NIH Public Access
Author Manuscript
AIDS Care. Author manuscript; available in PMC 2012 August 02.

Published in final edited form as:
AIDS Care. 2010 July ; 22(7): 874–885. doi:10.1080/09540120903483034.

Measuring depression levels in HIV-infected patients as part of routine clinical care using the 9-item patient health questionnaire (PHQ-9)

PK Crane[1,*], **LE Gibbons**[1], **JH Willig**[2], **MJ Mugavero**[2], **ST Lawrence**[2], **JE Schumacher**[2], **MS Saag**[2], **MM Kitahata**[1], and **HM Crane**[1]

[1]University of Washington

[2]University of Alabama Birmingham

Abstract

Little is known about the psychometric properties of depression instruments among persons infected with HIV. We analyzed data from a large sample of patients in usual care in two US cities (n=1467) using the 9-item Patient Health Questionnaire from the PRIME-MD (the PHQ-9). The PHQ-9 had curvilinear scaling properties and varying levels of measurement precision along the continuum of depression measured by the instrument. In our cohort, the scale showed a prominent floor effect and a distribution of scores across depression severity levels. Three items had differential item functioning (DIF) with respect to race (African-American vs. white); two had DIF with respect to sex, and one had DIF with respect to age. There was minimal individual-level DIF impact. Twenty percent of the difference in mean depression levels between African-Americans and whites was due to DIF. While standard scores for the PHQ-9 may be appropriate for use with individual HIV-infected patients in cross-sectional settings, these results suggest that investigations of depression across groups and within patients across time may require a more sophisticated analytic framework.

Background

Depression is the most prevalent mental disorder among HIV-infected individuals (Starace et al., 2002) with rates two to four times higher than those found in general populations (Ciesla & Roberts, 2001; Morrison et al., 2002; Starace et al., 2002). Depression and depressive symptoms are key predictors of poor adherence to HIV medications (Arnsten et al., 2002; Boarts, Sledjeski, Bogart, & Delahanty, 2006; Catz, Kelly, Bogart, Benotsch, & McAuliffe, 2000; DiMatteo, Lepper, & Croghan, 2000; Gordillo, del Amo, Soriano, & Gonzalez-Lahoz, 1999; Paterson et al., 2000; Singh et al., 1996; Starace et al., 2002; Waldrop-Valverde & Valverde, 2005) and negatively impact clinical outcomes (Bouhnik et al., 2005; Burack et al., 1993; Ickovics et al., 2001). It is thus important to measure and address depression and depressive symptoms in HIV-infected patients. The use of standardized instruments facilitates screening for depressive symptoms in busy clinical care settings (Staab et al., 2001). We sought to address measurement properties of a brief, commonly-used depression instrument, the 9-item Patient Health Questionnaire from the PRIME-MD (the PHQ-9) (Spitzer, Kroenke, & Williams, 1999), using data from HIV-infected patients in routine clinical care.

[*]Corresponding author: Paul K. Crane, MD MPH, General Internal Medicine, Harborview Medical Center, Box 359780, 325 Ninth Avenue, Seattle, WA 98104. (206) 744-1831 (phone). (206) 744-9917 (fax). pcrane@u.washington.edu.

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Current data on the psychometric properties of this instrument are somewhat limited. Reliability is an important psychometric property relating to the reproducibility and measurement precision of the test. While several studies have addressed overall average reliability of the PHQ-9 using Cronbach's alpha (Adewuya, Ola, & Afolabi, 2006; Cameron, Crawford, Lawton, & Reid, 2008; Dum, Pickren, Sobell, & Sobell, 2008; Hepner, Hunter, Edelen, Zhou, & Watkins, 2009; Hides et al., 2007; Lee, Schulberg, Raue, & Kroenke, 2007; Lotrakul, Sumrithe, & Saipanish, 2008; Omoro, Fann, Weymuller, Macharia, & Yueh, 2006; Stafford, Berk, & Jackson, 2007; Yeung et al., 2008), only one study among Kenyans (Monahan et al., 2009) has specifically included patients infected with HIV. Furthermore, using a single omnibus statistic such as Cronbach's alpha to describe reliability may be problematic if measurement precision varies across the scale.

Validity refers to the property that the test measures what it reports to measure, and is a second essential psychometric characteristic. Several studies have evaluated validity in a number of settings (Donnelly & Kim, 2008; Gilbody, Richards, Brealey, & Hewitt, 2007; Kroenke, Spitzer, & Williams, 2001; Lamers et al., 2008; Lee et al., 2007; Rogers, Adler, Bungay, & Wilson, 2005; Stafford et al., 2007).

A test may be reliable and valid but not be tolerable to patients; acceptability is a third characteristic of tests that should be investigated. One study addressed acceptability in routine use in psychiatric practices (Duffy et al., 2008).

A few recent studies have employed modern psychometric approaches to further characterize the PHQ-9 (Lamoureux et al., 2009; Williams et al., 2009). Differences in observed scores across groups do not necessarily indicate bias, since mean levels of depression may vary across groups. However, when controlling for the underlying level of depression, group membership should be unrelated to specific item responses; the item should work the same way across groups. When that is not the case, the item is said to have differential item functioning (DIF) (Camilli & Shepard, 1994). DIF is a critical area of inquiry when scales are used with heterogeneous populations. Two studies have addressed whether the PHQ-9 may have item-level bias (differential item functioning, or DIF) related to race/ethnicity (Hepner, Morales, Hays, Edelen, & Miranda, 2008; Huang, Chung, Kroenke, Delucchi, & Spitzer, 2006), but no other covariates were evaluated. One small study addressed DIF in older adults with vision problems, but its sample size was too small to make firm conclusions (Lamoureux et al., 2009). There have been no evaluations of the PHQ-9 for DIF among patients infected with HIV.

To address these gaps in our knowledge about the PHQ-9, we sought to assess psychometric properties of this instrument among individuals infected with HIV who responded to this instrument as part of routine clinical care. We also sought to assess the PHQ-9 for DIF related to several covariates, and to determine whether all or a portion of observed differences in depression scores across groups may have been due to DIF.

Methods

Study setting

This cross-sectional study was conducted among a convenience sample of patients from the University of Alabama at Birmingham (UAB) 1917 HIV/AIDS Clinical Clinic and from the University of Washington (UW) Harborview Medical Center Madison HIV Clinic. UAB and UW are sites of the Centers for AIDS Research Network of Integrated Clinical Systems (CNICS), an initiative that captures a broad range of information on disease management through collection of point-of-care data across multiple academic cohorts engaged in the

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

longitudinal care of HIV/AIDS (Kitahata et al., 2008). The Institutional Review Board reviewed and approved this study protocol at both sites.

