Critical Review on the Geriatric Depression Scale (GDS)

Introduction (GDS purpose, population & design)

- The Geriatric Depression Scale (GDS) was the first instrument designed to measure specifically depressive symptoms in the elderly population (Yesavage et al., 1982).
- It's a self-report measure of depression in a Yes/No format for elderly people (+65 years old). Suitable both for older people living independently in the community or institutionalized in acute or long-term care settings, and useful for screening depression not only in the physically healthy elderly, but also in the physically ill (Yesavage et al., 1982).
- The primary version was a 30-item scale, but it was later reduced to a 15-item form (Sheik & Yesavage, 1986, cited by Zhao, He, Yi & Yao, 2019, p. 1), as the long version with 30 items was proved to be both time-consuming and difficult for some patients to complete, so a 15-item version of the scale was developed with the items that proved to have a higher chance to identify depressive patients.
- The 30 main items of the scale were selected from a 100-item questionnaire that was given to normal and severely depressed elderly individuals in order to identify which questions were most highly correlated with the total scores, thus presenting a better capability to identify depression in the elderly population. The 30-items selected were then readministered in its self-rating form to a new group of 47 elderly subjects, that were all over 55 years old, both male and female, and were either non-depressed subjects living in the community with no history of mental illness or patients hospitalized for depression (from hospitals in Santa Clara County California) → Normative Population (Yesavage et al., 1982)
- On the long version of the scale, 20 items indicate the presence of depression when answered positively while the other 10 are indicative of depression when answered negatively. For the short version it is 10 and 5, respectively (Yesavage et al., 1982).
- Examples of the dimensions analyzed in the GDS-15 are: general depressive affect (seven items – e.g., «*Do you feel that your situation is hopeless?*»), life satisfaction (four items – e.g., «*Do you feel happy most of the time?*»), and withdrawal (three items – e.g., «*Do you prefer to stay at home rather than going out and doing new things?*») (Mitchell et al., 1993). These dimensions are not stable across the different versions, as some studies present two, three, and four factors (Zhao, He, Yi & Yao, 2019).
- Differently from other depression scales, GDS does not identify suicidal tendencies and it's only based on psychic symptoms, not on somatic ones, as these tend to be very common complaints in this range of population (Yesavage et al., 1982).
- One crucial factor is GDS's importance in the discrimination between depression and dementia (Smarr & Keefer, 2011).

GDS-30 Validation Phase

How did the original authors tested the scale reliability and validity?
(Yesavage et al., 1982)

1. Method

a) Selection of 3 groups
   - 1st group: normal elderly subjects that were functioning well in the community and had no history of mental illness (n=40)
   - 2nd group: elderly subjects diagnosed with mild depression (n=26)
   - 3rd group: elderly subjects diagnosed with severe depression (n=34)

- The subjects from the 2nd and 3rd group were both inpatients and outpatients, while the subjects from the 1st group were all recruited at local senior centers and housing projects; the patients from the 3 groups were both male and female
- The differentiation of the clinically depressed subjects into mild and severe depression groups was based on whether or not, during a clinical interview, the patients met the Research Diagnostic Criteria (RDC) for a major affective disorder – the 2nd group (mildly depressed subjects) had an average of 3.4 RDC criteria symptoms, while the 3rd group (severely depressed subjects) had an average of 5.9 RDC criteria symptoms. The RDC were chosen as the basis for classifying the level of depression in subjects because of a consensus among researchers that it appears to capture the essential aspects of depressive disorders.

b) Clinical Interviews
- The subjects in all groups were given a 30-60 min clinical interview conducted by the authors
- The interviews involved a rating of the Hamilton Rating Scale for Depression (HRS-D) and the administration of two self-rating scales – the Zung Self-Rating Depression Scale (SDS) and the GDS
- The order in which the scales were administered was randomly determined for each subject

2. Results

a) Reliability (Internal Consistency & Test-Retest Reliability)

