A study on the measurement of depressive symptoms: frequency versus intensity

Jessica Sundell

Handledare: Elisabet Borg
PSYKOLOGI III – Vetenskaplig undersökning, 15 poäng, VT 2017

STOCKHOLMS UNIVERSEITET
PSYKOLOGISKA INSTITUTIONEN
Depression is a common mental health disorder and is a major contributor to the overall global burden of disease. Improving the instruments that are used to detect and assist in the diagnosis of depression is of importance due to the impact that the disorder has on the individual and society. The study aimed to explore and improve the measurement of depression. Seventy-one participants’ results on the PHQ-9 (that measures frequency of depressive symptoms) and the Borg CR scale® (measuring intensity of depressive symptoms) were analysed. It was also of interest to measure mental health patients’ \( n = 16 \) self-rated depression and compare it to mental health staffs’ \( n = 7 \) observer-rated depression. The overall correlation between intensity and frequency of depressive symptoms was found to be positively strong \( (r = .695) \), the Borg CR scale showed high internal consistency when used for measuring intensity of depressive symptoms \( (.846) \), self-reported depression was lower than observer-rated depression \( (p = .167) \). In addition, inter-individual differences in frequency and intensity of depressive symptoms were also examined. Overall, the outcome of this study showed that intensity of depressive symptoms can differ considerably between individuals, despite similar overall depression scores.
Depression has been found to be a strong risk factor for suicide attempt and suicide (see for example, Crona, Öjehagen, Brådvik, Stenmarker & Hallberg, 2017; Zhang et al., 2005; Zhao et al., 2015). It can cause the affected person to suffer greatly and function poorly at work, at school and in the family. At its worst, depression can lead to suicide. Therefore, detecting and treating early depressive symptoms is essential in order to prevent the development of depression and prevention of suicide (Sousa, Zauszniewski, Mendes, & Zanetti, 2005).

In the Western world, depression is approximately twice as prevalent among women as it is among men (Grant et al., 2004; Grant et al., 2009; Kessler, 2003; Van de Velde et al., 2010). For example, one longitudinal study on depression and gender differences, found significant differences between genders on emotional regulation, which was found to be a risk factor for the occurrence of emotional disorders (Gomez-Baya et al., 2017). Specifically, lower perceived emotional repair in females partly explains their reports of higher depressive symptoms. One study found that the higher prevalence of depression among women than men is due to higher risk of first onset (Kessler, 2003). However, studies have found that suicide is higher among men than women, and that suicide plays a more probable resort for males than females when suffering from depression (Crona et al., 2017; Zhang et al., 2005; Zhao et al., 2015). One major cross-sectional study found that incidence rates of depression were greater among women, whereas incidence rates of most substance use disorders were greater among men (Grant et al., 2009).

Research has found that depression is often associated with substance use disorders, but the directional relationship is not often clear (whether depression precedes substance use or the other way around) (Grant et al., 2004; Grant et al., 2009; Moore et al., 2016; Lai et al., 2015). One study found that among individuals with substance use disorders who sought treatment in the past 12 months, approximately 40% also had a mood disorder (Grant et al., 2004). Furthermore, studies have concluded that depressive symptoms may present themselves differently depending on the individual’s substance use (Moore et al., 2016; Lai et al., 2015).

Improving the measurements that are used to detect and assist in the diagnosis of depression is of significance due to the impact that the disorder has on the individual and society. Globally, individuals who are depressed are often inaccurately diagnosed, and others who do not have the disorder are often misdiagnosed and prescribed antidepressants (WHO, 2016). One of the most frequently used ways to detect depression is through continuous rating scales, which are primarily self-reported, and are used for detection and to determine severity of depressive symptoms (Gonzalez & Esbitt, 2010; Tonge & Rowe, 2003). One of the problems with these types of scales is the assumption that all items are equally weighted. For example, a question of reduced appetite is a common question in depression scales, and are weighted as important as a question of suicidal thoughts and/or tendencies. Therefore, these types of measurements are not well adapted to differentiate in level of severity, and are often more adjusted to differentiate between depressed and non-depressed based on the scoring.

Some research has estimated that general practitioners incorrectly identify about 50% of depressed patients, and that only a limited proportion of cases receive adequate treatment (Kroenke, 2010; Mitchell, Vaze & Rao, 2009). Moreover, the conceptualisation of depression in scales is also an issue. To meet the criteria of
Unipolar depression or Major Depressive Disorder, the symptoms need to be present almost every day over a two week period (American Psychiatric Association, 2013). The criterion does not refer to the intensity of depressive symptoms.