Data sources

HIV-infected patients 18 years of age or older who attended clinic for a routine appointment between September 2005 and April 2009 were eligible for the study. Patients complete assessments roughly every 6 months; however, only the first assessment for each patient is considered here. Patients used tablet PCs (UW) and wall-mounted personal computers with touch screens (UAB) to complete an assessment including measures of depression (PHQ-9 from the PRIME-MD (Kroenke et al., 2001; Spitzer et al., 1999)) and drug use (Alcohol, Smoking, and Substance Involvement Screening Test (Newcombe, Humeniuk, & Ali, 2005; 2002)). As previously described (H. Crane et al., 2008a; H. Crane et al., 2007), we used web-based survey software developed specifically for patient-based measures to facilitate capture of patient-based assessments in real-time during routine care visits. Data were also obtained from the CNICS data repository which integrates comprehensive clinical data on the CNICS cohort from all outpatient and inpatient encounters including demographic, clinical, and socioeconomic information (Kitahata et al., 2008).

The PHQ-9

The items of this scale correspond to the features of depression enumerated in the Diagnostic and Statistical Manual of Mental Disorders (American Psychiatric Association Task Force on DSM-IV, 1994). The nine depression symptom items share a common stem: “Over the past 2 weeks, how often have you been bothered by any of the following problems?” The items share a common set of response options: 0, “Not at all”; 1, “Several days”; 2, “More than half the days”; and 3, “Nearly every day.” The content of the PHQ-9 items is shown in the left column of Table 2.

Psychometric analyses. Dimensionality

We addressed dimensionality using a combination of exploratory and confirmatory factor analyses (EFA and CFA) (Lai, Crane, & Cella, 2006; Reeve et al., 2007). We evaluated the ratio between Eigen values for the first and second factor in EFA. We evaluated model fit and item loadings with a single factor CFA. We used model-based modification indices and clinical intuition to guide us regarding residual correlations to include in a modified single factor CFA. We evaluated model fit for this modified single factor model and compared the magnitude of the loadings for the two CFA models.

Psychometric analyses. Test characteristic curve (TCC) and test information curve (TIC)

We used Parscale (Muraki & Bock, 2003) for item response theory (IRT) analyses. IRT assumes a single underlying (“latent”) trait is responsible for the observed correlations across responses to the items of the PHQ-9. In this paper we refer to that trait as the “level of depression”. We used the graded response IRT model (Samejima, 1969, 1997) for the ordinal response data of the PHQ-9.

We plotted the TCC, which shows the relationship of the standard total score on the Y-axis to the level of depression measured by the test on the X-axis. A linear TCC implies equal scaling across all total PHQ-9 scores. We plotted the TIC and the standard error of measurement (Mungas, Reed, Marshall, & Gonzalez, 2000). The TIC shows the amount of measurement precision at each depression level. This is somewhat analogous to Cronbach’s alpha, though the TIC shows the level of reliability at all levels of depression rather than summarizing this with a single omnibus statistic, as done with Cronbach’s alpha. To

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

enhance clinical comprehensibility, we transformed all IRT scores to have a mean of 100 and a standard deviation of 15.

Psychometric analyses. Differential item functioning (DIF)

We used a hybrid ordinal logistic regression-IRT package called difwithpar for DIF analyses (P. Crane, Gibbons, Jolley, & van Belle, 2006). Based on preliminary analyses we found a p-value of 0.03 was sensitive enough to detect DIF for several covariates with sufficient remaining anchor items, so we report results from analyses using this threshold for both uniform and non-uniform DIF. We addressed DIF related to the following covariates: sex, race (African-American vs. white; others were excluded from that analysis), age (<40, 40–49, >49 years), $CD4^+$ nadir (<200, 200–299, ≥300 cells/mm^3), recent substance use (within 3 months), and HIV transmission risk factor (men who have sex with men [MSM] vs. other). We identified DIF with respect to the unadjusted IRT depression score for each of these covariates. We also identified DIF with respect to all of these covariates at the same time. We present findings of item-level DIF presence with respect to each covariate and the cumulative DIF impact on individuals and groups.

Results

Item responses for the PHQ-9 were available for 1,467 participants. Complete item and covariate data were available for 1,452 participants (99%). The PHQ-9 was well accepted. The median completion time was 58 seconds (interquartile range 33–83 seconds; 95^{th} percentile 128 seconds, 99^{th} percentile 177 seconds, maximum time 199 seconds). Only two individuals complained about the assessment; both endorsed suicidality and objected to the clinical follow-up that resulted from this endorsement. There were no other complaints related to administration of the PHQ-9. Demographic characteristics of study participants stratified by race/ethnicity are summarized in Table 1.

Eigen values for the first two factors in EFA were 5.35 and 0.37, with a ratio of 14. A single-factor CFA model did not fit well; the confirmatory fit index (CFI) was 0.92 (threshold >0.95), the Tucker-Lewis Index (TLI) was 0.89 (threshold >0.95), and the root mean squared error of approximation (RMSEA) was 0.14 (<0.08 for adequate fit, <0.05 for good fit). We used modification indices to inform the specification of empirically-guided residual correlations. The largest residual correlation was 0.14. With these empirically guided residual correlations included, CFI was 0.98, TLI was 0.97, and RMSEA was 0.07. We compared the loadings for single-factor models with and without residual correlations. Standardized factor loadings on the primary factor ranged from 0.72 to 0.87 for the model without residual correlations and from 0.73 to 0.86 for the model with residual correlations. All differences in item loadings between the two models were less than 0.02. These analyses confirmed previous findings (Cameron et al., 2008; Dum et al., 2008; Hepner et al., 2008; Huang et al., 2006) that the PHQ-9 was sufficiently unidimensional to proceed with IRT analyses (See Appendix Figure 1).

The PHQ-9 TCC is shown in Figure 1. This demonstrates a distinctly curvilinear shape, such that 1 point difference at the bottom of the standard score scale is associated with a greater difference in depression levels than a 1 point difference in standard scores in the middle of the scale. The PHQ-9 TIC and standard error of measurement curve are shown in Figure 2.

A histogram of participant scores is shown in Figure 3. There is a prominent floor effect, as over 40% of the cohort endorsed none of the items of the PHQ-9. Above that level, scores ranged continuously from very mild levels of depression to very severe levels.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Item-level DIF findings are shown in Table 2. Three of the items had DIF with respect to race, one with respect to age, one with respect to recent substance use, and two with respect to sex. No item-level DIF was found with HIV transmission risk factor or $CD4^+$ cell count nadir. In all, six of the nine PHQ-9 items had DIF with respect to at least one covariate.

