Journal of Depression and Anxiety
ISSN: 2167-1044

Tusiime et al., J Depress Anxiety 2015, 4:2
DOI: 10.4184/2167-1044.1000184

Research Article Open Access

Examining the Psychometric Properties of the Beck Depression Inventory-Ii Using an Item Response Modelling Approach in an HIV Infected Population in Kampala, Uganda

Jayne Byakika Tusiime*, David R Bangsberg and Wilson Mark

Makerere University School of Public Health, Kampala, Uganda

Abstract

Background: Depression is prevalent among individuals living with HIV, with evidence suggesting that more than one third of people with HIV/AIDS may have mood disorders or clinically significant depressive symptoms. Sub-Saharan Africa bears the greatest burden due to HIV/AIDS. The social, economic and health impact of depression in sub-Saharan Africa is great. However, there are few scales for measuring depression that have been validated in this population. The Beck Depression Inventory (BDI-II) is one of the most widely used instruments for assessing depressive symptom severity. Although the BDI-II has been used in sub-Saharan Africa, the psychometric properties of the BDI have not been well studied in this region especially among HIV infected patients. The purpose of this analysis therefore, was to examine the psychometric properties of the BDI-II in a low income HIV-infected population using an item response modeling (IRM) approach.

Methods: Data for this analysis were obtained from a survey of adult members of the Mother-To-Child-Transmission Plus Program in Kampala, Uganda. The BDI-II was administered to every study participant at enrollment. Data were analyzed using both the Classical Test Theory (CTT) approach and the Item Response Modeling (IRM) approach.

Results: Mean depression score was 8.86 + 5.44. The Cronbach's alpha was 0.79 and the reliability coefficient was 0.86. The Wright map showed a good spread of the items over the entire span of the construct of depression. Differential item functioning was observed for some items. There was evidence for validity based on instrument content, internal structure and relations to other external variables.

Conclusion: In conclusion, this analysis demonstrated good psychometric properties of the BDI-II when used to screen for depression in a low income predominantly female HIV-infected population. These findings therefore support the use of the BDI-II in assessing depressive symptoms for HIV-infected patients in sub-Saharan Africa especially women.

Keywords: Beck depression inventory II; Validation; HIV-infected people; Sub-Saharan Africa

Introduction

Depression is prevalent among individuals living with HIV, with evidence suggesting that more than one third of people with HIV/AIDS may have mood disorders or clinically significant depressive symptoms [1-6]. A meta-analysis comparing rates of depression disorder in HIV-positive and at risk HIV- negative patients demonstrated a twofold increase in the prevalence of major depression in patients infected with HIV [3].

Sub-Saharan Africa (SSA) is home to over two thirds of all patients living with HIV, with AIDS being the leading cause of death [7]. It is estimated that about 1.6 million people in Uganda are living with HIV/AIDS with an HIV prevalence of 7.3%. There is a dearth of literature on depression among individuals living with HIV/AIDS in sub-Saharan Africa. The sparse literature shows elevated rates of depression among HIV-infected individuals relative to community samples [5,8-16] that is consistent with the developed world [14-17].

The social, economic and health impact of depression in sub-Saharan Africa is great. Depression is associated with mortality [18-21], work disability [20,22-24], lower quality of life [19,25-29], risk of heart disease [30] and high-risk behaviors for contracting HIV infection [31]. Depression is associated with lowered adherence to antiretroviral medication [32,33].

Depression in sub-Saharan Africa presents in forms (culture specific idioms, somatic, based on interpersonal relationships or spiritual in nature) that may obscure detection [34-36]. However, depression exists at possibly higher prevalence rates than in western countries [35] according to several studies [19-22,26-29,36-44] of community and non-HIV specific clinic populations, with generally higher rates for women than men [23,40]. Finally, though depression may present differently in sub-Saharan Africa compared to western countries, it [depression] is reasonably easy to elicit when present and sought [23].

Depression in SSA has been measured using standard tools developed in western countries like the Beck Depression Inventory (BDI) [21,32,40,45,46]; Hopkins Symptom Checklist (HSCL) [18,22,38,40,44,47]; the Center for Epidemiologic Studies Depression Scale (CES-D) [31,39]; the Edinburgh Postanatal Depression Scale (EPDS) [36,42]; the Patient Health Questionnaire (PHQ) [5,38]; the World Mental Health Survey of WHO Composite International Diagnostic Interview (WMH-CIDI) [19] and the Composite

***Corresponding author:** Jayne Byakika Tusiime, B.Pharm, M.Sc., Ph.D, Makerere University School of Public Health, Kampala, Uganda, Tel: +256 (0) 414 269 003; E-mail: tusjayne@hotmail.com

Received November 25, 2014; **Accepted** April 28, 2015; **Published** April 30, 2015

Citation: Tusiime JB, Bangsberg DR, Wilson M. (2015) Examining the Psychometric Properties of the Beck Depression Inventory-Ii Using an Item Response Modelling Approach in an HIV Infected Population in Kampala, Uganda. J Depress Anxiety 4: 184. doi:10.4184/2167-1044.1000184

International Diagnostic Interview (CIDI) [20]. There have been few attempts to validate these instruments in this culturally different setting [5,27,38-40,46] and to adapt the instrument to site specific culture and terminology [44]. While some may argue that instruments developed in one culture can be used in another provided careful attention is paid to conceptual translation [47], this has not always been done when using these standard instruments. The most common practice is to do a direct translation of the instrument from its original language to the local language [20,31]. This kind of translation does not always ensure conceptual translation and may even sometimes change the meaning of the concept under translation.

An important limitation of most standard instruments or diagnostic measures is that they may be culturally inappropriate unless proved otherwise [34]. Descriptions of certain health conditions like mental illness developed in one set of cultures may not be equally applicable to other cultures. Given this assumption, if the nature of emotions, thoughts, and behaviors can be expected to vary by culture [48], uncritical application of standard instruments cross-culturally might yield misleading or erroneous results where such instruments are locally inappropriate [49]. Inconsistent findings in any research effort may result from random processes and non-equivalent measures, procedures, or samples, but may also be explained by problems of low validity. Problems of validity are not new to epidemiology [50], but are more likely to occur in transcultural epidemiology which is defined here as research in which the views, concepts or measures of the investigator extend beyond the scope of one cultural unit to another [51].