Many of the measurements used to analyse behaviour ask the respondent to rate the frequency (e.g. almost every day, less than once a week, often etc.) (Blair & Burton, 1987). When asking for the frequency of a specific behaviour, emotion or thought (e.g. how often), one can often overlook intensity. For example, a person might have intrusive thoughts about how worthless he or she is up to several times a day, but might be able to disregard those thoughts and not invite them to become intense. Furthermore, studies that aim to improve scales of depression tend to use between-group differences as a source of measurement invariance (e.g. Whisman, Judd, Whiteford, & Gelhorn, 2013). Using solely between-subject differences might lead to overlooking specific within-subject differences, such as the intensity and concentration of depressive symptoms when evaluating the level of an individual’s depression. For example, one study found that that the specific depressive trait hopelessness was found to be a significantly higher predictor of suicide, in comparison to other depressive traits (Tsujii et al., 2014). This problem makes it especially important to research within-subject differences in depressive symptoms, instead of collecting an individual’s scores on a scale in to a total score (see e.g., The Patient Health Questionnaire, Beck Depression Inventory, Montgomery–Åsberg Depression Rating Scale, The Hamilton Depression Rating Scale).

Scales are not a substitute, but a complement to the clinical assessment and diagnosis of depression (Carneiro, Fernandes, & Moreno, 2015). Reliable and valid instruments are essential to validate the diagnosis of depression, assess progress during treatment, reduce bias caused by physical symptoms, and to measure outcomes (Bukumiric et al., 2016). Correlations between different depression scales are reported both in psychometric and clinical studies. The correlation coefficients are indicators of concurrent validity of the scales. Sufficiently high correlation coefficients suggest that the two instruments measure consistent aspects of depression (see meta analysis by Bukumiric et al., 2016; Zimmerman, Posternak, and Chelminski, 2004). For example, research have found correlations to be between \( r = .78 \) (Sayer et al., 1993) and and \( r = .85 \) (Brown, Schulberg, & Madonia, 1995) between the same depression scales (Beck Depression Inventory and the Hamilton Rating Scale for Depression).

Previous research has investigated the level of discrepancy between self-reported and observer-reported depression severity (e.g. Calamia & Bernstein, 2017; Stein et al., 2016; Tsujii et al., 2014; Wongpakaran, Wongpakaran, & van Reekum, 2013). Moreover, one study examined the level of agreement between patient’s, relatives and general practitioner’s perspective of unmet needs in elderly depressed patients (Stein et al., 2016). Their results indicated that as severity of depression increased, so did level of self-reported unmet needs in the patients. Tsujii and colleagues (2014) examined the level of discrepancy in self-reported and observer-rated depression in a sample with Major Depression. They found that patients with higher discrepancy had significantly higher traits of hopelessness and were more likely to have survived past suicide attempts. With regard to depressive symptom severity, significant differences have been found between self-reported symptoms given by a patient and those identified by an observer (e.g. a nurse, carer, psychologist or general practitioner). Such research has typically reported moderate agreement (i.e. \( r = .3 \) to \( .5 \)) in observer-rated versus self-
rated depression severity (e.g. Benedict, Fishman, McClellan, Bakshi, & Weinstock-Guttman, 2003; Calamia & Bernstein, 2017; McKinlay et al., 2008).

An impressive sample of 630 patients living with AIDS participated in the cross-sectional DHIVA (Dietitians in HIV/AIDS) study, to examine the degree of discrepancy between patient and physician’s perception of depression severity in the patients (Marando et al., 2016). The study found severe depression in 22.5% of the patient’s self-reports versus 4% identified by physicians. Furthermore, 135/155 (87%) of the self-reported severely depressed patients were considered as non or mildly/moderately depressed by their physicians. The researchers found that as self-reported depression increased, so did associations with treatment failure and unemployment, but also discrepancy in self-report versus physician’s report.

The PHQ-9 (Patient Health Questionnaire) is a widely used self-report instrument to detect and measure frequency of depressive symptoms (Kroenke, Spitzer & Williams, 2001). The scale uses the frequency of symptoms that factors into a total score of depression severity (None, Mild, Moderate, Moderately Severe and Severe). Scores range from 0 to 27, and a cut-score of ≥10 has been recommended for detecting cases of current Major Depressive Episode (Kroenke et al., 2009). It performs well in different cultures and has been translated into several languages in addition to a computerised version (Doris, Hans-Christoph, Christian, & David, 2016; Manea, Gilbody & McMillan, 2015). PHQ-9≥10 has a sensitivity of 88% and a specificity of 88% (Kroenke, Spitzer & Williams, 2001), while other studies have found a sensitivity of 81.3% and specificity 85.3% (Mitchell, Yadegarfar, Gill, & Stubbs, 2016). Another study found that PHQ-9 is sensitive (77%) but not specific (46%) at capturing depressive symptoms when compared to clinician diagnoses (Sawaya, Atoui, Hamadeh, Zeinoun, & Nahas, 2016). Many studies have confirmed good psychometric qualities of the PHQ-9 in primary care settings but also in the general population (Lowe, Kroenke, Herzog, & Grafe, 2004; Manea et al., 2015). However, psychometric studies have also warned clinician’s to not use the PHQ-9 as a way to confirm a clinical diagnosis of depression due to its brevity (Mitchell et al., 2016). Some research has suggested that a comprehensive assessment of depression should include both self-rating scales and observer-reported measures (Uher et al., 2012; Rush et al., 2006).

