A Comparison of Depressive Symptoms in Stroke and Primary Care: Applying Rasch Models to Evaluate the Center for Epidemiologic Studies-Depression Scale

A. Simon Pickard, PhD,[1] Mehul R. Dalal, PhD,[2] Donald M. Bushnell, MA[3]

[1]College of Pharmacy, University of Illinois at Chicago, Chicago, IL, USA; [2]Global Health Outcomes, GlaxoSmithKline, Philadelphia, PA, USA; [3]Health Research Associates, Inc., Seattle, WA, USA

ABSTRACT

Objectives: Clinical trials and community-based studies often include the Center for Epidemiologic Studies-Depression scale (CES-D) as a measure of depression outcome. We compared responses to symptom-related items on the CES-D by depressed stroke and primary-care patients for several purposes: 1) to illustrate the use of Item Response Theory (IRT)-based (Rasch) models for comparing scale functioning across different patient subgroups; and 2) to inform clinicians and outcome researchers about scale functioning and depressive symptomatology in stroke- compared with primary care-based depression.

Methods: Two data sources were analyzed, including 32 depressed patients who were 3 months poststroke, and 366 depressed primary-care patients. Presence of depression was based on a CES-D score 16 or higher. Rasch models were used to assess item fit and compare item hierarchies between depressed primary-care and stroke patients.

Results: Item hierarchies were similar for poststroke depression and primary care-based depression. Interpersonal disruption items were the most difficult to endorse for both groups. No items misfit the scale in primary-care depression. Items relating to restless sleep, unfriendliness, and crying slightly misfit the scale in stroke patients, that is, may measure a different trait. Differential item functioning (DIF) between the groups was identified for items relating to appetite, restless sleep, crying, and feeling disliked.

Conclusions: Results generally supported the use of the CES-D as measure of depression outcome, particularly in primary care-based depression. DIF may imply that slightly different clusters of depressive symptoms are reported by depressed stroke patients compared with primary care, but this is conjectural given the small stroke sample size and the same items have been previously associated with bias in studies of large nonstroke samples. This study found Rasch models to be useful tools to investigate scale performance for different clinical applications.

Keywords: cerebrovascular disease, CES-D, depression, psychometrics, stroke.

Introduction

Depression is a common occurrence after stroke, with prevalence estimates ranging from 10% to 64% [1–6]. Variations in the prevalence of poststroke depression (PSD) have been attributed to study/care setting, time lapse after stroke, and method of assessing depression [7]. Although mental health is typically evaluated by history taking and clinical assessment, patient self-reported questionnaires are increasingly used to evaluate patient outcomes [8]. Self-rated scales are not a substitute for clinical diagnosis of depression but they are useful as screening tools. Such scales include the Self-Rating Depression Scale [9], Beck Depression Inventory [10], and Center for Epidemiologic Studies-Depression (CES-D) scale [11].

Although the CES-D was developed for depression screening in the general population, it is often used in studies of stroke outcome [7,12–16]. Two investigations have provided some evidence to support interobserver reliability [17], construct validity [6,17], and sensitivity and specificity of the CES-D as a screen for depression after stroke [6]. Nevertheless, PSD may be etiologically different from depression in the general population [7], and studies comparing the symptom profiles of poststroke and functional depression have suffered from various methodological limitations [18].

One approach to understand scale equivalence in different groups or conditions, such as depression symptoms after stroke compared with primary care, is to use Item Response Theory (IRT)-based models. A fundamental principal that underlies the IRT family of models is that an individual's response to any given item reveals a level of ability in the trait being measured [19]. Unlike classical test theory-based approaches, IRT-based models can be used to evaluate item hierarchy along a single continuum (i.e., item dif-

Address correspondence to: A. Simon Pickard, College of Pharmacy, University of Illinois at Chicago, Room 164, MC 886, 833 South Wood Street, University of Illinois At Chicago, Chicago, IL 60612, USA. E-mail: pickard1@uic.edu