Individual-level DIF impact for each of the covariates and for all of the covariates simultaneously is shown in Figure 4. These box plots demonstrate that the items found with DIF were not associated with a salient amount of DIF impact for any of the 1,452 participants.

Group-level DIF impact is summarized in Figure 5. For covariates other than race, these box plots indicate no particular pattern across groups. However, the box plot indicates that accounting for DIF for African-Americans was associated with scores that were somewhat higher, while for whites, accounting for DIF was associated with scores that were somewhat lower.

The effect of DIF on mean depression levels estimated across groups defined by each of the covariates is shown in Table 3. As anticipated based on Figure 5, race had the largest impact. Using unadjusted standardized scores, the mean depression score for African-Americans was 6.73 points lower than the mean depression score for whites. Using scores that accounted for DIF, this difference in mean scores was only 5.41 points. The difference between these two values, 1.32 points, or 20% of the observed difference in mean scores between African-Americans and whites, is due to DIF.

Discussion

Using data from a large cohort of individuals with HIV infection, we found that the PHQ-9 was well accepted, as the median completion time was around 1 minute and all 1,467 participants completed the questionnaire in less than 3½ minutes. The PHQ-9 has curvilinear scaling properties (Figure 1) and varying levels of measurement precision (Figure 2). It had a notable floor effect in our clinical setting (Figure 3). PHQ-9 items had DIF related to race (3 items), sex (2 items), substance use (1 item) and age (1 item) (Table 2). This DIF was associated with negligible individual-level DIF impact (Figure 4) except for race, 20% of the difference between African-Americans and whites was attributable to DIF (Figure 5 and Table 3).

The finding that the PHQ-9 has curvilinear scaling properties suggests that it may be inadvisable to use standard PHQ-9 scores for several sorts of analyses. For example, the minimal clinically important difference for the PHQ-9 has been estimated to be 5 points (Lowe, Unutzer, Callahan, Perkins, & Kroenke, 2004). Figure 1 demonstrates that a 5 point difference in PHQ-9 scores represents a larger amount of difference in depression levels at the lowest end of the scale than in the middle part of the scale. For individuals with low PHQ-9 standard scores, 5 points represents a much more salient amount of depression change than for individuals with moderate PHQ-9 standard scores.

Curvilinear scaling is also important in the context of longitudinal data analyses (Patten, 2008; Patten & Schopflocher, 2009). Unless appropriate methods are used (P. Crane et al., 2008b; Proust, Jacqmin-Gadda, Taylor, Ganiayre, & Commenges, 2006), using standard scores may lead to biased estimates of rates of change. Using simulated longitudinal data, we have previously shown that failing to account for curvilinearity by using standard scoring resulted in biased estimates of the rate of change compared with an IRT approach that accounts for curvilinearity (P. Crane et al., 2008b).

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

DIF is present in a test item for a covariate if, when controlling for the trait or ability measured by the test, individuals in groups defined by the covariate have unequal probabilities of endorsing the item (Camilli & Shepard, 1994; Holland & Wainer, 1993). DIF analyses have been the mainstay of bias detection procedures in educational testing for several decades, and have become increasingly common in medical settings.

The analyses of DIF shown here are the most thorough for the PHQ-9 to date. The sample is both large enough and diverse enough to permit appropriate analyses of several covariates. Prior studies have either analyzed only race (Hepner et al., 2008; Huang et al., 2006) or included data on too few participants for robust DIF analyses (Lamoureux et al., 2009). In contrast to previous papers that analyzed race (Hepner et al., 2008; Huang et al., 2006), we found several PHQ-9 items had DIF with respect to race. The DIF impact findings suggest that these items were associated with minimal individual-level DIF impact. For clinicians caring for individual patients, that result is very reassuring, as DIF did not interfere with any particular person's score by very much. However, group-level DIF impact findings suggest that 20% of the difference in mean scores between African-Americans and whites was attributable to DIF. For social scientists studying factors associated with depression at a population level, that result suggests the need to use scores that account for DIF rather than naïve scores, since 1/5 of the difference across groups using unadjusted scores was due to DIF.

Several strengths and weaknesses should be considered when considering our findings. The sample is large and diverse, and was a convenience sample recruited from usual clinical care from busy clinics in two cities. The measure was well tolerated so there is minimal missing data. While the sample was diverse with respect to the covariates analyzed here, there were too few study participants who were neither white nor African-American to permit DIF analyses of other race/ethnicity groups. An earlier paper on DIF in the PHQ-9 evaluated DIF in Asian-Americans and whites (Huang et al., 2006) and did not find any items with DIF. Since we found several items with DIF related to race in African-Americans and whites while the earlier paper found none (Huang et al., 2006), it is difficult to be reassured regarding use of the PHQ-9 in Asian-Americans infected with HIV.

In conclusion, the PHQ-9 was well-tolerated in a large and diverse sample of individuals with HIV infection. Standard scores may be used in the care of individual patients, though caution should be exercised when evaluating change over time due to curvilinear scaling. Group-level mean differences between African-Americans and whites may be exaggerated due to DIF.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was presented in part at the International Conference on Outcomes Measurement, Bethesda, Maryland, September, 2008.

References

Adewuya AO, Ola BA, Afolabi OO. Validity of the patient health questionnaire (PHQ-9) as a screening tool for depression amongst Nigerian university students. J Affect Disord. 2006; 96(1–2): 89–93. [PubMed: 16857265]

American Psychiatric Association. Task Force on DSM-IV. Diagnostic and statistical manual of mental disorders : DSM-IV. 4th ed.. Washington, DC: American Psychiatric Association; 1994.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Arnsten JH, Demas PA, Grant RW, Gourevitch MN, Farzadegan H, Howard AA, et al. Impact of active drug use on antiretroviral therapy adherence and viral suppression in HIV-infected drug users. J Gen Intern Med. 2002; 17(5):377–381. [PubMed: 12047736]

Balfour L, Kowal J, Silverman A, Tasca GA, Angel JB, Macpherson PA, et al. A randomized controlled psycho-education intervention trial: Improving psychological readiness for successful HIV medication adherence and reducing depression before initiating HAART. AIDS Care. 2006; 18(7):830–838. [PubMed: 16971295]

Blanch J, Rousaud A, Hautzinger M, Martinez E, Peri JM, Andres S, et al. Assessment of the efficacy of a cognitive-behavioural group psychotherapy programme for HIV-infected patients referred to a consultation-liaison psychiatry department. Psychother Psychosom. 2002; 71(2):77–84. [PubMed: 11844943]