Context affects research validity. Evidence of measurement validity and reliability cannot be assumed to generalise across populations. This lack of generalisability may be especially problematic when the original measure is translated into another language, as is common in transcultural studies. Creating a culturally acceptable, comprehensible, relevant and semantically equivalent translation is difficult [52] making it essential to study the psychometric properties translated measures that might have changed during imperfect translations. While we acknowledge the limitations of these 'foreign' generated instruments, we do not say that these instruments are irrelevant for studying health conditions in other contexts but rather that their appropriateness will vary across cultures and by illness, and therefore needs to be explored. As such, instruments may need to be adapted to the new situations or in other cases completely new instruments developed.

The BDI is one of the most widely used instruments for assessing depressive symptom severity. Not only is it used for assessing the intensity of depression in psychiatrically diagnosed patients [53], but also for detecting depression in nonpsychiatric patients with conditions like HIV/AIDS, diabetes and other chronic conditions [32,40,45,54,55]. It has also been used to detect depressive symptoms in normal populations.

The BDI has been used to assess depressive symptoms in sub-Saharan Africa [39,40,45,56]. However, the psychometric properties of the BDI have not been well studied in SSA especially among HIV infected patients.

Classical Test Theory (CTT) has been the underpinning for most test construction and theory. Item response theory (IRT) is a new approach to test development [57,58]. It is an improvement of the CTT approach. The two main achievements of IRT is the ability to separately estimate parameters about the test items and the people taking the test at the same time under the same scale. CTT focuses primarily on test-level information, whereas IRT focuses on item-level information.

Using IRT, we set out to characterize the BDI in a low income HIV-infected population taking antiretroviral medications. It is particularly important to characterize depression in this population because depression is associated with decreased adherence and disease progression.

Methods

Study design

This analysis is based on data collected from a cross-sectional survey of 123 adults attending the Mother To-Child-Transmission Plus (MTCT Plus) Program at Mulago Hospital in Kampala, Uganda.

Study setting

The study setting has been described in detail elsewhere [32] Briefly, Mulago Hospital is Uganda's largest teaching, referral, and research hospital. The MTCT Plus program is located at Mulago Hospital under the Makerere University-Johns Hopkins University Research Collaboration.

Study procedures

Details of recruitment and data collection have been reported elsewhere [32]. Briefly, we recruited individuals who were newly initiating ART and individuals who were on chronic ART. In addition to the other data collected, the research assistants administered the BDI-II at the beginning of the study to detect depressive symptoms. The instrument took about 10 minutes to administer.

The instrument

The BDI-II is a 21-Guttman item test presented in multiple choice format which purports to measure presence and degree of depression in adolescents and adults. The BDI was derived from clinical observations about the attitudes and symptoms displayed frequently by depressed psychiatric patients and infrequently by nondepressed psychiatric patients [59]. The clinical observations were consolidated systematically into 21 symptoms and attitudes which could be rated from 0 to 3 in terms of intensity. The items were chosen to assess the intensity of depression and were not selected to reflect a particular theory of depression. The 21 symptoms and attitudes were: (1) Mood, (2) Pessimism, (3) Sense of Failure, (4) Lack of Satisfaction, (5) Guilt Feelings, (6) Sense of Punishment, (7) Self-dislike, (8) Self-accusation, (9) Suicidal Wishes, (10) Crying, (11) Irritability, (12) Social Withdrawal, (13) Indecisiveness, (14) Distortion of Body Image, (15) Work Inhibition, (16) Sleep Disturbance, (17) Fatigability, (18) Loss of Appetite, (19) Weight Loss, (20) Somatic Preoccupation, and (21) Loss of Libido.

Analyses

We conducted univariate analyses of the baseline characteristics of our study population, including socio-demographic and clinical variables. Categorical variables are presented as frequencies and percents while continuous variables are presented as medians and interquartile ranges [IQRs]. Depression responses for each subject were scored by summing the ratings given to each of the 21 items. The depression scores were then categorized into four categories according to a recommended guideline: none or minimal depression, <10; mild to moderate depression, 10-18; moderate to severe depression, 19-29; and severe depression, 30-63.

Classical test theory (CTT): Item and test characteristics were first evaluated using CTT analysis. Item parameters included the item mean

(item difficulty) and point biserial correlations. According to Varma (Varma, 2009), a point-biserial value of at least 0.15 is recommended. The internal consistency reliability of the test was calculated using Cronbach's alpha and adequate reliability was demonstrated with a reliability index of at least 0.70 [60]. The CTT analyses were performed using Conquest software [61].

Item response modeling: We used an item response modeling approach to calibrate the data [58]. Preference of IRT over CTT has been a result of several limitations of CTT [62]. The primary limitation of CTT is that the item and scale statistics apply only to the specific group of subjects who took the test. That means that if the scale is to be administered to people who are different in some way, such as being a member of a minority group or patients rather than students, then it is necessary to re-establish its psychometric properties and perhaps develop new norms. Similarly we would have to go through the same renorming process if any of the items were altered, or if items were deleted in order to develop a shorted version of the scale. The second problem is that it is impossible to separate out the properties of a test from the attributes of the people taking it. Thus the instrument's characteristics change as we test different groups. A third problem is the assumption that each item contributes equally to the final score. While this is also true for IRT, the important difference is that we test to see if this is acceptable, that is we test for fit of the items. In CTT, the items are simply summed up irrespective of how much each item correlates with the underlying construct. Fourth, it is also assumed that each item is measured on the same interval scale. This assumption often fails on two grounds: items are most often ordinal rather than interval, and the 'psychological distance' between response options differs from one item to the next [63]. IRT has been designed to overcome these limitations [64]. One advantage of IRT is that the scale that emerges from an IRT analysis truly has interval-level properties. Secondly, it provides a more precise estimate of measurement error. Third, when the model fits, it is invariant theoretically: that is, that item characteristics are independent of the sample from which they are derived.