10.1111/j.1524-4733.2006.00082.x

ficulty order), and test whether item hierarchies are consistent across different samples of patients and test occasions [20]. IRT-based models have emerged as powerful tools to assist in the development of item banks for computer adaptive testing, helping to identify items that are stable across different groups and provide maximum information about a trait, for example, fatigue or depression, without requiring large numbers of items [21]. In this respect, IRT-based models may be used to inform clinicians and outcomes researchers about whether differences exist between primary care- and stroke-based depression based on patterns of responses to the CES-D. In this study, we utilized IRT-based models to compare responses to symptom-related items on the CES-D by depressed stroke and primary-care patients for several purposes: 1) to illustrate the use of IRT-based (Rasch) models for comparing scale functioning across different patient subgroups; and 2) to inform clinicians and outcome researchers about CES-D scale functioning and depressive symptomatology in stroke- compared with primary care-based depression.

Subjects and Methods

This study utilized secondary data from two sources. Stroke patient data were collected as part of a study of health outcomes after stroke, which has been described in detail elsewhere [16]. Patient assessments of the CES-D collected at 3 months postrecruitment were analyzed. Patients who were diagnosed with ischemic stroke were recruited with 2 weeks of stroke before discharge from two large teaching hospitals in Edmonton, Canada. Participants had to be 18 years of age or older. Patients were not eligible if they had receptive aphasia, were cognitively impaired, unable to comprehend the English language, or had a very poor prognosis, based on consultation with attending clinicians.

Data on USA-based primary-care patients with depression were obtained from the Longitudinal Investigation of Depression Outcomes (LIDO) study. The LIDO project was a longitudinal study of depressive symptoms, quality of life and health services use among primary-care patients in Barcelona (Spain), Be'er Sheva (Israel), Melbourne (Australia), Porto Alegre (Brazil), St. Petersburg (Russia), and Seattle (the United States). At each site, consecutive adult (age 18–75) visitors were invited to participate in the study. After providing informed consent, participants completed a screening questionnaire that included the CES-D. The study has been extensively described elsewhere [22,23].

Measures

The CES-D is a 20-item scale designed to measure depressive symptoms experienced in the past week (Table 1) [11]. Responses are interpreted based on a simple summary score, calculated by summing the item responses. Patients were categorized as depressed if CES-D scores were 16 or higher, a threshold used in community-based studies [11,24] and in stroke [6,17]. Higher scores are associated with more frequent depressive symptoms.

Analysis

Among the family of IRT-based models used to characterize item function, the most common are the 1-parameter (Rasch), 2-parameter, and 3-parameter logistic models. The first parameter is item location, which represents the item's severity or difficulty con-

Table 1 CES-D items

Item statement	Abbreviation
I was bothered by things that usually don't bother me	Bothered
I did not feel like eating: my appetite was poor	Appetite
I felt that I could not shake off the blues even with help from family members	Blues
I felt I was just as good as other people	As good as
I had trouble keeping my mind on what I was doing	Concentration
I felt depressed	Depressed
I felt that everything I did was an effort	Effort
I felt hopeful about the future	Hopeful
I thought my life had been a failure	Failure
I felt fearful	Fearful
My sleep was restless	Restless
I was happy	Happy
I talked less than usual	Talked less
I felt lonely	Lonely
People were unfriendly	Unfriendly
I enjoyed life	Enjoy life
I had crying spells	Crying
I felt sad	Sad
I felt that people disliked me	Disliked
I could not get going	Get going

CES-D, Center for Epidemiologic Studies-Depression scale.

cerning the trait of interest, whereas the 2-parameter and 3-parameter models additionally take into account item discrimination and guessing, respectively. There are many Rasch models, all of which are derived from a basic probability model that assumes no relationship exists between a person's responses to different items after taking their ability into account. The correspondence between an individual's ability on a latent trait and the predicted response to an item is represented by an item characteristic curve [19], which has an ogival (S-shaped) form. Item location along the continuum of the measure is expressed in log odds, or logits. The 1-parameter Rasch model was selected for the present study because it can be used to select items that "fit," that is, have equal item characteristic curves, across different populations, and it requires smaller samples to derive stable parameter estimates compared with other models [21].

The model appropriate for ordered response categories, the Rasch Rating Scale model [25], was used to evaluate item hierarchy and item fit statistics on the CES-D. Item fit was evaluated using goodness-of-fit statistics, reported as infit mean-squares (MNSQ). MNSQ is the ratio of the observed to the predicted variance for an item, and informs how well each item functions within the scale [26]. Guidelines for rating scales suggest items with MNSQ values higher than 1.4 misfit the scale [27], indicating it either taps a different dimension or that it differs from other items in its ability to discriminate people [21].