Boarts JM, Sledjeski EM, Bogart LM, Delahanty DL. The differential impact of PTSD and depression on HIV disease markers and adherence to HAART in people living with HIV. AIDS Behav. 2006; 10(3):253–261. [PubMed: 16482405]

Bouhnik AD, Preau M, Vincent E, Carrieri MP, Gallais H, Lepeu G, et al. Depression and clinical progression in HIV-infected drug users treated with highly active antiretroviral therapy. Antivir Ther. 2005; 10(1):53–61. [PubMed: 15751763]

Burack JH, Barrett DC, Stall RD, Chesney MA, Ekstrand ML, Coates TJ. Depressive symptoms and CD4 lymphocyte decline among HIV-infected men. JAMA. 1993; 270(21):2568–2573. [PubMed: 7901433]

Cameron IM, Crawford JR, Lawton K, Reid IC. Psychometric comparison of PHQ-9 and HADS for measuring depression severity in primary care. Br J Gen Pract. 2008; 58(546):32–36. [PubMed: 18186994]

Camilli, G.; Shepard, LA. Methods for identifying biased test items. Thousand Oaks: Sage; 1994.

Carrico AW, Antoni MH, Weaver KE, Lechner SC, Schneiderman N. Cognitive-behavioural stress management with HIV-positive homosexual men: mechanisms of sustained reductions in depressive symptoms. Chronic Illn. 2005; 1(3):207–215. [PubMed: 17152183]

Catz SL, Kelly JA, Bogart LM, Benotsch EG, McAuliffe TL. Patterns, correlates, and barriers to medication adherence among persons prescribed new treatments for HIV disease. Health Psychol. 2000; 19(2):124–133. [PubMed: 10762096]

Ciesla JA, Roberts JE. Meta-analysis of the relationship between HIV infection and risk for depressive disorders. Am J Psychiatry. 2001; 158(5):725–730. [PubMed: 11329393]

Crane HM, Grunfeld C, Harrington RD, Uldall KK, Ciechanowski PS, Kitahata MM. Lipoatrophy among HIV-infected patients is associated with higher levels of depression than lipohypertrophy. HIV Med. 2008a; 9(9):780–786. [PubMed: 18754804]

Crane HM, Lober W, Webster E, Harrington RD, Crane PK, Davis TE, et al. Routine collection of patient-reported outcomes in an HIV clinic setting: the first 100 patients. Curr HIV Res. 2007; 5(1):109–118. [PubMed: 17266562]

Crane PK, Gibbons LE, Jolley L, van Belle G. Differential item functioning analysis with ordinal logistic regression techniques: DIFdetect and difwithpar. Med Care. 2006; 44 Suppl 3(11):S115–S123. [PubMed: 17060818]

Crane PK, Narasimhalu K, Gibbons LE, Mungas DM, Haneuse S, Larson EB, et al. Item response theory facilitated cocalibrating cognitive tests and reduced bias in estimated rates of decline. J Clin Epidemiol. 2008b; 61(10):1018–1027. e1019. [PubMed: 18455909]

Currier MB, Molina G, Kato M. A prospective trial of sustained-release bupropion for depression in HIV-seropositive and AIDS patients. Psychosomatics. 2003; 44(2):120–125. [PubMed: 12618534]

Currier MB, Molina G, Kato M. Citalopram treatment of major depressive disorder in Hispanic HIV and AIDS patients: a prospective study. Psychosomatics. 2004; 45(3):210–216. [PubMed: 15123845]

Dalessandro M, Conti CM, Gambi F, Falasca K, Doyle R, Conti P, et al. Antidepressant therapy can improve adherence to antiretroviral regimens among HIV-infected and depressed patients. J Clin Psychopharmacol. 2007; 27(1):58–61. [PubMed: 17224714]

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

DiMatteo MR, Lepper HS, Croghan TW. Depression is a risk factor for noncompliance with medical treatment: meta-analysis of the effects of anxiety and depression on patient adherence. Arch Intern Med. 2000; 160(14):2101–2107. [PubMed: 10904452]

Donnelly PL, Kim KS. The Patient Health Questionnaire (PHQ-9K) to screen for depressive disorders among immigrant Korean American elderly. J Cult Divers. 2008; 15(1):24–29. [PubMed: 19172976]

Duffy FF, Chung H, Trivedi M, Rae DS, Regier DA, Katzelnick DJ. Systematic use of patient-rated depression severity monitoring: is it helpful and feasible in clinical psychiatry? Psychiatr Serv. 2008; 59(10):1148–1154. [PubMed: 18832500]

Dum M, Pickren J, Sobell LC, Sobell MB. Comparing the BDI-II and the PHQ-9 with outpatient substance abusers. Addict Behav. 2008; 33(2):381–387. [PubMed: 17964079]

Gilbody S, Richards D, Brealey S, Hewitt C. Screening for depression in medical settings with the Patient Health Questionnaire (PHQ): a diagnostic meta-analysis. J Gen Intern Med. 2007; 22(11): 1596–1602. [PubMed: 17874169]

Gordillo V, del Amo J, Soriano V, Gonzalez-Lahoz J. Sociodemographic and psychological variables influencing adherence to antiretroviral therapy. AIDS. 1999; 13(13):1763–1769. [PubMed: 10509579]

Hepner KA, Hunter SB, Edelen MO, Zhou AJ, Watkins K. A comparison of two depressive symptomatology measures in residential substance abuse treatment clients. J Subst Abuse Treat. 2009

Hepner KA, Morales LS, Hays RD, Edelen MO, Miranda J. Evaluating differential item functioning of the PRIME-MD mood module among impoverished black and white women in primary care. Womens Health Issues. 2008; 18(1):53–61. [PubMed: 18069001]

Hides L, Lubman DI, Devlin H, Cotton S, Aitken C, Gibbie T, et al. Reliability and validity of the Kessler 10 and Patient Health Questionnaire among injecting drug users. Aust N Z J Psychiatry. 2007; 41(2):166–168. [PubMed: 17464695]

Holland, PW.; Wainer, H., editors. Differential item functioning. Hillsdale, N.J.: Erlbaum; 1993.