Analysis was done using Conquest [61] and ConstructMap [65] software. These two soft wares are based on Rasch's family of logistic models. Two models for ordinal data were considered: the rating scale model (RCM) [66] and the partial credit model (PCM) [67]. The PCM does not assume that the distances between ordinal responses are the same for all items (i.e the distance between response options a and b, b and c, etc. is not the same across all items). In contrast, the RCM assumes that the distances between ordinal responses are the same for all items (i.e. the distance between response options a and b, and b and c, etc is the same for all items). The best fitting model was assessed by comparing the deviance parameters as well as the weighted fit indices for the items and the respondents (i.e. infit statistics). Item misfit was established by infit mean square (MSQ) values outside the range of 0.75 to 1.33, with significant t values [58]. For the weighted t statistic, any value <-2.00 or > 2.00 was indicative of misfit. The likelihood ratio test (LRT) and MSQ were used to determine the best fitting model.

Using the estimated parameters, a Wright map was constructed. The Wright map is an empirical map based on respondents' self-reports [58]. The Wright map of person-item estimates displays the item locations and the extent to which the item locations reflected the full range of depressive symptoms. It is useful for identifying a set of items that span the range of the construct. This map served to qualitatively assess content representation by determining if the items assessed minimal, moderate or severe depression.

To investigate the reliability of the instrument, first the standard error of measurement for each person location was estimated. Using the standard error of measurement, the information for the whole instrument was calculated. The information for the whole instrument is the sum of the information for each item, where information for each item is the reciprocal of the square of the standard error of measurement. A quality-control index of consistency was developed. Internal consistency was assessed by calculating the reliability coefficient using the expected a posteriori (EAP) procedure.

Validity of the instrument was investigated using evidence based on instrument content, internal structure and relations to other variables [68]. These are the most recent standards for educational and psychological testing. To assess presence of internal structure, we investigated consistency of the items with the instrument as a whole using mean locations of each group. In addition, differential item functioning (DIF) was investigated to determine if items functioned in a similar way for respondents across important subgroups. Finally, we explored evidence based on relations to other external variables that the construct [depression] should predict. In particular, we explored the relationship between depression and adherence to antiretroviral medications.

Ethical considerations: All participants provided informed consent and all procedures were approved by the Faculty of Medicine Research and Ethics Committee and the Institutional Ethics Review Board of Makerere University, the Uganda National Council of Science and Technology, and the University of California, San Francisco Committee on Human Subjects.

Results

One hundred twenty three patients were included in the analysis. The study population was predominantly female with low income. Patient characteristics are summarized in Table 1. Fifty six percent (n=63) scored 10 or less on the whole instrument Table 2.

Item response modeling

The Partial Credit Model was used for analysis. Compared to the rating scale model, the partial credit model fitted the data better as one would expect, given the nature of the alternative responses to the items. (X2=172.0582; df=24, p value<0.0001)

Characteristic	n (%)
Sex	
Female	92 (75)
Male	31 (25)
Median age (years, IQR)	32 (28-36)
Education attainment	
Primary	62 (50.4)
Post Primary	61 (49.6)
Median income/mo (USD, IQR)	75 (25-100)
CD4 cell count <250 cells/ul	63 (51)
CD4 cell count ≥250 cells/µl	60 (49)

Table 1: Patient characteristics.

Category	Frequency N=112* (%)
None to mild (<10)	63 (56)
Moderate (10-18)	45 (40)
Moderate to severe (19-29)	4 (4)
*includes only participants with complete data	

Table 2: Distribution of depression scores.

Item fit

Weighted fit statistics showed a good fit for most of the items. The weighted mean square range was 0.83 to 1.43. As an effect size, there is no absolute limit to what is a good means weighted square value, but previous researchers have indicated that 0.75 is a reasonable lower bound and 1.33 is a reasonable upper bound [58]. Item 11 (agitation) showed the worst fit. Infit mean squares are given in Figure 1.

Wright map

The right-hand side of the map (Figure 2) shows the calibrated item threshold locations. The kth Thurstone threshold is the point at which the probability of the scores below k is equal to the probability of the scores k and above. Because there are four response categories, there

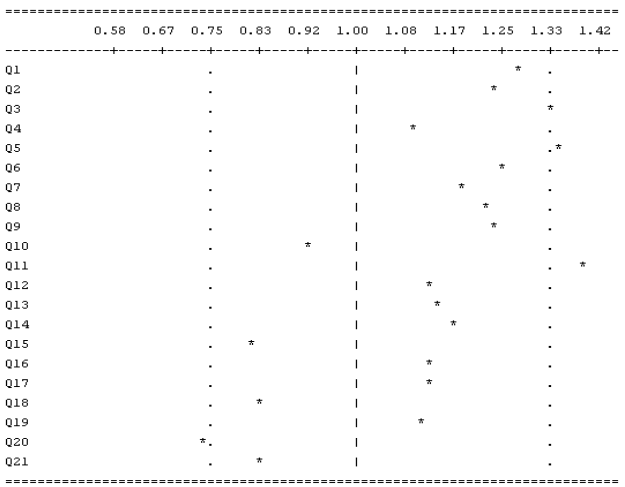

Figure 1: Infit mean squares.

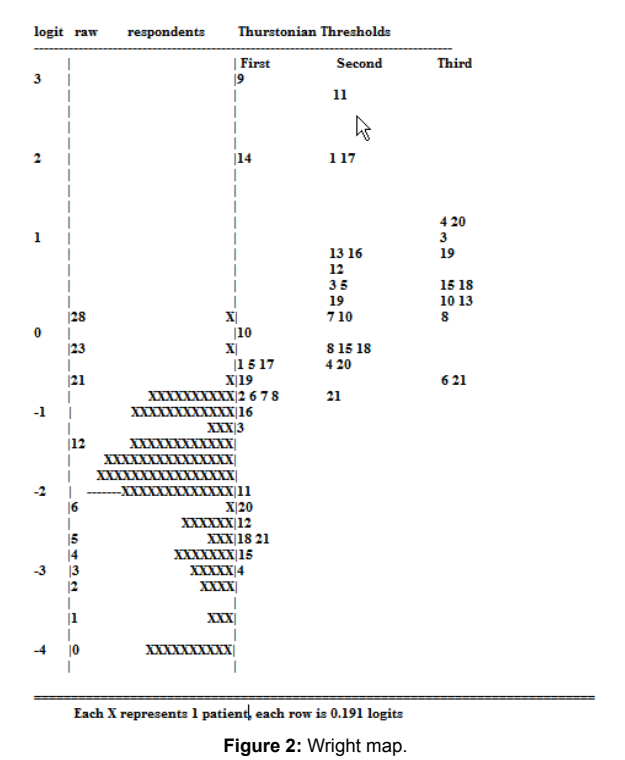

Figure 2: Wright map.