Item hierarchies were compared between the stroke and primary-care groups by examining mean item calibrations and item order. Differential item functioning (DIF), which refers to an item lacking equivalence in performance/functioning in different groups or settings, was identified statistically by conducting *t*-tests on mean item calibrations between the stroke and primary care-based patient groups [28]. The *t*-statistic was computed as a ratio of difference between item difficulty estimates and the square root of the sum of squared standard errors at a 95% confidence level. Sample sizes larger than 30 have been considered adequate for demonstrating item calibration stability within ±1 logit with 95% confidence [29]. Because the identification of DIF using *t*-tests is dependent on sample size, an alternative criterion of 0.5 logit between item calibrations is often used to determine DIF [21, 29, 30]. Rasch analysis was conducted with Winsteps® [31].

Results

Of the 101 stroke patients with complete item responses to the CES-D at 3 months (three had incomplete data), 32 patients (32%) were defined as depressed according to CES-D scores. Stroke patients defined as depressed had a mean age of 69 years (SD 14) and 59% were female. The mean age of primary care-based patients with depression was 42 years (SD 15), and the majority (67%) of patients were female [22].

Item hierarchies based on mean logit calibrations were strongly correlated between stroke and primary care-based groups (Spearman's rank order coefficient, $r_s = 0.75$). Items relating to interpersonal disruption feelings ("I felt that people disliked me" and "people were unfriendly") were hardest to endorse in both groups. Misfitting items, that is, MNSQ higher than 1.40, in PSD included "my sleep was restless," "I had crying spells," "people were unfriendly," and "I felt just as good as other people." No items misfit the scale in the primary care-based depression group. In comparing item functioning between the two groups, four items demonstrated statistically significant DIF: "my sleep was restless," "I felt that people disliked me," "I did not feel like eating," and "I had crying spells." Each of these items identified with statistically significant DIF demonstrated a logit difference of approximately 0.5 or more across the two groups. (Table 2)

Discussion

Although classical test theory-based methods have generally supported construct validity and internal consistency reliability of the CES-D in stroke and community-based studies, such methods can not facilitate the evaluation of whether items are equivalent in meaning to different individuals [19]. We investigated the fit and difficulty level of the CES-D items along the continuum of the measure using the 1-parameter IRT (Rasch) model. The CES-D scale functioning was found to be quite impressive in primary care-based depression, with no items misfitting the scale, although only three items slightly misfit the scale in stroke. DIF observed between depressed stroke and primary-care patients may imply that slightly different clusters of depressive occur in stroke compared with primary-care patients, but this is conjectural given the small size of the stroke sample. In addition, the same items have been previously associated with bias in studies of large nonstroke samples.

Item-response bias in the CES-D has been examined in large samples containing 708 cancer patients and 504 caregivers of the chronically ill elderly [32], and using data (n = 2340) from the Established Populations for Epidemiologic Studies of the Elderly (EPESE) [33]. Use of confirmatory factor analytic models in caregivers of the chronically ill elderly and cancer patients identified two items with gender bias ("I had crying spells" and "I talked less") and three additional items with psychometric problems ("people are unfriendly," "people dislike me," and "I thought my life had been a failure") [32]. The EPESE study indicated the CES-D would have greater validity after removal of the two

Table 2 CES-D item calibrations and DIF in depressed stroke and primary-care patients