The Borg CR scale® (CR100, centiMax®) (hereafter called centiMax) was developed to measure perceptions, experiences and emotions (Borg & Borg, 2001; Borg, 2007). It is primarily used in medicine, ergonomics and sport research. The centiMax (figure 1) is a scale from “Nothing at all” (0) to “Absolute maximum” (●) (a rare perception at a value somewhat above 100). On the centiMax scale the term “centiMax” denotes the measurement of the perceptual entity in “centigrades” (cM) of a maximal intensity (e.g. of a maximal perceived exertion). The visual scale is constructed in a manner to avoid floor- and ceiling effects. The scale also possesses verbal anchors next to the numerical values and a deliberate visual design to assist the respondent to make authentic scoring of their perceptions. Several numbers has verbal anchors to them, which the reader can refer to when using the scale (see figure 1). The verbal anchors are placed in congruence with numbers on a ratio scale. There are several advantages to using ratio scaling, partially because it gives you equal distances between scores and has an absolute zero. Furthermore, when using scores from a ratio scale, one can imply that an experience (e.g. a depressive symptom) is twice as strong as another (not just “more than” or “less than” which is common with ordinal data).
The first aim of this study was to explore and improve the measurement of depression. For this reason, a non-clinical sample was asked to complete a well-established depression instrument that measures frequency of depressive symptoms (PHQ-9). Additionally, they were asked to respond to the same items but with the help of a category-ratio scale that measures perceived intensity (centiMax). The second aim of this study was to investigate the level of discrepancy between patient’s self-rated depression, and mental health staff’s observer-rated depression. By using the self-rated depression scale PHQ-9 and perceived intensity scale centiMax, the patients self-rated their perceived depression, and their mental health staff used the same scales but as an observer-rating one. More specifically, the aim is to examine the association between perceived intensity of depressive symptoms (using the centiMax ratio data) and frequency of depressive symptoms (using the PHQ-9 ordinal data).

Method

Setting and participants
An in-patient treatment facility south of Stockholm was selected as recruitment setting for the study. The facility is a support housing for male patients with a minimum age of 20 years. The facility specialises on men with multiple diagnosis such as ADHD/ADD, Asperger’s Syndrome, Bipolar disorder, depression, anxiety, substance abuse, personality disorders, and Schizophrenia. Meanwhile the patient’s reside at the support house, they receive daily support and care by psychiatric nurses and mental health staff. The facility also provides support including contacts with medical and psychiatric care, handling of medications, and maintenance of hygiene and economy. There is a zero tolerance of bringing alcohol or drugs into the facility, however the housing accommodates patients that are under the influence of alcohol and/or drugs that choose to continue their substance abuse outside of the facility. The researcher was granted permission to give out the survey by the support housing management, under the condition that the patient’s remain anonymous in the study. Additionally, the author of this study is a known member of the mental health staff.
The sample consisted of 40 male patients out of whom 16 participated in the study. The mental health staff (consisting of 7 full-time employees out of whom 7 participated), which helps patients with medications, social support, contact with family and health care services to name a few, was asked to participate. A convenience sample of 71 individuals also partook in the study.

Out of the 40 patients that were currently residing at the facility, 16 agreed to participate (15 denied to participate, 3 were not present at the time of the data collection, 6 were deemed to have reduced mental and/or physical capability to answer the survey). The patients were all males with a mean age of 51.31 years ($SD = 11.66$), ranging from 32 to 72 years of age. The 16 participants provided full written consent to participate. Due to ethical reasons, the participant’s diagnoses are not presented in the current study, and the participants will only be described as a sample of patients that is being treated for mental health disorders. The data gathered on the patients were collected and stored in a way so that their responses cannot be linked to them as individuals.

The inclusion criterion for the mental health staff was a full-time employment that had lasted for at least six months. The staff consisted of 4 females (57%) and 3 males (43%). Their age ranged from 39 to 57, with a mean age of 50 years ($SD = 7.05$). The mental health staff was asked to rate how depressed they believe that the patients were.

A non-clinical convenience sample ($n = 71$) also participated in the study. There were 42 males (59.2%) and 29 females (40.8%) with a mean age of 31.62 years of age ($SD = 10.92$ years) ranging from 18-58 years of age. The participants were informed that the survey is anonymous and voluntary, that the data gathered was for research purposes only, and that they could choose to terminate the survey at any point.

**Materials**

The Patient Health Questionnaire (PHQ-9) is a widely used instrument and consists of nine items matching the 4th edition Diagnostic and Statistical Manual of Mental Disorders criteria for a Major Depressive Episode (MDE). There has been little change in the diagnostic criteria from edition 4 to 5, and the PHQ-9 scoring is not affected by the changes (Volker et al. 2016). Respondents are asked to rate the items on a scale of 0 (“not at all”), 1 (“several days”), 2 (“more than half of the days”), to 3 (“nearly every day”). The questions are asked in relation to how much/often the depressive symptoms have bothered them over the past 2 weeks. PHQ-9 scores of 5 indicate mild, 10 indicate moderate, 15 indicate moderately severe and a score of $>20$ indicates severe depression. For this specific study, after the first page that contained information on confidentiality and research purpose, the participants (patients, mental health staff and non-clinical sample) rated their depressive symptoms using the PHQ-9 scale.

