung Self-Rating Scale for Depression (Zung SDS)

The Zung SDS is a self-report measure that measures depressive symptoms (Zung, 1965). Psychometric studies report good discriminant and concurrent validity (Thurber et al., 2002; Schafer, 2006). Reliability has been reported to be sufficient (Gabrys and Peters, 1985). In our sample, internal consistency (McDonald's omega) was good ($\omega = 0.839$).

1.3. Procedure

The Ethical Committee - Psychology of the Faculty of Psychology and Neuroscience (ECP128 07_05_2013), Maastricht University approved the study. Individuals were screened for in- and exclusion criteria using the SCID-I.

Eligible individuals signed informed consent, filled in a demographics questionnaire, and entered a waitlist-period during which they were randomly assigned to one of four conditions (see Fig. 1). The 10 participants who dropped out before week five, before the time of the second measurement occasion for the monthly conditions, were excluded a priori. The main source of attrition was starting treatment. Individuals who dropped out at a later stage were included in the analyses, hence, the final $N = 60$. The duration of the study was nine weeks. Participants in the monthly conditions logged into the study's questionnaire portal once every week, as a control for effort and attention. All participants completed the Zung SDS at week nine only.

1.4. Statistical analysis

Mixed regression was used to examine the impact of weekly and monthly questionnaire completion (hypothesis 1 and 2). BDI-II and QIDS-SR scores were Z-transformed using baseline means and standard deviations pooled per measurement instrument to compare scores across different instruments. Effect-coding was used for completion frequency (i.e., monthly = −1, weekly = 1) and time (i.e., baseline = −1, week nine = 1). For the mixed models, group, time (i.e., week one and week nine), and group by time variables were dummy-coded and entered as fixed effects in the model. The random and fixed parts of the model were adjusted to improve model fit, identifiability, and parsimony (e.g., see Baayen et al., 2008). Bayesian ANOVA was used to enable null-hypothesis, or equality testing (hypothesis 3). Bayes' factors were calculated to examine the equality of groups on Zung SDS scores at week nine.

2. Results

Participant characteristics and variable means are available in Appendix A. Missing value analysis rejected the hypothesis that missing values were not missing completely at random across time points for BDI-II scores χ^2 (3, $N = 32$) = 1.09, $p = .78$ and QIDS-SR scores χ^2 (5, $N = 27$) = 5.41, $p = .37$. A first-order autoregressive model (AR1) was the best fitting structure for the repeated part of the model, see Appendix B. A random intercept or slope was not included in the model.

To test hypothesis 1, that weekly completion of self-report depression questionnaires results in a stronger decrease in self-reported depression compared to monthly completion, we applied mixed regression modeling to predict depression scores at week nine. The interaction between completion frequency and time was not significant ($F(1, 54.57) = 0.19$, $p = .66$). The main effect of time was significant ($F(1, 54.57) = 7.52$, $p = .008$, $d = -0.72$), whereas the main effect of completion frequency was not ($F(1, 57.79) = 0.03$, $p = .86$). A follow-up Bayesian independent samples *t*-test of completion frequency on post-test depression scores demonstrated that it is 3.64 times more likely that a frequency effect is absent than present ($B_{10} = 0.28$). This indicates that self-reported depression scores decreased over time, irrespective of the condition.

To examine instrument specificity on the effects of time and completion frequency, another mixed regression model was run, test hypothesis 2. Next to completion frequency and time, instrument (i.e., QIDS-SR or BDI-II) and their respective two-way and three-way interaction variables were entered in the model. The three-way interaction of instrument by completion frequency by time was not significant ($F(1, 52.19) = 0.58$, $p = .45$). To interpret the two-way interaction, the three-way interaction term was subsequently removed from the model. The effects of instrument by completion frequency ($F(1, 55.46) = 5.08$ $p = .03$), instrument by time ($F(1, 53.31) = 5.33$, $p = .03$) and time ($F(1, 53.31) = 8.94$, $p = .004$) were significant. Other effects in the model were not significant (p's > 0.23). Simple effects analysis revealed that for QIDS-SR completers time was significant ($F(1, 25.96) = 8.22$, $p = .008$, $d = -1.13$) and completion frequency was not significant ($F(1, 26.31) = 2.10$, $p = .16$). For BDI-II completers, time ($F(1, 28.43) = 0.70$, $p = .41$) and completion frequency ($F(1, 29.14) = 3.82$, $p = .06$) were not significant. As visualized in Fig. 2, findings indicate that the decrease in depression scores over time was specific for QIDS-SR completers, irrespective of completion frequency.