Huang FY, Chung H, Kroenke K, Delucchi KL, Spitzer RL. Using the Patient Health Questionnaire-9 to measure depression among racially and ethnically diverse primary care patients. J Gen Intern Med. 2006; 21(6):547–552. [PubMed: 16808734]

Ickovics JR, Hamburger ME, Vlahov D, Schoenbaum EE, Schuman P, Boland RJ, et al. Mortality, CD4 cell count decline, and depressive symptoms among HIV- seropositive women: longitudinal analysis from the HIV Epidemiology Research Study. JAMA. 2001; 285(11):1466–1474. [PubMed: 11255423]

Kitahata MM, Rodriguez BG, Haubrich R, Boswell S, Mathews WC, Lederman MM, et al. Cohort profile: the Centers for AIDS Research (CFAR) Network of Integrated Clinical Systems (CNICS). Int J Epidemiol. 2008; 37(5):948–955. [PubMed: 18263650]

Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med. 2001; 16(9):606–613. [PubMed: 11556941]

Lai JS, Crane PK, Cella D. Factor analysis techniques for assessing sufficient unidimensionality of cancer related fatigue. Qual Life Res. 2006; 15(7):1179–1190. [PubMed: 17001438]

Lamers F, Jonkers CC, Bosma H, Penninx BW, Knottnerus JA, van Eijk JT. Summed score of the Patient Health Questionnaire-9 was a reliable and valid method for depression screening in chronically ill elderly patients. J Clin Epidemiol. 2008; 61(7):679–687. [PubMed: 18538262]

Lamoureux EL, Tee HW, Pesudovs K, Pallant JF, Keeffe JE, Rees G. Can Clinicians Use the PHQ-9 to Assess Depression in People with Vision Loss? Optom Vis Sci. 2009

Lee PW, Schulberg HC, Raue PJ, Kroenke K. Concordance between the PHQ-9 and the HSCL-20 in depressed primary care patients. J Affect Disord. 2007; 99(1–3):139–145. [PubMed: 17049999]

Lotrakul M, Sumrithe S, Saipanish R. Reliability and validity of the Thai version of the PHQ-9. BMC Psychiatry. 2008; 8:46. [PubMed: 18570645]

Lowe B, Unutzer J, Callahan CM, Perkins AJ, Kroenke K. Monitoring depression treatment outcomes with the patient health questionnaire-9. Med Care. 2004; 42(12):1194–1201. [PubMed: 15550799]

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Monahan PO, Shacham E, Reece M, Kroenke K, Ong'or WO, Omollo O, et al. Validity/reliability of PHQ-9 and PHQ-2 depression scales among adults living with HIV/AIDS in western Kenya. J Gen Intern Med. 2009; 24(2):189–197. [PubMed: 19031037]

Morrison MF, Petitto JM, Ten Have T, Gettes DR, Chiappini MS, Weber AL, et al. Depressive and anxiety disorders in women with HIV infection. Am J Psychiatry. 2002; 159(5):789–796. [PubMed: 11986133]

Mungas D, Reed BR, Marshall SC, Gonzalez HM. Development of psychometrically matched English and Spanish language neuropsychological tests for older persons. Neuropsychology. 2000; 14(2): 209–223. [PubMed: 10791861]

Muraki, E.; Bock, D. PARSCALE for Windows (Version 4.1). Chicago: Scientific Software International; 2003.

Newcombe DA, Humeniuk RE, Ali R. Validation of the World Health Organization Alcohol, Smoking and Substance Involvement Screening Test (ASSIST): report of results from the Australian site. Drug Alcohol Rev. 2005; 24(3):217–226. [PubMed: 16096125]

Omoro SA, Fann JR, Weymuller EA, Macharia IM, Yueh B. 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. 2006; 36(3):367–381. [PubMed: 17236703]

Paterson DL, Swindells S, Mohr J, Brester M, Vergis EN, Squier C, et al. Adherence to protease inhibitor therapy and outcomes in patients with HIV infection. Ann Intern Med. 2000; 133(1):21–30. [PubMed: 10877736]

Patten SB. Confounding by severity and indication in observational studies of antidepressant effectiveness. Can J Clin Pharmacol. 2008; 15(2):e367–e371. [PubMed: 18840922]

Patten SB, Schopflocher D. Longitudinal epidemiology of major depression as assessed by the Brief Patient Health Questionnaire (PHQ-9). Compr Psychiatry. 2009; 50(1):26–33. [PubMed: 19059510]

Proust C, Jacqmin-Gadda H, Taylor JM, Ganiayre J, Commenges D. A nonlinear model with latent process for cognitive evolution using multivariate longitudinal data. Biometrics. 2006; 62(4): 1014–1024. [PubMed: 17156275]

Reeve BB, Hays RD, Bjorner JB, Cook KF, Crane PK, Teresi JA, et al. Psychometric evaluation and calibration of health-related quality of life item banks: plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). Med Care. 2007; 45 Suppl 1(5):S22–S31. [PubMed: 17443115]

Rogers WH, Adler DA, Bungay KM, Wilson IB. Depression screening instruments made good severity measures in a cross-sectional analysis. J Clin Epidemiol. 2005; 58(4):370–377. [PubMed: 15862723]

Rousaud A, Blanch J, Hautzinger M, De Lazzari E, Peri JM, Puig O, et al. Improvement of psychosocial adjustment to HIV-1 infection through a cognitive-behavioral oriented group psychotherapy program: a pilot study. AIDS Patient Care STDS. 2007; 21(3):212–222. [PubMed: 17428189]

Samejima F. Estimation of latent ability using a response pattern of graded scores. Psychometrika Monograph, No. 17. 1969

Samejima, F. Graded response model. In: van der Linden, WJ.; Hambleton, RK., editors. Handbook of modern item response theory. NY: Springer; 1997. p. 85-100.

Scogin F, Shah A. Screening older adults for depression in primary care settings. Health Psychol. 2006; 25(6):675–677. [PubMed: 17100495]

Screening for depression: recommendations and rationale. Ann Intern Med. 2002; 136(10):760–764. [PubMed: 12020145]

Singh N, Squier C, Sivek C, Wagener M, Nguyen MH, Yu VL. Determinants of compliance with antiretroviral therapy in patients with human immunodeficiency virus: prospective assessment with implications for enhancing compliance. AIDS Care. 1996; 8(3):261–269. [PubMed: 8827119]

Spitzer RL, Kroenke K, Williams JB. Validation and utility of a self-report version of PRIME-MD: the PHQ primary care study. Primary Care Evaluation of Mental Disorders. Patient Health Questionnaire. JAMA. 1999; 282(18):1737–1744. [PubMed: 10568646]

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Staab JP, Datto CJ, Weinrieb RM, Gariti P, Rynn M, Evans DL. Detection and diagnosis of psychiatric disorders in primary medical care settings. Med Clin North Am. 2001; 85(3):579–596. [PubMed: 11349474]

Stafford L, Berk M, Jackson HJ. Validity of the Hospital Anxiety and Depression Scale and Patient Health Questionnaire-9 to screen for depression in patients with coronary artery disease. Gen Hosp Psychiatry. 2007; 29(5):417–424. [PubMed: 17888808]