The participants were also asked to fill out The Borg CR scale® (centiMax) with instructions referring to perceived exertion. They were asked to use the instructions meant for rating perceived exertion as a point of reference, when rating depressive symptoms derived from the PHQ-9. The items used together with the centiMax scale were directly taken from the PHQ-9, although framed in a different way. Instead of asking how frequently the symptoms have occurred during the past 2 weeks, the items used together with the centiMax scale asked how intense these feelings had been (in centigrades/cM). The sixth item on the PHQ-9 (Feeling bad about yourself - or that you are a failure or have let yourself or your family down) was made into two separate items
on the centiMax scale (item 6 and item 6.1), due to it asking for two different depressive emotions (low self-esteem and feelings of failure). The questionnaire that was given to the patients had half of a page describing how to use the centiMax scale together with centiMax figure (as seen in figure 1 above). The depressive symptoms were located next to the figure for ease of scoring. Furthermore, the brief instructions also informed the patients that a full instruction to the scale could be read on the following page (making the total patient questionnaire four pages long).

The non-clinical sample was asked to fill out an online survey. The first page of the survey followed general guidelines about online research information gathering (i.e. research purpose, contact information and statements on confidentiality and anonymity). The second page consisted of the PHQ-9 questions, which were set up as a rating scale. The third page was information about how the participants should score their symptoms on the centiMax scale. The last page to fill out was the centiMax scale, which participants filled out with numbers in the boxes next to the symptoms.

Procedure
The researcher approached the patients during two consecutive days. They were informed that filling out the survey was voluntary, confidential and part of a research assignment. They were also informed that their responses would not be shared with anyone else among the staff apart from the researcher. Those who agreed to participate first had to read and sign a paper of informed consent, and after that they were given instructions together with the questionnaires. A total of 16 patients filled out the PHQ-9, and six completed the centiMax scale (ten were excluded due to not completing the centiMax scale).

Due to the low response rate on the centiMax scale, the mental health staff was asked to only fill out the PHQ-9. The mental health staff was asked to rate (using the PHQ-9) how depressed they think that the patients were. More specifically, the staff was informed that the patients had filled out the PHQ-9, and were now asked to fill out the PHQ-9 with a specific patient in mind. They were asked to rate their personal opinion of how frequently the patients suffer from depressive symptoms (not how they believed that the patients had rated themselves). Each member of the staff filled out 4-5 questionnaires each, so that every patient was rated by two separate staff members. Moreover, the staff was not informed about who of the patients that had filled out the PHQ-9 (apart from the patients whom they rated). The staff was also asked not to consult with their colleagues when rating the patients.

The non-clinical sample was asked to fill out an online survey through convenient sampling. The participants were recruited via social media and were friends and/or acquaintances to the author. They were given information about the survey (research purpose and anonymity of the test) via a private message from the researcher, in addition to a website link that would direct them to the online survey.

Statistical analysis and exclusion criteria
The statistical analysis of the results was completed using SPSS version 24. Cronbach alpha coefficients were used to evaluate internal consistency, and the relationships between scale items were determined with Pearson’s correlations. Intra-class correlations were run to examine inter-rater reliability among the mental health staff. Friedman’s test and Mann-Whitney U test was also run on the data to measure
differences between groups. Participants that did not complete an item in one of the questionnaires had its data for that questionnaire removed (e.g. if only one question was skipped on the centiMax scale, their responses on the PHQ-9 was still kept).

Results

Frequency versus intensity
To evaluate how well the centiMax scale correlates with the PHQ-9 scale, several analyses were performed. The non-clinical sample had a mean total score of 7.3 (SD = 4.9) on the PHQ-9 scale. The average rating on the scale items was .81 (SD = .54). The sample’s score indicated that 29.5% rated themselves as having minimal depression, 46.6% rated themselves as having mild depression, 15.4% moderate depression, 5.6% moderately severe, and 2.8% severe depression. The PHQ-9 scale had a high level of internal consistency, as determined by a Cronbach’s alpha of 0.811. The individual item medians, means and standard deviations can be seen in table 1.

Table 1. Medians, means and standard deviations for frequency of depressive symptoms on the PHQ9 scale (n = 71).

<table>
<thead>
<tr>
<th>Mdn</th>
<th>M</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Low interest</td>
<td>0</td>
<td>.44</td>
</tr>
<tr>
<td>2. Feeling depressed</td>
<td>1</td>
<td>1.3</td>
</tr>
<tr>
<td>3. Sleep disturbances</td>
<td>2</td>
<td>1.42</td>
</tr>
<tr>
<td>4. Loss of energy</td>
<td>1</td>
<td>.97</td>
</tr>
<tr>
<td>5. Eating disturbances</td>
<td>0</td>
<td>.63</td>
</tr>
<tr>
<td>6. Low self-esteem/ feelings of failure</td>
<td>1</td>
<td>.9</td>
</tr>
<tr>
<td>7. Difficulty concentrating</td>
<td>1</td>
<td>.79</td>
</tr>
<tr>
<td>8. Psychomotor disturbances</td>
<td>0</td>
<td>.44</td>
</tr>
<tr>
<td>9. Suicidal ideations</td>
<td>0</td>
<td>.39</td>
</tr>
</tbody>
</table>

The participants had a mean item rating of 18.1 cM (SD = 15.4) on the centiMax scale. Moreover, the centiMax scale showed high internal consistency, with a Cronbach’s alpha of 0.846. The medians, means and standard deviations for each item on the centiMax scale can be seen in table 2.