↳ Internal Consistency

- *Average Inter-Item Correlation* (i.e., examination of the extent to which scores on one item are related to scores on all other items in a scale, by comparing correlations between all pairs of questions that test the same construct by calculating the mean of all paired correlations) → The GDS scored 0.36, suggesting that while the items are reasonably homogenous, they do contain sufficiently unique variance so as to not be isomorphic with each other (Yesavage et al., 1982).
- *Average Item-Total Correlation* (i.e., analysis of the scale's items and if they conform to measure the same result, by taking the average inter-item correlations and calculating a total score for each item, then averaging these) → The GDS rated as 0.56, suggesting that all of the items on this scale do, in fact, measure a common latent variable (Yesavage et al., 1982).
- *Split-Half Correlation* (which divides items that measure the same construct into two tests, which are applied to the same group of people, then calculates the correlation between the two total scores) → The GDS was found to have a 0.94 split-half correlation (Yesavage et al., 1982).
- *Cronbach's alpha* (which calculates an equivalent to the average of all possible split-half correlations and indicates how well a set of variables or items were used to assess the desired aspect, thus being a coefficient utilized in order to provide an overall measure of the internal consistency of the scale) → In the GDS case, the computed value of the alpha coefficient was 0.94, suggesting a high degree of internal consistency (Yesavage et al., 1982).

↳ Test-Retest Reliability (which evaluates the scale consistency in results across time, meaning, if the test gives the same results in similar circumstances) → It was found that GDS has a 0.85 correlation ($p < 0.001$), revealing a high test-retest reliability (Yesavage et al., 1982).

b) Criterion Validity (Concurrent & Predictive Validity)

↳ The analysis of the validity of the GDS requires to focus on the criterion validity, which is the «*scale's ability to differentiate between depressed individuals and non-depressed individuals*» (Stiles & McGarrahan, 1998, p. 95). The criterion validity functions as a guide of how well a test correlates with an accepted standard or comparison (in this case, the RDC).

↳ *Assumption 1* (Concurrent Validity): if both the classification variable (classification of subjects as nondepressed, mildly depressed, or severely depressed) and the GDS are valid indices of depression, it would be expected that nondepressed subjects receive the lowest GDS scores whereas severely depressed subjects should score the highest on this scale.

- *Analysis of variance* → the classification variable served as a between-subjects factor while the subjects' total scores on GDS served as the dependent measure. Similar analysis were also performed on the SDS and HRS-D.
    - The means were ordered as predicted
    - The main effect for the classification variable was highly significant in each analysis
    - *t*-tests conducted between each pair of means showed that subjects classified as normal scored significantly lower on each of the scales compared to the mildly and severely depressed subjects while the severely depressed group scored higher than each of the other two groups (all $p < 0.001$)
- *Conclusion*: these findings provided evidence for the validity of the GDS as a measure of depression, as well as they validated the SDS and HRS-D.

↳ *Assumption 2* (Concurrent Validity): the authors also determined the relative strength with which GDS is related to the RDC - given RDC wide acceptance and the lack of a better set of criteria, the failure of a scale to correlate well with the RDC probably reflects more upon the scale in question than the RDC.

They firstly computed the correlation of each depression scales with the classification variable derived from these criteria, and then, following Ferguson (1971) (cited by Yesavage et al., 1982, p. 45), they compared the magnitude of each correlation to the other two.

- *Correlation between the classification variable and the GDS* → $r = 0.82$, $p < 0.001$
- *Correlation between the classification variable and the SDS* → $r = 0.69$, $p < 0.001$
- *Correlation between the classification variable and the and HRS-D* → $r = 0.83$, $p < 0.001$

Comparing each of this correlations to the others showed that, whereas those associated with the GDS and the HRS-D did not differ significantly from each other [$t(97) < 1$], both of these were significantly greater in magnitude than that associated with the SDS:

- *GDS vs SDS* → $t(97) = 3.83$, $p < 0.001$
- *HRS-D vs SDS* → $t(97) = 3.85$, $p < 0.011$

- *Conclusion 1*: GDS seems to discriminate effectively between the normal, mildly depressed, and severely depressed subjects.
- *Conclusion 2*: Despite the differences in content between the RDC and the GDS, the GDS total score was found to still correlate as strongly with the number of RDC symptoms as the HRS-D (which content corresponds more closely with these criteria), thus, emphasizing the subjective aspects of depression rather than the somatic and behavior aspects does not seem to have detracted from the validity of the GDS.
- *Conclusion 3*: Also, despite the differences in content between the GDS and HRS-D (i.e., absence of somatic symptoms on GDS and reliance upon them in HRS-D), GDS appears to be as valid as the HRS-D – this may be explained in part by the fact that both scales assay mood dysphoria and other psychological symptoms of depression, which seem to best discriminate between the depressed and nondepressed aged.