Starace F, Ammassari A, Trotta MP, Murri R, De Longis P, Izzo C, et al. Depression is a risk factor for suboptimal adherence to highly active antiretroviral therapy. J Acquir Immune Defic Syndr. 2002; 31(Suppl 3):S136–S139. [PubMed: 12562037]

Waldrop-Valverde D, Valverde E. Homelessness and psychological distress as contributors to antiretroviral nonadherence in HIV-positive injecting drug users. AIDS Patient Care STDS. 2005; 19(5):326–334. [PubMed: 15916495]

WHO ASSIST Working Group. The Alcohol, Smoking and Substance Involvement Screening Test (ASSIST): development, reliability and feasibility. Addiction. 2002; 97(9):1183–1194. [PubMed: 12199834]

Williams RT, Heinemann AW, Bode RK, Wilson CS, Fann JR, Tate DG. Improving measurement properties of the Patient Health Questionnaire-9 with rating scale analysis. Rehabil Psychol. 2009; 54(2):198–203. [PubMed: 19469610]

Yeung A, Fung F, Yu SC, Vorono S, Ly M, Wu S, et al. Validation of the Patient Health Questionnaire-9 for depression screening among Chinese Americans. Compr Psychiatry. 2008; 49(2):211–217. [PubMed: 18243896]

Yun LW, Maravi M, Kobayashi JS, Barton PL, Davidson AJ. Antidepressant treatment improves adherence to antiretroviral therapy among depressed HIV-infected patients. J Acquir Immune Defic Syndr. 2005; 38(4):432–438. [PubMed: 15764960]

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

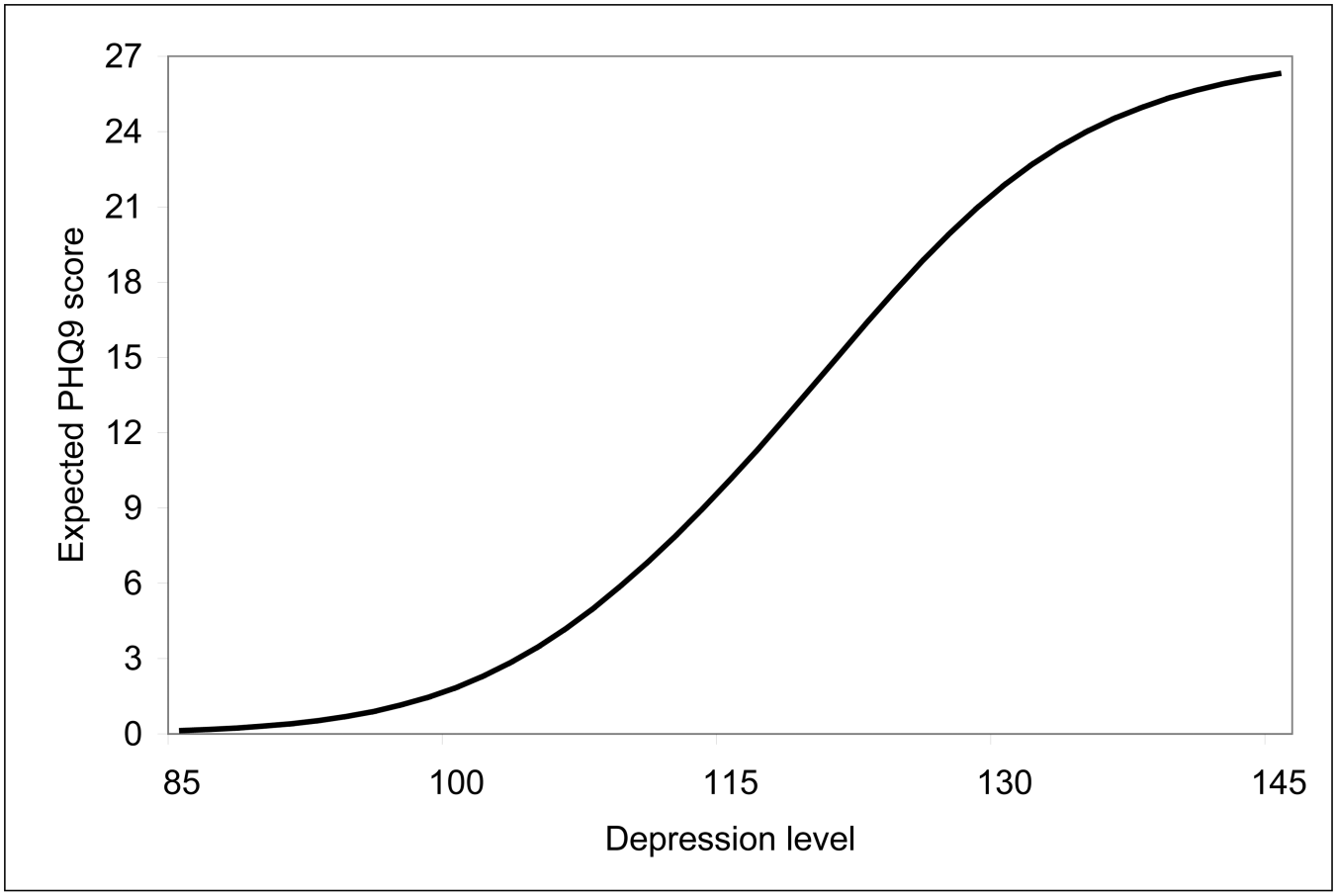

Figure 1.
Test characteristic curve for the PHQ-9*
* This graph plots the most likely standard PHQ-9 score associated with each level of depression. Item response theory scores were transformed to have a mean of 100 and a standard deviation of 15. At the mean level of depression in this cohort (depression level of 100), the expected PHQ-9 score is only 2 points. This means that the PHQ-9 provides very little discrimination among individuals with very low levels of depression. The sigmoid shape of the test characteristic curve implies that differences between standard scores have different implications depending on the starting value. For example, for individuals with severe levels of depression (standard PHQ-9 scores over 20), the curve is flatter than for individuals with moderate levels of depression (standard PHQ-9 scores over 10). The same 5 standard score points implies more change in depression for someone whose baseline score was 25 than for someone whose baseline score was 15. Curvilinear scaling has implications for longitudinal analyses (where rates of change are not uniform across levels of depression) and for cross-sectional regression analyses (where a single coefficient is used to describe the relationship between a 1-unit change in depression and some other variable).