Table 2. Medians, means and standard deviations for intensity of depressive symptoms on the centiMax scale (n = 71).

<table>
<thead>
<tr>
<th>Mdn</th>
<th>M</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Low interest</td>
<td>2</td>
<td>12.8</td>
</tr>
<tr>
<td>2. Feeling depressed</td>
<td>2</td>
<td>13.8</td>
</tr>
<tr>
<td>3. Sleep disturbances</td>
<td>1</td>
<td>12.6</td>
</tr>
<tr>
<td>4. Loss of energy</td>
<td>15</td>
<td>24.3</td>
</tr>
<tr>
<td>5. Eating disturbances</td>
<td>10</td>
<td>20</td>
</tr>
<tr>
<td>6.1 Feelings of failure</td>
<td>15</td>
<td>24.8</td>
</tr>
<tr>
<td>7. Difficulty concentrating</td>
<td>12</td>
<td>20.7</td>
</tr>
<tr>
<td>8. Psychomotor disturbances</td>
<td>2</td>
<td>14.4</td>
</tr>
<tr>
<td>9. Suicidal ideations</td>
<td>0</td>
<td>11.5</td>
</tr>
</tbody>
</table>

To determine the relationship between the centiMax and PHQ-9 items, a Pearson product-moment correlation was performed. For this analyses, those patients who completed the centiMax scale were included making the total sample n = 77. The assumption of linearity was met, but one outlier was found and kept in the analyses. The variables were not normally distributed, as assessed by Shapiro-Wilk's test (p < .05). It
was assumed that the outlier and non-normality of the distributions was due to restriction of range rather than actual outliers, and the Pearson’s correlational analyses were run nonetheless. There was a positively strong correlation between the participant’s mean total scores on the PHQ-9 scale and their mean item ratings on the centiMax scale, \( r = .695 \) (\( p < .001 \), \( n = 75 \)). A scatterplot summarizes the participants’ total scores on the PHQ-9 and mean item rating on the centiMax (see figure 2).

Figure 2. Scatterplot depicting mean item ratings on the centiMax and total scores on the PHQ-9 (\( n = 77 \)).

A Pearson’s correlational analysis was also performed between each of the items on the PHQ-9 scale and the corresponding items on the centiMax scale. The correlational analysis was performed to measure the agreement between the scales. As seen in table 3, there were non-existing to strong positive correlations between the level of intensity of depressive symptoms (measured by the centiMax scale) and frequency of depressive symptoms (measured by the PHQ-9 scale).

<table>
<thead>
<tr>
<th>centiMax</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>6.1</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>PHQ-9</td>
<td>.651*</td>
<td>.062</td>
<td>.000</td>
<td>.511*</td>
<td>.724*</td>
<td>.565*</td>
<td>.519*</td>
<td>.457*</td>
<td>.481*</td>
<td>.669*</td>
</tr>
</tbody>
</table>

*) Correlations were significant with \( p < .001 \)

A Mann-Whitney U test was run to determine if there were differences in centiMax and PHQ-9 score between males and females. The test was chosen because the assumption of normality was not met. Distributions of the scores for males and females were similar, as assessed by visual inspection. Median PHQ-9 score was statistically significantly higher in males (\( Mdn = 6.5 \)) than in females (\( Mdn = 4 \)), \( U = 452, z = -2.58, p = .010 \), but not on the centiMax scale where males (\( Mdn = 15.8 \)) scored slightly higher than females (\( Mdn = 14.5 \)) \( U = 630, z = -.694, p = .488 \).

*Self-rated versus observer-rated*
The six patients who completed the centiMax scale rating, had a mean item rating of 38.8 cM ($SD = 24.9$). The patients had a mean sum score of 8.9 ($SD = 8.6$) on the PHQ-9 scale. The medians, means and standard deviations for each item on the PHQ-9 scale can be seen in table 4.

Table 4. Patient’s medians, means and standard deviations for items on the PHQ-9 scale ($n = 16$).

<table>
<thead>
<tr>
<th>Item</th>
<th>Mdn</th>
<th>M</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Low interest</td>
<td>1.5</td>
<td>1.1</td>
<td>1.2</td>
</tr>
<tr>
<td>2. Feeling depressed</td>
<td>0</td>
<td>.9</td>
<td>1.1</td>
</tr>
<tr>
<td>3. Sleep disturbances</td>
<td>.5</td>
<td>1.1</td>
<td>1.3</td>
</tr>
<tr>
<td>4. Loss of energy</td>
<td>1</td>
<td>1.1</td>
<td>1.3</td>
</tr>
<tr>
<td>5. Eating disturbances</td>
<td>1</td>
<td>1.4</td>
<td>1.3</td>
</tr>
<tr>
<td>6. Low self-esteem/ feelings of failure</td>
<td>0</td>
<td>.6</td>
<td>1.1</td>
</tr>
<tr>
<td>7. Difficulty concentrating</td>
<td>0</td>
<td>.8</td>
<td>1.2</td>
</tr>
<tr>
<td>8. Psychomotor disturbances</td>
<td>0</td>
<td>.8</td>
<td>1.2</td>
</tr>
<tr>
<td>9. Suicidal ideations</td>
<td>0</td>
<td>.8</td>
<td>1.3</td>
</tr>
</tbody>
</table>