↳ *Sensitivity, Specificity & Cut-Off Score* (Predictive Validity)

- Computing indices of sensitivity and specificity for the measure gives us information on the percentage of individuals correctly and incorrectly classified using particular scores on this measure.
- In this case, according to Stiles and McGarrahan (1998), sensitivity deals with the correct recognition of subjects that were determined as depressed by the comparison with as external or independent criterion; while specificity deals with the correct identification of subjects that were determined as not depressed by the same external or independent criterion. Therefore, a low sensitivity results in depressed persons being missed using a criterion and classified

incorrectly as nondepressed; and a low specificity results in nondepressed persons being incorrectly labelled as suffering from depression.

- According to Zedeck (2014) the cut-off score is a value held to delimitate the lowest point at which a certain category is attained, and it's of huge importance as it influences the sensitivity and specificity. As stated by Stiles and McGarrahan (1998), "*the higher the cutoff score, typically the lower the sensitivity and higher the specificity… Similarly, the lower the cutoff score, the higher the sensitivity and lower the specificity*" (p. 95).
- Indeed, in a sample of elderly persons drawn from the same centers as those used in their study, the authors' found that a cut-off score of 11 on the GDS yielded a 84% sensitivity rate and a 95% specificity rate; while more stringent cut-off score of 14 yielded a slightly lower, 80%, sensitivity rate, but resulted in the complete absence of nondepressed persons being incorrectly classified as depressed (i.e. a 100% specificity rate)
- Based on these finding, the authors suggested that a score of 0-10 should be viewed as within the normal range, while a score of 11 or greater should be viewed as being a possible indicator of depression.

c) Construct Validity (Convergent Validity)

↳ Assumption: given previous findings indicating that the SDS (Zung, 1965; Hedlund & Vieweg, 1979, cited by Yesavage et al., 1982, p. 45) and HRS-D (Carroll et al., 1973; Hamilton, 1960, 1967; Biggs et al., 1978; Knesevich et al., 1977, cited by Yesavage et al., 1982, p. 45) are valid measures of depression, positive correlations between these measures and the GDS would provide evidence for the scales' convergent validity.

- *Correlation between the GDS and the SDS* → $r = 0.84$, $p < 0.001$
- *Correlation between the GDS and the HRS-D* → $r = 0.83$, $p < 0.001$

Conclusion: these analysis provided evidence of the convergent validity of GDS.

Recent Reviews

a) Construct Validity (Convergent Validity)

- A later comparison of the ability to differentiate nondepressed, mildly depressed, and severely depressed individuals (diagnosed according to the RDC) showed the GDS-30 to be comparable to the HRS-D (*F*-scores of 99.48 and 110.63, respectively) and superior to the SDS (*F*-score of 44.75) (Spitzer et al., 1978; Yesavage et al., 1883, cited by Balsamo et al., 2018, p. 2030).
- High correlations were also found between the GDS-30 and other scales by Snyder et al. (2000)(cited by Balsamo et al., 2018, p. 2030) in a clinical sample of older adults:
  - *Between GDS-30 and SDS* → $r = 0.88$, $p < 0.001$
  - *Between GDS-30 and CES-D* → $r = 0.82$, $p < 0.001$
  - *Between GDS-30 and HRS* → $r = 0.77$, $p < 0.001$
  - *Between GDS-30 and CPRS-D* → $r = 0.86$, $p < 0.001$
  - *Between GDS-30 and the BDI* → $r = 0.78$, $p < 0.0001$
- Regarding its correlation with anxiety and the quality of life**,** both the correlations between the GDS-30 and the Spielberger State-Trait Anxiety Inventory-Trait Scale (STAI-Trait) ($r = 0.47$, $p < 0.01$), and with the Quality of Life inventory (QOLI) ($r = 0.49$, $p < 0.01$) in a clinical sample affected by Generalized Anxiety Disorder were high (Frisch et al., 1992; Frisch, 1994; Snyder et al., 2000; Spielberger et al., 1983, cited by Balsamo et al., 2018, p. 2030).

b) Criterion Validity (Concurrent & Predictive Validity)