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

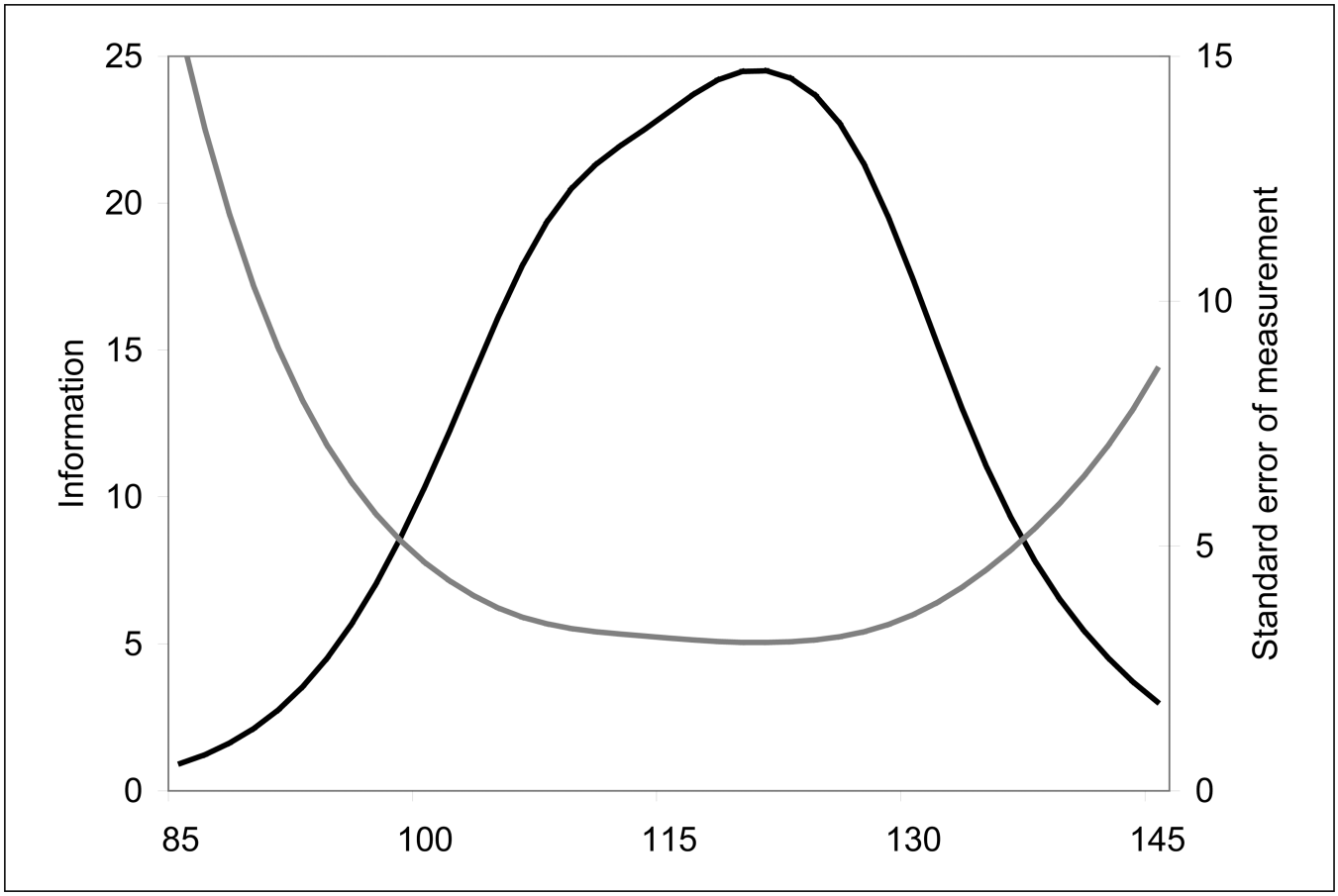

Figure 2.
Test information curve and standard error of measurement for the PHQ-9*.
* The black curve shows the amount of measurement precision ("information") at each depression level. The gray curve shows the standard error of measurement associated with each depression level. The PHQ-9 is characterized by fairly good reliability for individuals with depression levels from 100–130, while below 100 the reliability of the instrument is quite limited, and above 130 or so the reliability again begins to diminish. The clinical implication of this finding is that PHQ-9 scores between 100 and 135 or so are characterized by a standard error of 5 points or fewer, while scores below 100 and above 135 are characterized by larger standard errors. This means those with low levels of depression (<100 points) and high levels of depression (>135 points) are measured less accurately than those with moderate levels of depression. These results are to be contrasted with Crohnbach's alpha, which would provide a single omnibus statistic summarizing reliability as if it were a constant across the range of depression measured by the test. Item response theory (IRT) output provides both the point estimate of the individual's score along with the standard error associated with that score. Clinicians should become used to seeing both of these results reported.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