The PHQ-9 found minimal depression in five of the patients’ self-rating, mild depression in five, moderate depression in three, moderately severe depression in one and severe depression in two of the patients. Six of the patients were considered as mildly depressed by mental health staff, six were rated as moderately depressed, one was rated moderately severe, and two of the patients were rated as severely depressed. The mental health staff’s mean observer-ratings on depressive symptoms in the patient’s can be seen in table 5.

Table 5. Mental health staff’s means and standard deviations of their total score rating on the PHQ-9 scale ($n = 7$).

<table>
<thead>
<tr>
<th>Staff member</th>
<th>$M$</th>
<th>$SD$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (rated five patients)</td>
<td>15.5</td>
<td>7.2</td>
</tr>
<tr>
<td>2 (rated five patients)</td>
<td>16.5</td>
<td>4.4</td>
</tr>
<tr>
<td>3 (rated five patients)</td>
<td>11.8</td>
<td>4.3</td>
</tr>
<tr>
<td>4 (rated four patients)</td>
<td>4.3</td>
<td>1.9</td>
</tr>
<tr>
<td>5 (rated five patients)</td>
<td>8.8</td>
<td>2.6</td>
</tr>
<tr>
<td>6 (rated four patients)</td>
<td>9.8</td>
<td>6.4</td>
</tr>
<tr>
<td>7 (rated five patients)</td>
<td>7.6</td>
<td>2.5</td>
</tr>
</tbody>
</table>

To estimate the level of inter-rater reliability among the mental health staff, intra-class correlations were calculated. ICC estimates were calculated based on a mean rating, consistency, and one-way random-effects model. The inter-rater reliability among the mental health staff was $r = .77$. In figure 3, the reader can visually interpret the staff-ratings.
To measure any differences between patient’s self-ratings and staff-ratings, a between-subject statistical analysis was required. Since the assumptions of a repeated measures ANOVA were violated, a Friedman’s test was run to determine if there were differences in the patient’s self-rated depression and staff-rating (by either their first or second rater). Depression was rated lower by the patient’s ($Mdn = 7$), than their mental health staff ($Mdn = 10.5$ to $11.5$), but the differences were not statistically significant, $\chi^2(2) = 3.57, p = .167$.

One of the advantages with ratio data is the possibility of more detailed comparisons of symptoms. This may, for example, be done in symptoms profiles. Two individuals (participant ID 24 and participant ID 49) were arbitrarily chosen because they were found to have similar mean item rating on the centiMax ($m = 15.6$ cM and $m = 17.7$ cM respectively). They were also found to have equal PHQ-9 total score (participant ID 24 and 49 both scored a total of 8, suggesting “mild depression”). On item 6 on the centiMax scale (“low self-esteem”), the participants responded differently (participant ID 49 responded with $112$ cM, and participant ID 24 responded with $0$ cM). On the corresponding item on the PHQ-9 scale, participant ID 24 responded with 1 (“several days”), and participant ID 49 with a 2 (“more than half of the days”). In figure 4, the reader can visually interpret the two participant’s differences in symptom severity.
Figure 4. Participant ID 24 and 49 differences in symptom severity on the centiMax scale, with identical total scores on the PHQ-9 (both scored as “mildly depressed”).

Discussion

The aim of this study was to explore and improve the measurement of frequency and intensity of depressive symptoms. For this purpose, a non-clinical sample answered a well-established depression scale that measures frequency of depressive symptoms (PHQ-9). Additionally, the aim was to test those items on a category-ratio scale that measures perceived intensity of depressive symptoms (centiMax). More specifically, the objective was to validate the centiMax scale against the PHQ-9 for measuring severity of depressive symptoms. Furthermore, the aim was also to validate those scales on a clinical sample and as observer-rating scales. More specifically, the discrepancy between mental health patient’s self-rated severity of depressive symptoms and mental health staff’s observer-rated severity of depressive symptoms was also measured. However, due to a low patient response rate on the centiMax scale, the discrepancy between patients and mental health staff was only compared using the PHQ-9.