Item	Item Description	Point-biserial correlation for each category			
		0	1	2	3
1	Mood	-0.36	0.35	0.05	-
2	Pessimism	-0.53	0.53	-	-
3	Sense of failure	-0.43	0.40	0.11	0.05
4	Lack of satisfaction	-0.55	0.22	0.25	0.34
5	Guilt feelings	-0.50	0.40	0.26	-
6	Sense of punishment	-0.50	0.32	0.38	-
7	Self-dislike	-0.53	0.39	0.31	-
8	Self-accusation	-0.46	0.20	0.20	0.41
9	Suicidal wishes	0.03	-0.03	-	-
10	Crying	-0.37	0.12	0.20	0.35
11	Irritability	-0.45	0.39	0.32	-
12	Social withdrawal	-0.53	0.38	0.28	-
13	Indecisiveness	-0.22	0.22	0.05	-
14	Distortion of body image	-0.17	0.17	-	-
15	Work inhibition	-0.37	0.14	0.13	0.33
16	Sleep disturbance	-0.24	0.21	0.09	-
17	Fatigability	-0.49	0.48	0.09	-
18	Loss of appetite	-0.44	0.21	0.15	0.33
19	Weight loss	-0.41	0.32	0.09	0.32
20	Somatic preoccupation	-0.44	0.19	0.24	0.29
21	Loss of libido	-0.49	0.17	0.13	0.35

Table 3: Item point-biserial correlations for the BDI-II.

can only be three possible threshold values: 1, 2 and 3. A threshold of 1 refers to the point at which response levels 1, 2, and 3 together become more likely than level 0; a threshold of 2 is the point at which levels 2, 3 and 4 together become more likely than levels 0 and 1 and so on. From the Wright map we can see that there is generally a good spread of the items over the construct "feelings of depression". The items cover the lower, middle and higher end of the construct. However, from the map there are some gaps in the construct that are not well covered by the items. Specifically, we notice a lack of enough items that cover severe depression. There is a wide gap between item thresholds 14.1 and 11.2. Likewise there is need for more items between thresholds 14.1 and 4.3. On the lower end of the scale there is a gap between 3.1 and 11.1. The consequence of these gaps is that the instrument may not adequately discriminate among respondents up at the high end. However, given that the concern is not the exact level of depression but whether one is depressed or not given a cut off point, this may not be so much of a problem and the current items may suffice.

The locations of the bars of the histogram (i.e. the Xs) are the estimated locations of the respondents on the variable. The respondents tend to cluster in the lower half of the scale (below logit score 0) reflecting the fact that most patients had minimal to moderate depression (Table 3). There were few respondents at the extremes. The mean depression score was 8.86 with a standard deviation of 5.44.

Evidence for Reliability

Standard Error of Measurement (SEM): When using an instrument on an individual respondent, the sem is the most important tool for assessing the usefulness of that estimate of location (Figure 3). If the sem is too large, the measurer will not be able to make intelligible interpretation of the results. For this sample, the mean of the sem was 0.569 with a standard deviation of 0.407. This gives a 95% confidence interval width of 1.596 logits. This is about 23% of the width of the entire Wright map from maximum to minimum locations. This implies that the instrument is not accurate for individual usage. To see this, let

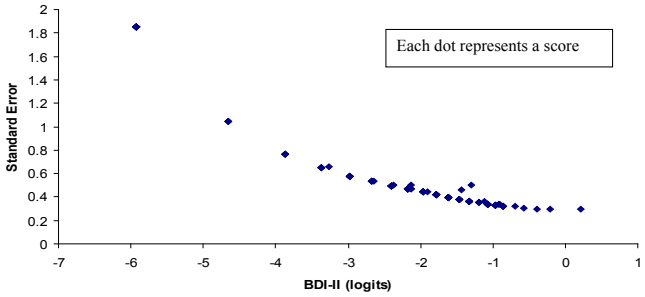

Figure 3: The standard error of measurement for the BDI-II instrument.

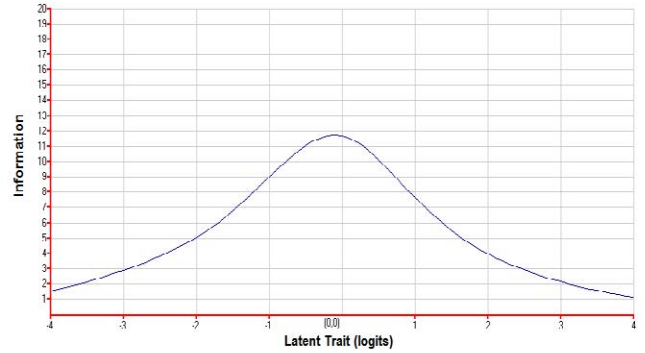

Figure 4: Test information curve.

us consider the second threshold. The confidence interval spans (for the second threshold) the range from items 13 and 16 to items 4 and 20 covering a wide range of depression states indeed. Thus, this instrument is probably not very useful for accurate clinical diagnosis, but it may well be useful for initial screening or as a basis for group measures.

Test information: The uncertainty in an estimate of person location is usually characterized for individuals using the standard error of measurement. Derived from the standard error of measurement, the information curve for the BDI-II instrument is given in Figure 4. Together with Figure 3, these graphs show that the most sensitive part of the instrument is from approximately -1.5 to +1.5 logits. Comparing with the Wright map, this range contains thirty six out of the forty seven items thresholds (77%) observed in these data. Thus, the instrument's range of maximum sensitivity makes general sense with respect to the items.

Internal consistency coefficients: The Cronbach's alpha was 0.77 but when only complete cases were used the Cronbach's alpha coefficient was 0.79. The expected a posteriori (EAP) reliability was 0.86.