- *Predictive Validity* → In the meta-analysis realized by Krishnamoorthy, Rajaa and Rehman (2020), the sensitivity and specificity of GDS-30 were found to be 82% and 76%, respectively, with near higher diagnostic accuracy (AUC=0.85); while the sensitivity and specificity of GDS-15 were found to be 86% and 79%, respectively, also with higher diagnostic accuracy (AUC = 0.90). Both this results were closely similar to the antecedent ones from the review conducted by Wancata et al. (2006) and to previous reviews on diagnostic accuracy of GDS-15.
- *Predictive Validity* → Most of the studies with GDS-30 had a cut-off value of 10 or 11, while most of the studies with GDS-15 had a cut-off value of 5 or 6 (Smarr & Keefer, 2011).
- *Concurrent Validity* → The majority of studies used individual clinical interviews accompanied by the DSM as the gold standard (i.e., external or independent criterion), but even though the clinical interview provides an accurate finding in terms of diagnosis, this was not a stable choice as the gold standard in all researches (Wancata et al., 2006; Davison, McCabe & Mellor, 2009; Lozupone et al., 2016; Santos, Nunes, Kislaya, Gil & Ribeiro, 2019).

Critics

→ In the original authors' study, differences in the content and format of the three scales should be considered when making comparisons between them and the criterion → the criterion (RDC) is more heavily represented on the HRS-D, so this scale would be expected to be more strongly related to the RDC, and the group classification variable, than the other two scales (GDS and SDS) which were in disadvantage in the analyses undertaken in this study because they do not measure all of the symptoms comprising the RDC, while measuring others (e.g., diurnal symptom variation) which are not reflected in these criteria.

→ Allen-Burge et al. (1994) (cited by Balsamo et al., 2018, p. 2031) reported gender effects on the GDS, with poorer detection of depression in males.

→ Evidence supporting the use of the GDS with cognitively impaired individuals, were mixed:

- Feher et al. (1992) (cited by Balsamo et al., 2018, p. 2030) confirmed it as a valid measure of mild-to-moderate depression in Alzheimer's patients with mild-to-moderate dementia. Several authors also argue that the GDS should be used very cautiously as a screening instrument in a population in which dementia is prevalent or in persons known to have dementia, as these patients tend to disavow memory loss and to deny depressive symptoms on the GDS, which makes the use of the GDS in patients with severe dementia not recommended (Korner et al., 2006; Sheikh et al., 1986; Wancata et al., 2006, cited by Balsamo et al., 2018, p. 2030). To support this, the correlation of the GDS with the CSDD (Cornell Scale for Depression in Dementia) was found to be relatively high (r = 0.77 , p < 0.01) in patients with mild dementia diagnosed with a score of 22 or less on the MMSE (Mini-Mental State Exam), but weaker (r = 0.37 , p = 0.17) with increased cognitive impairment (Agrell et al., 1989; Ott et al., 1992, cited by Balsamo et al., 2018, p. 10). Debruyne et al. (2009) recommend the administration of GDS always together with the MMSE to evaluate the patients' mental capacities, given that they consider the GDS to don't be always reliable in assessing patients with other mental comorbidities.
- On the other hand, the GDS has been found to have moderate sensitivity (82.6%) and specificity (81.3%) in an inpatient, mostly cognitively impaired, geriatric sample (Bentz et al., 2008, cited by Balsamo et al., 2018, p. 2031). Also, in another study, it was found to differentiate depressed from nondepressed elderly undergoing cognitive treatment for senile dementia – these subjects were classified as demented by criteria of Folstein et al. (1975) (cited by Yesavage et al., 1982, p. 46) MMSE and it was found that the subjects categorized as depressed by a therapist blind to GDS scores received a mean score of 14.72 (S.D. = 6.13) on the GDS *vs* a mean of only 7.49 (S.D. = 4.26)

for nondepressed subjects, t (41) = 4.4, $p < 0.001$ – nevertheless, these results should only be viewed as suggestive since the number of subjects was small (n=43).