The non-clinical participants had a mean item rating of 18.1 cM ($SD = 15.4$) on the centiMax, which insinuates that the average depressive symptom perception was between “Weak” and “Moderate”. Moreover, the sample had a mean total score of 7.3 ($SD = 4.9$) on the PHQ-9 scale, suggesting mild depression. The results from the non-clinical sample indicated high internal consistency (measured with Cronbach’s alpha) on both the centiMax (0.846) and the PHQ-9 (0.811). However, the results from the item correlations between the scales suggested that some participants had frequent depressive symptoms, but at a low intensity (and also the opposite: high intensity and low frequency) (see table 5). The lowest correlation was found between the intensity and frequency on the item referring to sleep disturbances. There are many possible reasons for that. One can imagine an individual who has increased or decreased sleep,
but the intensity and frequency of that symptom might not depend on each other, and pose a zero correlation. Another symptom of depression (feeling depressed) also showed minimal correlation between intensity and frequency, which suggests that intensity is independent of frequency in this particular sample’s depressed mood. Furthermore, the highest correlation was found between intensity and frequency on the item referring to eating disturbances. A possible implication of these results is that it may be useful to measure both aspects of depression (frequency and intensity). The overall correlation between items on the centiMax scale and items on the PHQ-9 scale was $r = .695$, insinuating a strong positive correlation between the scales. The correlation coefficients are indicators of concurrent validity of the scales, and that the two scales measure consistent aspects of depression (Bukumiric et al., 2016; Zimmerman, Posternak, and Chelminski, 2004). Moreover, in hindsight, considering previous research on the factors such as daily tiredness and personality traits when rating retrospectively (Aire, Anu, & Jüri, 2016), the PHQ-9 could be improved (by not only the use of cM scoring) by asking frequency in a more direct manner (e.g. number of actual occasions instead of ordinal rating).

A Mann-Whitney U test was run to measure differences in depressive symptoms between males and females, found that the mean total score on the PHQ-9 was significantly higher in males ($Mdn = 6.5$) than females ($Mdn = 4$). However, the difference was not statistically different on the centiMax scale where males ($Mdn = 15.8$) scored slightly higher than females ($Mdn = 14.5$). This is partly inconsistent with previous research that has found depression to be more prevalent among women (Grant et al., 2009; Kessler, 2003; Van de Velde et al., 2010). It is also interesting that the difference between genders was significant when measured with the PHQ-9, but not when measured with the centiMax. This can possibly be explained by the different scores that ratio versus ordinal data provides, and should be further examined in future studies. However, the non-clinical sample is not representative of the general population (due to the restriction of range and the convenient sampling). It would be of interest to further study these gender differences and how depressive symptoms manifest differently in males and females.

The patients had a mean total score of 8.9 ($SD = 8.6$) on the PHQ-9 scale. According to the PHQ-9 interpretation of scoring, it indicates “mild depression” and suggests “Physician uses clinical judgment about treatment, based on patient's duration of symptoms and functional impairment” (PHQscreensers, 2013). The six patients who filled out the centiMax scale had a mean item rating of 38.8 cM ($SD = 24.9$). That score insinuates that the average depressive symptom perception was fairly close to the centiMax verbal anchor “Fairly strong”. With this type of data, we can conclude that someone with a rating of 40 on item 4 (“loss of energy”) perceives twice as much “loss of energy” as an individual with a rating of 20 on the same item. On the PHQ-9 scale there are no equal intervals, e.g. a score of 2 (“more than half of the days”) is not necessarily twice as much as a score of 1 (“several days”).

To measure if there were significant differences between the patient’s self-rated depression and the mental health staff’s observer-rated depression, a Friedman’s test was run. The results found that the mental health staff rated the patient’s depression as higher than the patient’s rated themselves. The differences were not statistically significant, which can be interpreted as a non-significant discrepancy in perceived depression between the patients and mental health staff. These findings are not
consistent with previous research that has found a significant discrepancy between self- and informant-rated depression (e.g. Marando et al., 2016; Stein et al., 2016), but also consistent with other research that has found moderate agreement between the two perspectives (Calamia & Bernstein, 2016; Wongpakaran et al., 2013). There are many factors that could have impacted the findings. For example, observer-rating scales can reflect different types of expertise between the observer and patients (e.g. an observer might fall back on their experience while patients may have better access to their subjective feelings) (Demyttenaere & De Fruyt, 2003). Furthermore, the staff-ratings might have been subjected to the raters' different interpretations of the rating scale (Hoyt & Kerns, 1999). The staff-ratings could also have been subjected to their different evaluations of the same targets (Hoyt & Kerns, 1999). Consequently, observer-ratings poses threats to the validity and reliability of the results when it comes to rating internalized, complex symptoms such as depressive ones (Demyttenaere & De Fruyt, 2003; Hoyt & Kerns, 1999; Rush et al., 2006).

It was also of interest to illustrate the benefits of ratio data and attempt to make a symptom profile. Two individuals (participant ID 24 and participant ID 49) were found to have similar mean item rating on the centiMax (15.6 cM and 17.7 cM respectively). They were also found to have equal PHQ-9 total score (participant ID 24 and 49 both scored a total of 8, insinuating “mild depression”). For example, on item 6 on the centiMax scale (“low self-esteem”) the two individuals responded differently (participant ID 49 responded with 112 cM, and participant ID 24 responded with 0 cM). Participant ID 49 responded that low self-esteem was perceived as above “extremely strong”, and participant ID responded that “low self-esteem” was perceived equal to “nothing at all”. On the corresponding item on the PHQ-9 scale participant ID 24 responded with 1 (“several days”), and participant ID 49 with a 2 (“more than half of the days”). The centiMax symptom profile of the individuals’ symptoms illustrates how two individuals with similar scores can have different intensity of symptoms. The PHQ-9 method of summarizing all symptom scores into a total score do not take into consideration how symptoms severity look different among individuals with similar sum scores. Future studies should examine how frequency and intensity of depressive symptoms differ between for example males and females, early versus late onset of depression, or between different antidepressants.