Evidence for validity: The structure of the discussion about validity below is based on the 1999 Standards for Educational and Psychological Testing [68].

Evidence based on instrument content: While the author of the original BDI [59] did not explicitly apply IRM procedures in developing this instrument, these procedures [IRM] were employed implicitly. IRM is a new approach to instrument development that was not in existence then. The first step in developing an instrument using IRM is to develop a construct map. A construct map is a content structure in which the relative behaviors of a generic respondent having various amounts of the construct are arranged in order on one side, and potential items are arranged in order of endorsability on the other. Beck et al designed the instrument with the intent of building in strong content. The content was based on systematic observations and records of the characteristic attitudes and symptoms of depressed patients [59]. He selected a group of these attitudes and symptoms that appeared to be specific for these depressed patients and which were consistent with the descriptions of depression contained in the psychiatric literature. On the basis of this procedure, he constructed an inventory composed of 21 categories of symptoms and attitudes. He arranged responses in each category using a Guttman-like scale that endeavors to cover the entire spectrum of depressive symptoms. This intention of structure is illustrated in the Wright map (Figure 2) that fairly covers the different levels of depressive symptoms. Based on these observations, we say that there is evidence of validity based on instrument content.

Evidence for validity based on response processes: Evidence based on the response process consists of studies of how respondents react to the items and the instrument as a whole. These might consist of 'think alouds', 'exit interviews' or 'cognitive interviews' with samples of respondents. Such evidence was not available with our secondary analysis of the data, so this aspect of the framework is not relevant at this time.

Evidence for validity based on internal structure: One major criterion for internal or construct validity is an a priori theoretically based hypothesis about the order of item endorsability, the ease with which respondents rate items strongly. The Wright map is a very good tool for investigating this [58]. No such expectations were made for the BDI. The authors clearly stated that "the items were chosen on the basis of their relationship to the overt behavioral manifestations of depression and do not reflect any theory regarding the etiology or the underlying psychological processes in depression' [59]. This is also illustrated in the Wright map (Figure 2) where there is no discernible empirical order to the items of the BDI. Instead, the feature of the BDI-II items that maps out the BDI variable is the transitions between the response categories. One thing we would expect is a similar pattern of item locations at each threshold level but this is not the case. Hence this source of validity evidence is not available.

An important piece of evidence that an item is functioning as expected is that the increasing response levels of the item are operating consistently with the instrument as a whole. With the item response modeling approach, one way that this consistency can be made manifest is that respondents higher on the construct would, in general, also score higher on each item. In terms of the Wright map, one would consider the locations of the respondents within each score group for an item: if the mean location for each group tends to increase as the scores increase, it seems reasonable to say that this particular expectation that comes from the items design has been fulfilled. This analysis showed that for most items, the mean increases as the score rises (Table 4).

A standard requirement of the items design is that, across important subgroups, items function in a similar way for respondents who are at the same location--- that is they should exhibit no evidence of differential item functioning (DIF). DIF is, therefore, said to be present when respondents at the same location on the variable give different responses across different subgroups. For this analysis, we investigated the effect of CD4 cell count on item functioning. Respondents were divided into two subgroups: those with CD4 cell count below 250 cells/uL and those with CD4 cell count above 250 cells/uL. A cut off of 250 was used because this is the level around which an HIV-infected patient in this population is deemed to have AIDS and thus started on antiretroviral therapy. Overall, respondents with CD4 cell count levels below 250 were 0.5 logits higher on the depression trait. This does not show DIF but evidence of differential impact. This result is not

Item #	Description of the item	Response Categories			
		0	1	2	3
1	Mood	-2.17	-1.39	-1.43	-
2	Pessimism	-2.28	-1.28		-
3	Sense of failure	-2.29	-1.44	-1.52	-1.43
4	Lack of satisfaction	-2.91	-1.73	-1.44	-0.41
5	Guilt feelings	-2.20	-1.23	-0.93	-
6	Sense of punishment	-2.19	-0.85	-1.23	-
7	Self-dislike	-2.27	-1.27	-1.13	-
8	Self-accusation	-2.20	-1.49	-1.19	-0.67
9	Suicidal wishes	-1.99	-2.09	-	-
10	Crying	-2.05	-1.45	-0.86	-0.40
11	Irritability	-2.44	-1.59	0.00	-
12	Social withdrawal	-2.63	-1.58	-1.16	-
13	Indecisiveness	-2.06	-1.34	-1.56	-
14	Distortion of body image	-2.01	-1.38	-	-
15	Work inhibition	-2.57	-1.83	-1.58	-0.74
16	Sleep disturbance	-2.16	-1.66	-1.52	-
17	Fatigability	-2.22	-1.22	-1.08	-
18	Loss of appetite	-2.60	-1.77	-1.54	-0.74
19	Weight loss	-2.20	-1.38	-1.42	0.00
20	Somatic preoccupation	-2.53	-1.72	-1.45	-0.55

Table 4: Mean locations of respondents within each score group.

Item #	Item description	Logit difference	DIF (Y/N)
5	Guilty feelings	-0.670	Y
11	Agitation	-1.742	Y
13	Indecisiveness	-0.738	Y
14	Worthlessness	+1.024	Y
15	Loss of energy	+0.790	Y
17	Irritability	+0.672	Y

Table 5: Testing for DIF.

surprising, as being diagnosed with HIV would usually be associated with negative feelings. Test results for items with DIF are shown in Table 5. In terms of DIF, a recommended standard of effect size is as follows: a logit difference value less than 0.426 is "negligible", a value between 0.426 and 0.638 is "intermediate" and a value over 0.638 is "large" [58]. Overall, there was statistically significant DIF for the item set (Chi sq=52.02, df=20, p-value <0.0001). Applying these criteria to the statistically significant logit differences in Table 5 shows that items 5, 11, 13, 14, 15, and 17 had large DIF. The rest of the items had no DIF.

Evidence based on relations to other variables: Where there are other external variables that the construct should (according to theory) predict, a strong relationship between the instrument under scrutiny and these external variables can be used as validity evidence. In another paper from the same study (Byakika-Tusiime et al., 2009), we demonstrated that depression was significantly associated with adherence to antiretroviral therapy [OR=0.32, 95% CI 0.11-0.93].