	Stroke (n = 32)		Primary care (n = 366)		Differences in item calibration	
Item	Mean logit calibration (SEM)	Infit MNSQ	Mean logit calibration (SEM)	Infit MNSQ	|*t*-stat|	|Logit difference|
Disliked*	1.65 (0.31)	0.96	0.88 (0.07)	0.89	2.42	0.77
Unfriendly	1.10 (0.25)	1.53†	0.84 (0.08)	1.21	0.99	0.26
Failure	0.53 (0.21)	0.85	0.50 (0.07)	0.99	0.14	0.03
Crying*	0.36 (0.20)	1.59†	0.84 (0.07)	1.07	2.27	0.48
Fearful	0.24 (0.20)	0.88	0.53 (0.07)	0.96	1.37	0.29
Blues	–0.02 (0.19)	0.69	0.21 (0.07)	0.69	1.14	0.23
Appetite*	–0.06 (0.19)	1.01	0.59 (0.08)	1.39	3.15	0.65
Effort	–0.10 (0.19)	1.04	–0.45 (0.06)	0.87	1.76	0.35
Talked less	–0.10 (0.19)	1.11	0.19 (0.07)	1.07	1.43	0.28
Restless*	–0.17 (0.19)	1.63†	–0.78 (0.07)	1.34	3.01	0.61
Sad	–0.20 (0.19)	0.61	–0.26 (0.07)	0.62	0.30	0.06
As good as	–0.28 (0.19)	1.37	–0.09 (0.07)	1.40	0.94	0.48
Concentration	–0.28 (0.19)	1.02	–0.23 (0.07)	0.93	0.25	0.29
Depressed	–0.28 (0.19)	0.59	–0.34 (0.06)	0.65	0.30	0.23
Get going	–0.28 (0.19)	1.17	–0.63 (0.06)	0.97	1.76	0.35
Bothered	–0.31 (0.19)	0.50	0.05 (0.07)	0.94	1.78	0.36
Hopeful	–0.35 (0.19)	1.15	–0.45 (0.07)	1.37	0.49	0.10
Happy	–0.49 (0.19)	0.68	–0.69 (0.07)	0.97	0.99	0.20
Lonely	–0.49 (0.19)	1.00	–0.19 (0.06)	0.84	1.51	0.30
Enjoy life	–0.49 (0.19)	0.93	–0.52 (0.07)	0.91	0.15	0.03

*DIF *t*-statistic *P*-value = 0.05; †MNSQ > 1.4.

CES-D, Center for Epidemiologic Studies-Depression scale; DIF, differential item functioning; MNSQ, mean-squares; SEM, standard error of the mean.

interpersonal items and the crying item, and did not find support for gender bias on the item "I talked less" [33]. Thus, despite the small size of the stroke sample in the present study, the same three items from the EPESE study were identified as psychometrically problematic using of Rasch models. Thus, the psychometric issues raised by items in the present study do not appear to be unique to stroke.

Given the widespread use of the CES-D, items from the scale are likely to be prominent in the development of item banks for clinical and research studies that utilize computer adaptive testing. Absence of DIF is the first step in selecting core sets of items for item banks used in computer adaptive testing [21], and this comparison of item functioning between depression after stroke and in primary care demonstrated that Rasch models can help to identify items that are desirable as core items in an item bank intended to assess depression outcomes across different patient groups. The identification of DIF can serve as a tool to help clinicians to identify clinically relevant characteristics or attributes that differ across patient groups, such as sleep-related problems in depressed stroke compared with other depressed patients groups, as well as for hypothesis generation for future studies. This investigation also illustrates how Rasch models may be utilized to help evaluate scale functioning in different clinical conditions, and potentially lead to refinements in the measure if substantial gains in measurement precision can be realized.

Enhanced measurement precision and less measurement error when evaluating depression outcomes offers the benefit of smaller sample sizes needed to detect significant differences between groups, reducing the resources and efforts needed from both clinical and outcomes researchers when designing and implementing studies. Nevertheless, the extent of the psychometric flaws should be considered in the context of cost and effort involved in scale revisions, because revisions affect interpretability [34], and would necessitate further validation studies of ability to predict interview-based clinical diagnoses. Given that only one item was identified as uniquely psychometrically problematic in the stroke subgroup, stroke-specific changes to the CES-D scale are not recommended.