→ Despite of the GDS being suitable for the detection of depression among the elderly persons living independently in the community or institutionalized in acute or long-term care settings, some authors argue that the psychometric properties of the GDS when it is used with institutionalized elderly in nursing homes are not as satisfactory as with the community elderly. According to Stiles and McGarrahan (1998), "*elders living in the community are better able to assess themselves and respond to the GDS questions more accurately than their nursing home counterparts*" (p. 100). Additionally, differences in the sensitivity and specificity of the scale were found to depend on the type of settings:

- In both the GDS-15 and GDS-30, the mean specificity was significantly higher among in-patients than among out-patients and nursing home residents, and the same results were verified for the GDS-30 mean sensitivity, while it was significantly higher in nursing homes and among in-patients than among the out-patients in GDS-15 (Wancata el al., 2006).

→ By the time this scale was first developed, the authors recognized that more research was needed on the expression of depression within elderly subjects (e.g., somatic symptoms), and Debruyne et al. (2009) later argued that some comorbidities that are very common in this population were not taken into consideration on this primary study.

References

Balsamo, M., Cataldi, F., Carlucci, L., Padulo, C., & Fairfield, B. (2018). Assessment of late-life depression via self-report measures: a review. *Clinical interventions in aging*, *13*, 2021.

Davison, T. E., McCabe, M. P., & Mellor, D. (2009). An examination of the "gold standard" diagnosis of major depression in aged-care settings. *The American journal of geriatric psychiatry*, *17*(5), 359-367.

Debruyne, H., Van Buggenhout, M., Le Bastard, N., Aries, M., Audenaert, K., De Deyn, P. P., & Engelborghs, S. (2009). Is the geriatric depression scale a reliable screening tool for depressive symptoms in elderly patients with cognitive impairment? *International Journal of Geriatric Psychiatry*, *24*(6), 556-562

Geriatric depression scale (GDS). (n.d.). https://www.apa.org

Krishnamoorthy, Y., Rajaa, S., & Rehman, T. (2020). Diagnostic accuracy of various forms of geriatric depression scale for screening of depression among older adults: Systematic review and meta-analysis. *Archives of Gerontology and Geriatrics*, *87*, 104002.

Lozupone, M., Veneziani, F., Galizia, I., Lofano, L., Montalbò, D., Arcuti, S., Tortelli, R., Barulli, M. R., Capozzo, R., Bonfiglio, C., Panza, F., Seripa, D., Todarello, O., & Logroscino, G. (2016). Validity of the Geriatric Depression Scale-30 against the gold standard diagnosis of depression in older age: The GreatAGE Study. *European Psychiatry*, *33*(S1), S85-S85.

Mitchell, J., Mathews, H. F., & Yesavage, J. A. (1993). A multidimensional examination of depression among the elderly. *Research on Aging*, *15*(2), 198-219.

Santos, A. J., Nunes, B., Kislaya, I., Gil, A. P., & Ribeiro, O. (2019). Estudo de validação em Portugal de uma versão reduzida da Escala de Depressão Geriátrica. *Análise Psicológica*, *37*(3), 405-415.

Smarr, K. L., & Keefer, A. L. (2011). Measures of depression and depressive symptoms: Beck Depression Inventory-II (BDI-II), Center for Epidemiologic Studies Depression Scale (CES-D), Geriatric Depression Scale (GDS), Hospital Anxiety and Depression Scale (HADS), and Patient Health Questionnaire-9 (PHQ-9). *Arthritis care & research*, *63*(S11), S454-S466.

Stiles, P. G., & McGarrahan, J. F. (1998). The Geriatric Depression Scale: A comprehensive review. *Journal of Clinical Geropsychology*.

Wancata, J., Alexandrowicz, R., Marquart, B., Weiss, M., & Friedrich, F. (2006). The criterion validity of the Geriatric Depression Scale: a systematic review. *Acta Psychiatrica Scandinavica*, *114*(6), 398-410.

Yesavage, J. A., Brink, T. L., Rose, T. L., Lum, O., Huang, V., Adey, M., & Leirer, V. O. (1982). Development and validation of a geriatric depression screening scale: a preliminary report. *Journal of psychiatric research*, *17*(1), 37-49. https://doi.org/10.1016/0022-3956(82)90033-4

Zedeck, S. (Ed.). (2014). *APA dictionary of statistics and research methods*. Washington, DC: American Psychological Association.

Zhao, H., He, J., Yi, J., & Yao, S. (2019). Factor structure and measurement invariance across gender groups of the 15-item Geriatric Depression Scale among Chinese elders. *Frontiers in Psychology*, *10*, 1360.