Evidence based on consequences of using an instrument: The final form of validity evidence relates to consequences. We did not have any data on consequences of using this instrument hence our analysis did not evaluate this form of evidence.

Discussion

There have been little data regarding the validity of the Beck Depression Inventory in an HIV-infected population despite the fact that the BDI is commonly used to assess symptoms of depression in this population.

The findings of this analysis showed that the BDI-II has good psychometric properties when used in a low income HIV-infected population in Uganda. These findings corroborate those found in other populations [69,70]. However to our knowledge, this is the first attempt to validate the BDI-II in an HIV-infected population in sub-Saharan Africa. It is also the first attempt to use item response modeling to calibrate this instrument. Previous calibrations were based on classical test theory and now IRM further confirms the robustness of the BDI-II instrument.

The weighted mean square indicates that most of the items are fitting within reasonable bounds. Thus, the overall finding is that the BDI-II data fit the Partial Credit Model reasonably well. The Wright map generally revealed a relatively good spread of the items over the construct "feelings of depression" although there were a few gaps identified in the instrument that may need bridging with more items or item responses. The biggest gap was observed at the top end of the instrument—that measures severe depression. The need to add more items or item responses for adequate discrimination of respondents at these levels would largely depend on the ultimate purpose of this instrument. If the purpose is to detect respondents with severe depression then it should probably be augmented with more items/item responses at the upper end. On the other hand if the purpose is to screen for minor to moderate or moderate to severe depression, then the current items seem to suffice.

The majority of respondents in this sample had mild to moderate depression and the distribution of the respondents shows that many respondents in this sample were below -1.5 logits. However, the test information curve tells us that the most sensitive part of this instrument is the region of moderate to severe depression (-1.5 to +1.5 logits). This implies that the instrument is not functioning optimally for quite a large proportion of this sample. However, since the ultimate purpose of this instrument is to detect levels of depression that would require intervention (moderate to severe depression) then this is not a big problem. That is, the instrument captures respondents with moderate to severe depression, which is the main concern.

The findings of this analysis provide evidence for the validity of the instrument's usage. Build of this evidence has been based on test content, internal structure, and relations to other variables. This is an alternative to the traditional way of assessing validity; that is, criterion, content and construct validity [68]. We have shown evidence based on instrument content, evidence based on internal structure and evidence based on relations to other variables.

We found significant differential item functioning with CD4 cell count subgroups. DIF is an undesirable characteristic of any test. Items 5, 11, 13, 14, 15, and 17 had large DIF. This finding should be taken with caution. There is need to establish that the DIF is indeed not a result of random fluctuations and would thus require doing repeated samplings to confirm this. If indeed DIF is established, the best strategy is to develop alternative items that do not exhibit DIF. Another less popular alternative is to use two different calibrations for the two groups. However together looking at all the items exhibiting DIF, those exhibiting positive DIF cancel out those with negative DIF. Hence overall we can say that there is no significant DIF effect.

Internal consistency coefficients in this sample were good. We found a Cronbach's alpha of 0.79 and an EAP reliability coefficient of 0.86. This demonstrates low measurement error. The reliability

coefficients obtained from this analysis are similar to what has been found elsewhere [69]. Within nonpsychiatric populations coefficient alphas of the range from 0.73 to 0.92 have been observed.

There were a few limitations to this analysis. First, the sample size was rather small. We would require a bigger sample size to confirm these findings. Secondly, the study sample was mainly female (75%). This may limit generalizability of our findings.

In conclusion, this analysis demonstrated good psychometric properties of the BDI-II when used to screen for depression in a low income predominantly female HIV-infected population. The data fitted the partial credit model reasonably well suggesting that the currently used scoring method of the items is reasonable. These findings therefore, support the use of the BDI-II in assessing depressive symptoms for HIV-infected patients in sub-Saharan Africa especially women. The quality of item calibration must be supervised continuously. We cannot expect the items in an instrument to retain their calibrations indefinitely or to work equally well for every person with whom they may be used.

Acknowledgements

We acknowledge all members of the research team of the parent study who recruited the study participants, collected and managed the data: Mary Kasango, Annet Kawuma, Muzamir Baryamushanga, Grace Manyangwa and Irene Zawedde. We also wish to thank the project administrator Sarah Nakandi and the driver Ibrahim Kiviri. Last but not least we thank all the patients in the MTCT+ program who participated in the parent study.

Funding

Funding was provided by the Bill and Melinda Gates Grant; the R0-1 MH54907, R21 AA015897, K24 AA015287 grants; and by the Fogarty AIDS International Training and Research Program/University California Berkeley (1D43 TW00003).