It is important to note that scale-related issues identified using Rasch models assume unidimensionality of the scale. The factor structure of the CES-D in non-stroke patients has been described as having four underlying factors: depressive affect, positive affect, somatic and interpersonal disruption [11]. Nevertheless, McDowell and Newell [35] cautioned against the interpretation of factor analysis of the CES-D, stating that although factor analyses have been relatively consistent they should not be used as a basis for identifying subscores because the factors intercorrelate, the scale as a whole has high internal consistency, and not all items load significantly on the factors. Furthermore, some CES-D items are causal indicators [36]; items that relate to causes of depression. Nevertheless, the converse need not apply: patients with depression may not experience the same set of symptoms. Thus, psychometric evaluations of the CES-D should be tempered by considerations such as the multidimensional nature of depression and inclusion of clinically relevant items that do not significantly/consistently load on the same factors. At the same time, the high internal consistency of the CES-D and the fact that interpretation is based on a summary score that gives equal weighting to each item argue in favor of the exploration of the CES-D as a unified scale using Rasch models.

Generalizability of the results to stroke patients may be limited by the timing of the assessment, which was assessed at approximately 3 months poststroke. The nature of PSD may change over the course of recovery and PSD does not remain constant throughout the poststroke period [5]. In relation to other studies assessing depression at 3 months poststroke, the 25% to 30% prevalence of depression observed in other studies was comparable to the 32% prevalence in the present study [1,7,37]. Use of the CES-D to define depression rather than clinical diagnosis is an acknowledged limitation. Because of the small stroke sample size, the present study has been described as a preliminary investigation.

In summary, the application of the Rasch model to the CES-D scale generated results that generally supported the validity of the CES-D scale, particularly in primary care-based depression. Items related to crying, unfriendliness, and feeling disliked that demonstrated DIF in the comparison between stroke- and primary care-based depression are not a stroke-specific concern because bias on those items has been previously identified in large nonstroke samples. Although Rasch models have limitations and require caution in their interpretation when applied to a condition such as depression, they can provide unique insight into the validity and reliability of outcome measures in different patient groups in clinical trials as well as in community-based studies.

Source of financial support: AstraZeneca provided investigator-initiated grant support for the collection of the data on stroke patients.

References

1 Berg A, Palomaki H, Lehtihalmes M, et al. Poststroke depression: an 18-month follow-up. Stroke 2003;34: 138–43.

2 Kotila M, Numminen H, Waltimo O, et al. Depression after stroke: results of the FINNSTROKE Study. Stroke 1998;29:368–72.

3 Andersen G, Vestergaard K, Riis J, et al. Incidence of post-stroke depression during the first year in a large unselected stroke population determined using a valid standardized rating scale. Acta Psychiatr Scand 1994;90:190–5.

4 Shima S, Kitagawa Y, Kitamura T, et al. Poststroke depression. Gen Hosp Psychiatry 1994;16:286–9.

5 Hosking SG, Marsh NV, Friedman PJ. Poststroke depression: prevalence, course, and associated factors. Neuropsychol Rev 1996;6:107–33.

6 Parikh RM, Eden DT, Price TR, et al. The sensitivity and specificity of the Center for Epidemiologic Studies Depression Scale in screening for post-stroke depression. Int J Psychiatry Med 1988;18:169–81.

7 Aben I, Verhey F, Honig A, et al. Research into the specificity of depression after stroke: a review on an unresolved issue. Prog Neuropsychopharmacol Biol Psychiatry 2001;25:671–89.

8 Jenkinson C, Fitzpatrick R, Garratt A, et al. Can item response theory reduce patient burden when measuring health status in neurological disorders? Results from Rasch analysis of the SF- 36 physical functioning scale (PF-10). J Neurol Neurosurg Psychiatry 2001;71:220–4.

9 Zung WW, Richards CB, Short MJ. Self-rating depression scale in an outpatient clinic. Further validation of the SDS. Arch Gen Psychiatry 1965;13:508–15.

10 Beck AT, Ward CH, Mendelson M, et al. An inventory for measuring depression. Arch Gen Psychiatry 1961;4:561–71.

11 Radloff LS. The CES-D Scale: a self-report depression scale for research in general population. Appl Psych Meas 1977;1:385–401.

12 Beekman AT, Penninx BW, Deeg DJ, et al. Depression in survivor of stroke: a community-based study of prevalence, risk factors and consequences. Soc Psychiatry Psychiatr Epidemiol 1998;33:463–70.

13 Colantonio A, Kasl SV, Ostfeld AM, et al. Psychosocial predictors of stroke outcomes in an elderly population. J Gerontol 1993;48(Suppl.):S261–8.