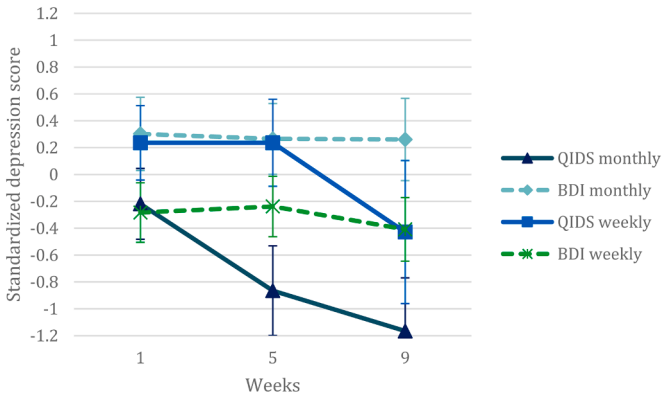

Fig. 2. Standardized depression scores over time per condition. *Note.* QIDS = Quick Inventory of Depressive Symptomatology self-report; BDI = Beck Depression Inventory.

Finally, a Bayesian ANOVA was performed to test the null-hypothesis, hypothesis 3, that instrument completion frequency was not related to self-reported depression scores after nine weeks on the self-report instrument that was not repeatedly completed as part of the study, the Zung SDS. The model's Bayes factor of $B_{10} = 0.14$ indicates moderate evidence in favor of the null hypothesis. More precisely, this means that for the Zung SDS score, the data is $1/B_{10} = 7{,}14$ times more likely to have occurred under the null-hypothesis than under the alternative hypothesis. Findings thus suggest that there is no difference in depressive symptom scores irrespective of condition when measured with an instrument that was not repeatedly completed.

3. Discussion

This study examined the differential effects of weekly and monthly completion of the BDI-II and QIDS-SR. Based on the findings of previous studies, we predicted that weekly completion would lead to lower scores on both questionnaires, compared to monthly completion. Conversely, we predicted that completion of an unfamiliar depression self-report instruments at the end of the study would not differ between experimental conditions. Contrary to our hypotheses, the main findings show that self-reported depression scores were not related to completion frequency on either measure. The QIDS-SR showed a decline in depression scores irrespective of measurement frequency, in both the weekly and monthly conditions.

The fact that depression scores were unaffected by completion frequency of the BDI-II contradicts several studies that have demonstrated a retest effect (Hatzenbuehler et al., 1983; Wolfner et al., 1998; Richter et al., 1997). One previous study reported a decrease in weekly BDI-II scores during a 6 to 24 weeks waiting list period in patients with chronic MDD (Renner et al., 2016). However, this study was not explicitly designed to test retest effects. Richter et al. (1997) used a mixed clinical sample and did not find any retest effect in a subgroup of patients with MDD. Longwell and Truax (2005) had a similar design to the current study and found retest effects on the BDI-II. However, again the main difference appears to be the use of a non-clinical sample as well as the exclusion of participants who were severely depressed (i.e., with

Table 1
Demographic characteristics and descriptive statistics of study variables.

	BDI-II weekly(n = 17)	BDI-II monthly(n = 15)	QIDS-SR weekly(n = 13)	QIDS-SR monthly(n = 15)
Age in years, mean (S.D.)	42.35 (13.29)	38.67 (12.82)	41.69 (14.04)	38.67 (15.62)
Female gender, n (%)	12 (70.6)	9 (60.0)	7 (53.8)	11 (73.3)
Education, n (%)				
Low	2 (11.8)	6 (40.0)	4 (30.8)	6 (40.0)
Middle	12 (70.6)	8 (53.3)	7 (53.8)	8 (53.3)
High	3 (17.6)	1 (6.7)	2 (15.4)	1 (6.7)
Partner, yes, n (%)	10 (58.8)	6 (40.0)	8 (61.5)	10 (66.7)
Employment, yes, n (%)	6 (35.3)	4 (26.7)	4 (30.8)	5 (33.3)
BDI-II/QIDS-SR score W1	35.33 (10.83)	41.71 (13.76)	17.55 (3.30)	16.15 (3.13)
BDI-II/QIDS-SR score W5	35.53 (12.02)	40.93 (13.25)	17.91 (3.91)	14.38 (4.01)
BDI-II/QIDS-SR score W9	32.80 (11.52)	41.21 (14.40)	15.64 (5.46)	12.92 (4.46)
Zung SDS score W9	56.47 (9.00)	58.50 (8.31)	58.55 (6.25)	55.93 (8.56)

Note. There were no significant differences between the conditions on any of these variables.

Table 2
Model comparison for the random part structure.

Model structure	−2LL	#par	Sign. worse?
Unstructured	320.40	11	–
AR1	326.89	10	No
CS	326.89	10	No
Scaled identity	343.22	9	Yes

Note. −2LL = −2 log likelihood level. #par = number of parameters in the model. Selected model is shown in bold. Each model is compared to its less structured predecessor.

BDI-II scores of 25 or more).

For the QIDS-SR, more research is needed to interpret the finding that depression scores declined over time irrespective of measurement frequency. Future studies could help clarify the existence of and reasons for potential measurement artifacts by interviewing the participants about their process after filling in the self-report questionnaires.