References

1. Benton TD (2008) Depression and HIV/AIDS. Curr Psychiatry Rep 10: 280-285.
2. Bing EG, Burnam MA, Longshore D, Fleishman JA, Sherbourne CD, et al. (2001) Psychiatric disorders and drug use among human immunodeficiency virus-infected adults in the United States. Arch Gen Psychiatry 58: 721-728.
3. Ciesla JA, Roberts JE (2001) Meta-analysis of the relationship between HIV infection and risk for depressive disorders. Am J Psychiatry 158: 725-730. Dew MA,
4. Becker JT, Sanchez J, Caldararo R, Lopez OL, et al. (1997) Prevalence and predictors of depressive, anxiety and substance use disorders in HIV-infected and uninfected men: a longitudinal evaluation. Psychol Med 27: 395-409.
5. Monahan PO, Shacham E, Reece M, Kroenke K, Ong'or WO, et al. (2009) Validity/reliability of PHQ-9 and PHQ-2 depression scales among adults living with HIV/AIDS in western Kenya. J Gen Intern Med 24: 189-197.
6. Morrison MF, Petitto JM, Ten Have T, Gettes DR, Chiappini MS, et al. (2002) Depressive and anxiety disorders in women with HIV infection. Am J Psychiatry 159: 789-796.
7. Pretorius C, Glaziou P, Dodd PJ, White R, Houben R (2014) Using the TIME model in Spectrum to estimate tuberculosis-HIV incidence and mortality. AIDS 28 Suppl 4: S477-487.
8. Kaharuza FM, Bunnell R, Moss S, Purcell DW, Bikaako-Kajura W, et al. (2006) Depression and CD4 cell count among persons with HIV infection in Uganda. AIDS Behav 10: S105-111.
9. Keogh P, Allen S, Almedal C, Temahagili B (1994) The social impact of HIV infection on women in Kigali, Rwanda: a prospective study. Soc Sci Med 38: 1047-1053.
10. Kiima DM, Njenga FG, Okonji MM, Kigamwa PA (2004) Kenya mental health country profile. Int Rev Psychiatry 16: 48-53.
11. Myer L, Smit J, Roux LL, Parker S, Stein DJ et al. (2008) Common mental disorders among HIV-infected individuals in South Africa: prevalence, predictors, and validation of brief psychiatric rating scales. AIDS Patient Care STDS 22: 147-158.
12. Olley BO, Gxamza F, Seedat S, Theron H, Taljaard J, et al. (2003) Psychopathology and coping in recently diagnosed HIV/AIDS patients--the role of gender. S Afr Med J 93: 928-931.
13. Olley BO, Seedat S, Nei DG, Stein DJ (2004) Predictors of major depression in recently diagnosed patients with HIV/AIDS in South Africa. AIDS Patient Care STDS 18: 481-487.
14. Justice AC, McGinnis KA, Atkinson JH, Heaton RK, Young C, et al. (2004) Psychiatric and neurocognitive disorders among HIV-positive and negative veterans in care: Veterans Aging Cohort Five-Site Study. AIDS 18 Suppl 1: S49-59.
15. Perry S, Fishman B (1993) Depression and HIV. How does one affect the other? JAMA 270: 2609-2610.
16. Stoskopf CH, Kim YK, Glover SH (2001) Dual diagnosis: HIV and mental illness, a population-based study. Community Ment Health J 37: 469-479.
17. Ickovics JR, Hamburger ME, Vlahov D, Schoenbaum EE, Schuman P et al. (2001) Mortality, CD4 cell count decline, and depressive symptoms among HIV-seropositive women: longitudinal analysis from the HIV Epidemiology Research Study. Jama 285: 1466-1474.
18. Antelman G, Kaaya S, Wei R, Mbwambo J, Msamanga GI, et al. (2007) Depressive symptoms increase risk of HIV disease progression and mortality among women in Tanzania. J Acquir Immune Defic Syndr 44: 470-477.
19. Gureje O, Kola L, Afolabi E (2007) Epidemiology of major depressive disorder in elderly Nigerians in the Ibadan Study of Ageing: a community-based survey. Lancet 370: 957-964.
20. Mogga S, Prince M, Alem A, Kebede D, Stewart R, et al. (2006) Outcome of major depression in Ethiopia: population-based study. Br J Psychiatry 189: 241-246.
21. Wild LG, Flisher AJ, Lombard C (2004) Suicidal ideation and attempts in adolescents: associations with depression and six domains of self-esteem. J Adolesc 27: 611-624.
22. Bolton P, Neugebauer R, Ndogoni L (2002) Prevalence of depression in rural Rwanda based on symptom and functional criteria. J Nerv Ment Dis 190: 631-637.
23. Tomlinson M, Swartz L, Kruger LM, Gureje O (2007) Manifestations of affective disturbance in sub-Saharan Africa: key themes. J Affect Disord 102: 191-198.
24. Jelsma J, Mielke J, Powell G, De Weerdt W, De Cock P (2002) Disability in an urban black community in Zimbabwe. Disabil Rehabil 24: 851-859.
25. Hughes J, Jelsma J, Maclean E, Darder M, Tinise X (2004). The health-related quality of life of people living with HIV/AIDS. Disabil Rehabil, 26: 371-376.
26. Jelsma J, Maart S, Eide A, Ka'Toni M, Loeb M (2007) The determinants of health-related quality of life in urban and rural isi-Xhosa-speaking people with disabilities. Int J Rehabil Res 30: 119-126.
27. Kaaya SF, Fawzi MC, Mbwambo JK, Lee B, Msamanga GI, et al. (2002) Validity of the Hopkins Symptom Checklist-25 amongst HIV-positive pregnant women in Tanzania. Acta Psychiatr Scand 106: 9-19.
28. Adewuya AO, Ola BA, Afolabi OO (2006) Validity of the patient health questionnaire (PHQ-9) as a screening tool for depression amongst Nigerian university students. J Affect Disord 96: 89-93.
29. Omoro SA, Fann JR, Weymuller EA, Macharia IM, Yueh B (2006) Swahili translation and validation of the Patient Health Questionnaire-9 depression scale in the Kenyan head and neck cancer patient population. Int J Psychiatry Med 36: 367-381.
30. Rosengren A, Hawken S, Ounpuu S, Sliwa K, Zubaid M, et al. (2004) Association of psychosocial risk factors with risk of acute myocardial infarction in 11119 cases and 13648 controls from 52 countries (the INTERHEART study): case-control study. Lancet 364: 953-962.
31. Smit J, Myer L, Middelkoop K, Seedat S, Wood R, et al. (2006) Mental health and sexual risk behaviours in a South African township: a community-based cross-sectional study. Public Health 120: 534-542.
32. Byakika-Tusiime J, Crane J, Oyugi JH, Ragland K., Kawuma A et al. (2009) Longitudinal Antiretroviral Adherence in HIV+ Ugandan Parents and Their Children Initiating HAART in the MTCT-Plus Family Treatment Model: Role of Depression in Declining Adherence Over Time. AIDS Behav.
33. Starace F, Ammassari A, Trotta MP, Murri R, De Longis P, et al. (2002) Depression

is a risk factor for suboptimal adherence to highly active antiretroviral therapy. J Acquir Immune Defic Syndr 31 Suppl 3: S136-139.

34. Bass JK, Bolton PA, Murray LK (2007) Do not forget culture when studying mental health. Lancet 370: 918-919.

35. Abiodun OA, Bola AO, Olutayo OA (2007) Prevalence of major depressive disorders and a validation of the beck depression inventory among nigerian adolescents. Eur Child Adolesc Psychiatry 16: 287-292

36. Uwakwe R, Okonkwo JE (2003) Affective (depressive) morbidity in puerperal Nigerian women: validation of the Edinburgh Postnatal Depression Scale. Acta Psychiatr Scand 107: 251-259.