14 Kim P, Warren S, Madill H, et al. Quality of life of stroke survivors. Qual Life Res 1999;8:293–301.

15 Ramasubbu R, Robinson RG, Flint AJ, et al. Functional impairment associated with acute poststroke depression: the Stroke Data Bank Study. J Neuropsychiatry Clin Neurosci 1998;10:26–33.

16 Pickard AS, Johnson JA, Feeny DH, et al. Agreement between self- and proxy assessment in stroke: a comparison of generic HRQL measures. Stroke 2004;35:607–12.

17 Shinar D, Gross CR, Price TR, et al. Screening for depression in stroke patients: the reliability and validity of the Center for Epidemiologic Studies Depression Scale. Stroke 1986;17:241–5.

18 Whyte EM, Mulsant BH. Post stroke depression: epidemiology, pathophysiology, and biological treatment. Biol Psychiatry 2002;52:253–64.

19 Hays RD, Morales LS, Reise SP. Item response theory and health outcomes measurement in the 21st century. Med Care 2000;38:II28–42.

20 Haley SM, McHorney CA, Ware JA. Evaluation of the MOS SF-36 physical functioning scale (PF-10): I. Unidimensionality and reproducibility of the Rasch item scale. J Clin Epidemiol 1994;47:671–84.

21 Lai JS, Cella D, Chang CH, et al. Item banking to improve, shorten and computerize self-reported fatigue: an illustration of steps to create a core item bank from the FACIT-Fatigue Scale. Qual Life Res 2003;12:485–501.

22 Herrman H, Patrick DL, Diehr P, et al. Longitudinal investigation of depression outcomes in primary care in six countries: the LIDO study. Functional status, health service use and treatment of people with

depressive symptoms. Psychol Med 2002;32:889–902.

23 Bech P, Lucas R, Amir M, et al. Association between clinically depressed subgroups, type of treatment and patient retention in the LIDO study. Psychol Med 2003;33:1051–9.

24 Weissman MM, Sholomskas D, Pottenger M, et al. Assessing depressive symptoms in five psychiatric populations: a validation study. Am J Epidemiol 1977;106:203–14.

25 Andrich D. A rating formulation for ordered response categories. Psychometrika 1978;43:561–73.

26 Smith EV. Understanding Rasch measurement: detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals. J Appl Meas 2002;3:205–30.

27 Wright BD, Linacre JM, Gustafson J-E, et al. Reasonable mean-square fit values. Rasch Measurement Transaction 1994;8:370. Available at http://www.rasch.org/rmt/rmt836.htm, accessed November 23, 2005.

28 Smith RM. Pre/post comparisons in Rasch measurement. In: Wilson M, Englehard G, Draney K, eds., Objective Measurement: Theory Into Practice, Vol. 4. Norwood: Ablex Publishing Corporation, 1997.

29 Linacre M. Sample size and item calibration stability. Rasch Measurement Transaction 1994;7:328.

30 Lee KO. Variance in mathematics and reading across grades. Rasch Measurement Transaction 1992;6: 222–3.

31 Linacre JM, Wright BD. A user's guide to WINSTEPS: Rasch Model Computer Program. Rasch Measurement. Chicago: MESA Press, 2001.

32 Stommel M, Given BA, Given CW, et al. Gender bias in the measurement properties of the center for epidemiologic studies depression scale (CES-D). Psychiatry Res 1993;49:239–50.

33 Cole SR, Kawachi I, Maller SJ, Berkman LF. Test of item-response bias in the CES-D scale: experience from the New Haven EPESE study. J Clin Epidemiol 2000;53:285–9.

34 Orlando M, Sherbourne CD, Thissen D. Summed-score linking using item response theory: application to depression measurement. Psychol Assess 2000;12: 354–9.

35 McDowell I, Kristjansson E. Depression. In: McDowell I, Newell C, eds., Measuring Health: A Guide to Rating Scales and Questionnaires. New York: Oxford University Press, 1996.

36 Fayers PM, Hand DJ. Factor analysis, causal indicators and quality of life. Qual Life Res 1997;6:139–50.

37 Wade DT, Legh-Smith J, Hewer RA. Depressed mood after stroke: a community study of its frequency. Br J Psychiatry 1987;151:200–15.