Furthermore, a notable distinction should be made between observer and clinician-rated scales of depression. Other studies that have examined discrepancy between self-rated and observer-rated depression has had a more clinically skilled observer (e.g. physician, psychologist or nurse) (e.g. Calamia & Bernstein, 2017; McKinlay et al., 2008). While the mental health staff does have education (usually a one year non-university degree) and experience with psychiatric patients, they may lack the clinical expertise on how depressive symptoms manifest itself. While studies have confirmed that complete assessment of depression should include both observer-rated and self-report (Uher et al., 2012), while other studies have confirmed that neither scales provide adequate assessment (Rush et al., 2006).

Previous research has discussed the many difficult aspects of recognizing and differentiating between depression, anxiety, alcohol and substance use (Grant et al., 2004; Lai et al., 2015). Some of the patient’s symptoms (such as symptoms as a result of their psychopathology and/or substance abuse) might interfere with recognizing depression by the mental health staff. A significant part of the mental health staff’s
work (at the particular facility where the research was performed) is dedicated to control and avert the patient’s from substance abuse, which can lead the staff to overlook other clinically relevant aspects such as depression. Moreover, the patients that demand more resources from the staff (in terms of more medical and/or health resources, averting from substance abuse etc.) might also lead to other, less demanding patients to not have their depressive symptoms recognized. Moreover, the psychopathology, comorbidity and substance use among the patients may have reduced their ability and/or willingness to provide authentic responses. In addition, it is not unlikely that some of the patients were under the influence of a drug and/or alcohol when answering the questionnaires, which could have impacted their responses.

The generalizability of the results should be examined. Firstly, the convenient sampling of the non-clinical sample poses a threat to the generalizability of the results. Because a majority of the non-clinical sample were direct friends or acquainted with the researcher, it poses a risk of selection bias by the author. It also poses risks by the participants to exhibit demand characteristics (despite that the survey was anonymous).

Furthermore, depression is a complex construct and two individuals with equal self-rating scores might manifest their symptoms very differently (as obvious from the example with the symptom profile), thus making it difficult to receive a reliable observer rating (Uher et al., 2012). The self-rating by the patients also possess limitations. For example, poorer patient memory and functional abilities can reduce the ability to provide valid self-reports (Calami & Bernstein, 2017).

Previous studies have found that observer-rated depression have been both higher and lower than self-rating, so it is difficult to say whether these results are congruent or not with previous research. Some research has found perceived lower severity in observer-rated depression, in comparison to self-rating (e.g. Marando et al., 2016: Stein et al., 2016). For example, Hartmann, Fritzsche, and Lincoln (2013) found that patients who rated their depressive symptoms as less severe than the clinicians demonstrated more negative symptoms such as decreased affective rapport and blunted affect. Patients who rated their depression symptoms as being more severe were found to self-report more symptoms of other psychopathology. Thus, factors contributing to discrepancies between self-rated and observer-rated depression need further examining.

The comparison of self-rated depression and observer-rated depression was limited by the small sample of patients that agreed to participate. This can be due to various reasons. The patient’s symptoms of psychopathology could possibly have impacted their willingness. Moreover, since the researcher was a known member of the staff, the low number of patient’s that participated in the study could have been due to fear of stigmatization. More patients might have participated if the researcher was unknown to them in addition to being assured that their responses would not be shared with any of the staff members. Or possibly lesser patients might have participated if the researcher was unknown to them.

Furthermore, the patient’s low response rate on the centiMax scale can also be due to various reasons. Order effects (as the centiMax was presented on the last page), the length of the questionnaire (it being 4 pages long for the patient’s to read), and the patient’s mental health status (e.g. reduced ability to concentrate for a longer amount of time in the patients with ADHD, side effects of antipsychotic medications etc.) may
have contributed. In hindsight, the mental health staff’s rating on the centiMax scale would have still been useful data, despite the low patient respondents.

The participant’s specific diagnosis and current psychopathology could have impacted the way they rated their own symptoms, in other words individual’s subjective representations of their illness can impact the way the perceive it (e.g. Baines & Wittkowski, 2013, a systematic review on the impact of illness representations in patients with psychopathology). Future studies should examine how this self-rated depression is related to patient’s illness representations.

In conclusion, this study found a positively strong overall correlation between frequency of depressive symptom (measured by the PHQ-9) and intensity of depressive symptoms (measured by the centiMax). The centiMax scale showed high internal consistency when used for measuring intensity of depressive symptoms, and was also found useful for making inter-individual comparisons of symptom severity. Moreover, gender differences in depression were found to be significantly different when measured with the PHQ-9, but not with the centiMax. A small discrepancy was found between mental health staff’s observer-rated depression and mental health patient’s self-rated depression.

References


Kroenke, K. (2010). Review: GPs accurately diagnose about 50% of patients with depression and accurately classify 81% of non-depressed patients. *Annals Of Internal Medicine, 152*, 4-13.