The present study has a few limitations. First, the study fully relied on self-report measurements. It would have been beneficial to include a clinician-rated measures to aid the interpretation of the time effects found in the QIDS-SR. Second, due to dropout, our sample was slightly smaller than the power calculation had required, limiting power for finding small effects.

In conclusion, we found no evidence for retest effects of commonly used self-report depression scales. If replicated, these findings suggest that it is unlikely that treatment effects of previous studies were affected by retest effects. To our knowledge, no other study experimentally examined the role of completion frequency in self-reported depression in patients with MDD. Future studies should continue to examine the QIDS to identify potential reasons for a time effect during repeated measurements.

Declaration of Competing Interest

None.

Funding sources

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. This research project was facilitated and supported by Maastricht University.

Appendix A

Table 1

Appendix B

Table 2

References

American Psychiatric Association, 1994. Diagnostic and Statistical Manual of Mental Disorders (4th ed.). Author, Washington, DC.

Baayen, R.H., Davidson, D.J., Bates, D.M., 2008. Mixed-effects modeling with crossed random effects for subjects and items. J Mem Lang 59 (4), 390–412. https://doi.org/10.1016/j.jml.2007.12.005.

Beck, A., Ward, C., Mendelson, M., Mock, J., Erbaugh, J., 1961. An inventory for measuring depression. Arch. Gen. Psychiatry 4 (6), 561–571. https://doi.org/10.1001/archpsyc.1961.01710120031004.

Beck, A., Steer, R., Brown, G., 1996. Manual For the BDI-II. Psychological Corporation, San Antonio, TX.

Ferrari, A.J., Charlson, F.J., Norman, R.E., Patten, S.B., Freedman, G., Murray, C.J.L., Vos, T., Whiteford, H.A, 2013. Burden of depressive disorders by country, sex, age, and year: findings from the Global Burden of Disease Study 2010. PLoS Med 10 (11). https://doi.org/10.1371/journal.pmed.1001547.

Gabrys, J.B., Peters, K., 1985. Reliability, discriminant and predictive validity of the Zung Self-Rating Depression Scale. Psychol Rep 57 (3f), 1091–1096. https://doi.org/10.2466/pr0.1985.57.3f.1091.

Hatzenbuehler, L.C., Parpal, M., Matthews, L., 1983. Classifying college students as depressed or nondepressed using the Beck Depression Inventory: an empirical analysis. J Consult Clin Psychol 51 (3), 360–366. https://doi.org/10.1037/0022-006X.51.3.360.

Longwell, B.T., Truax, P., 2005. The differential effects of weekly, monthly, and bimonthly administrations of the Beck Depression Inventory-II: psychometric properties and clinical implications. Behav Ther 36 (3), 265–275. https://doi.org/10.1016/S0005-7894(05)80075-9.

Renner, F., Arntz, A., Peeters, F.P., Lobbestael, J., Huibers, M.J., 2016. Schema therapy for chronic depression: results of a multiple single case series. J Behav Ther Exp Psychiatry 51, 66–73. https://doi.org/10.1016/j.jbtep.2015.12.001.

Richter, P., Werner, J., Bastine, R., Heerlein, A., Kick, H., Sauer, H., 1997. Measuring treatment outcome by the Beck Depression Inventory. Psychopathology 30 (4), 234–240. https://doi.org/10.1159/000285052.

Rush, A.J., Trivedi, M.H., Ibrahim, H.M., Carmody, T.J., Arnow, B., Klein, D.N., Markowitz, J.C., Ninan, P.T., Kornstein, S., Manber, R., Thase, M.E., Kocsis, J.H., Keller, M.B., 2003. The 16-Item Quick Inventory of Depressive Symptomatology (QIDS), Clinician Rating (QIDS-C), and Self-Report (QIDS-SR): a psychometric evaluation in patients with chronic major depression. Biol. Psychiatry 54 (5), 573–583. https://doi.org/10.1016/s0006-3223(02)01866-8.

Thurber, S., Snow, M., Honts, C.R., 2002. The Zung self-rating depression scale convergent validity and diagnostic discrimination. Assessment 9 (4), 401–405. https://doi.org/10.1177/1073191102238471.

Wang, Y.-.P., Gorenstein, C., 2013. Psychometric properties of the Beck Depression Inventory-II: a comprehensive review. Revista Brasileira de Psiquiatria 35 (4), 416–431. https://doi.org/10.1590/1516-4446-2012-1048.

Wolfner Ahava, G., Iannone, C., Grebstein, L., Schirling, J., 1998. Is the Beck Depression Inventory reliable over time? An evaluation of multiple test-retest reliability in a nonclinical college student sample. J Pers Assess 70 (2), 222–231. https://doi.org/10.1207/s15327752jpa7002_3.

Zung, W.W., 1965. A self-rating depression scale. Arch. Gen. Psychiatry 12 (1), 63–70. https://https://doi.org/10.1001/archpsyc.1965.01720310065008.