37. Adewuya AO, Ola BA, Dada AO, Fasoto OO (2006) Validation of the Edinburgh Postnatal Depression Scale as a screening tool for depression in late pregnancy among Nigerian women. J Psychosom Obstet Gynaecol 27: 267-272.

38. Bolton P, Wilk CM, Ndogoni L (2004) Assessment of depression prevalence in rural Uganda using symptom and function criteria. Soc Psychiatry Psychiatr Epidemiol 39: 442-447.

39. Hamad R, Fernald LC, Karlan DS, Zinman J (2008) Social and economic correlates of depressive symptoms and perceived stress in South African adults. J Epidemiol Community Health 62: 538-544.

40. Kagee A (2008) Symptoms of depression and anxiety among a sample of South African patients living with a chronic illness. J Health Psychol 13: 547-555.

41. Rahim SI, Cederblad M (1989) Epidemiology of mental disorders in young adults of a newly urbanized area in Khartoum, Sudan. Br J Psychiatry 155: 44-47.

42. Rochat TJ, Richter LM, Doll HA, Buthelezi NP, Tomkins A, et al. (2006) Depression among pregnant rural South African women undergoing HIV testing. JAMA 295: 1376-1378.

43. Verdeli H, Clougherty K, Bolton P, Speelman L, Lincoln N, et al. (2003) Adapting group interpersonal psychotherapy for a developing country: experience in rural Uganda. World Psychiatry 2: 114-120.

44. Wilk CM, Bolton P (2002) Local perceptions of the mental health effects of the Uganda acquired immunodeficiency syndrome epidemic. J Nerv Ment Dis 190: 394-397.

45. Oyugi JH, Byakika-Tusiime J, Ragland K, Laeyendecker O, Mugerwa R, Kityo C, et al. (2007) Treatment interruptions predict resistance in HIV-positive individuals purchasing fixed-dose combination antiretroviral therapy in Kampala, Uganda. Aids, 21: 965-971.

46. Dew MA, Becker JT, Sanchez J, Caldararo R, Lopez OL et al. (1997) Prevalence and predictors of depressive, anxiety and substance use disorders in HIV-infected and uninfected men: a longitudinal evaluation. Psychol Med, 27: 395-409.

47. Patel V, Abas M, Broadhead J, Todd C, Reeler A (2001) Depression in developing countries: lessons from Zimbabwe. BMJ 322: 482-484.

48. Kleinman A (1988) Rethinking psychiatry: from cultural category to personal experience. New York City: Free Press.

49. Simon GE, Goldberg DP, Von Korff M, Ustün TB (2002) Understanding cross-national differences in depression prevalence. Psychol Med 32: 585-594.

50. Dohrenwend BP (1990) 'The problem of validity in field studies of psychological disorders' revisited. Psychol Med 20: 195-208.

51. Prince RH (1997) What's in a name? Transcultural Psychiatry 3: 151-154.

52. Van Ommeren M (1999) Preparing instruments for transcultural research: use of theTranslation Monitoring Form with Nepali-speaking Bhutanese refugees. Transcultural Psychiatry 36: 285-301.

53. Piotrowski C, Sherry D, Keller JW (1985) Psychodiagnostic test usage: a survey of the society for personality assessment. J Pers Assess 49: 115-119.

54. Kalichman SC, Sikkema KJ, Somlai A (1995) Assessing persons with human immunodeficiency virus (HIV) infection using the Beck Depression Inventory: disease processes and other potential confounds. J Pers Assess 64: 86-100.

55. Lustman PJ, Clouse RE, Griffith LS, Carney RM, Freedland KE (1997) Screening for depression in diabetes using the Beck Depression Inventory. Psychosom Med 59: 24-31.

56. Deribe K, Woldemichael K, Wondafrash M, Haile A, Amberbir A (2008) Disclosure experience and associated factors among HIV positive men and women clinical service users in Southwest Ethiopia. BMC Public Health 8: 81.

57. Moosa M, Jeenah F, Vorster M (2005) HIV in South Africa: Depression and CD4 count. South African Journal of Psychiatry 11: 12-15.

58. Wilson M (2005) Constructing Measures: An Item Response Modelling Approach. Mahwah: Lawrence Erlbaum Associates, Inc.

59. Beck AT, Ward CH, Mendelson M, Mock J, Erbaugh J (1961) An inventory for measuring depression. Arch Gen Psychiatry 561-571.

60. Nunnally JC, Berstein IH (1994) Psychometric Theory. New York: McGraw-Hill.

61. Wu ML, Adams RJ, Wilson MR, Haldane SA (2007) ACER ConQuest: Generalised Item Response Modelling Software (Version 2.0). Camberwell: ACER Press:Australian Council for Educational Research Ltd.

62. Shea RL, Norcini JJ, Webster GD (1988) An application of item response theory to certifying examinations in internal medicine. Evaluation and the Health Professions 1: 283-305.

63. Bond TG, Fox CM (2001) Applying the Rasch Model: Fundamental measurement in the human sciences. NJ: Lawrence Erlbaum Associates, Mahwah.

64. Embretson SE, Reise SP (2000) Item resposne theory for psychologists. NJ: Lawrence Erlbaum Associates, Mahwah.

65. Kennedy C, Wilson M, Draney K., Tutunciyan S, Vorp R (2008) ConstructMap. Berkeley: Berkeley Evaluation & Assessment Research (BEAR) Center, USA.

66. Andrich DA (1978) A rating formulation for ordered response categories. Psychometrika, 561-573.

67. Masters GN (1982) A Rasch model for partial credit scoring. Psychometrika 47: 149-174.

68. American Educational Research Association, American Psychological Association, & Education. NCFMI (1999). Standards for educational and psychological testing. Washington, DC: American Educational Research Association.

69. Beck A, Steer R, Garbin M (1988) Psychometric properties of the Beck Depression Inventory: Twenty five years of evaluation. Clinical Psychology Review 8: 77-100.

70. Steele GI (2003) The development and validation of Xhosa translations of Beck Depression Inventory, the Beck Anxiety Inventory and the Beck Hopelessness Scale. Rhodes University, Grahamstown.