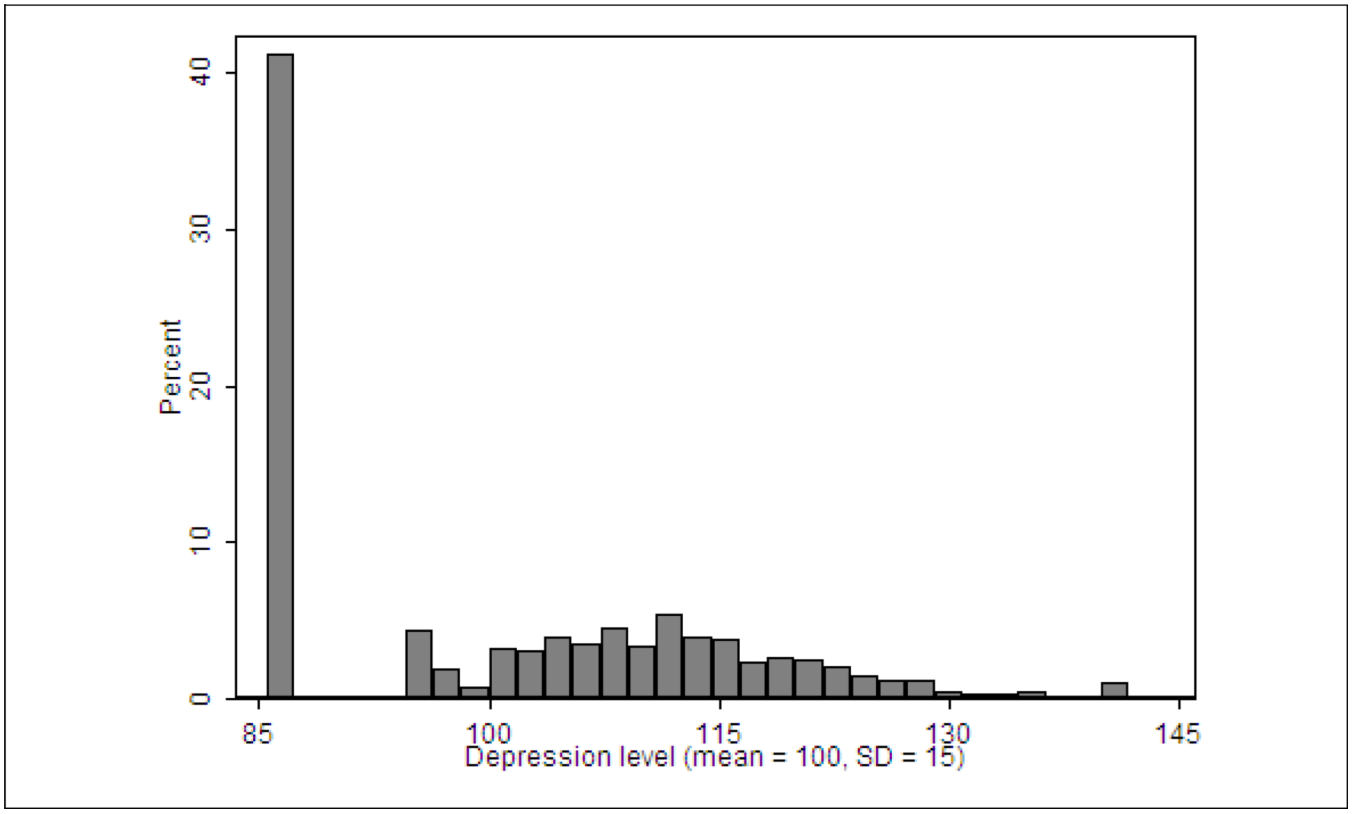

Figure 3.
Histogram of depression levels in this cohort (N=1452).

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

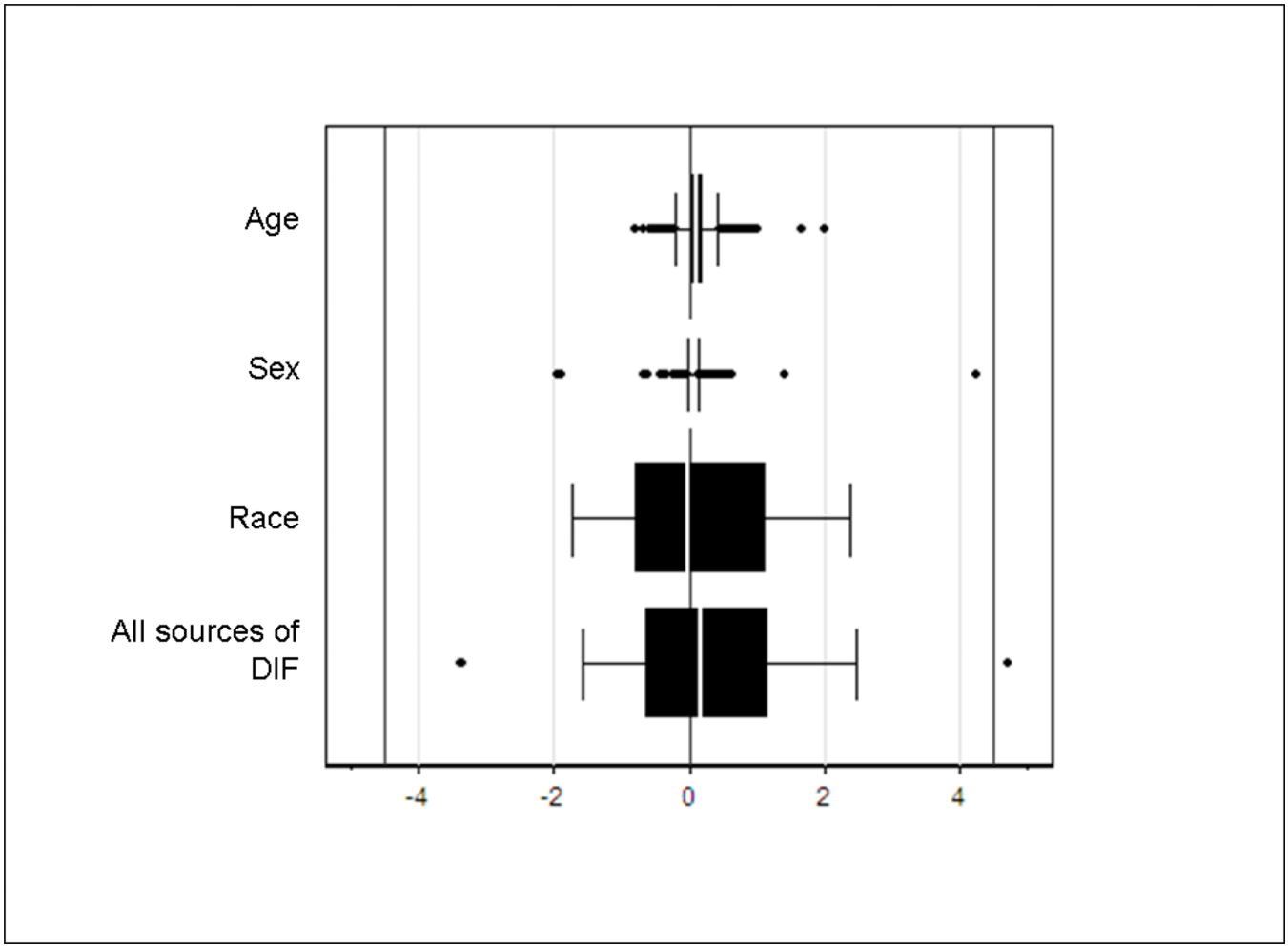

Figure 4.

Individual-level DIF impact with respect to each of the covariates and with respect to all of the covariates simultaneously*

* No items had DIF related to substance use, nadir CD4 count, or transmission risk factor. Vertical lines are placed at 0 (indicating no DIF) and at +4.5 and −4.5, indicating the median standard error of measurement (SEM). The first three box-and-whisker plots delineate individual-level DIF impact associated with each of the covariates evaluated in turn, while the last plot delineates individual-level DIF impact associated with all the covariates considered here. The values summarized in the box plots are the differences between the unadjusted IRT score and IRT scores that accounted for DIF associated with each covariate (first three plots) or with multiple covariates (last plot). A difference of 0 (the middle reference line) would mean that DIF made no difference for that person. Large positive values indicate that scores accounting for DIF were higher than scores that ignored DIF, which means that ignoring DIF resulted in underestimates of depression severity. Large negative values indicate that scores accounting for DIF were lower than scores that ignored DIF; thus ignoring DIF resulted in overestimates of depression severity. These box-and-whisker plots are indexed by 1× and 2× the median SEM of the PHQ-9 among these participants. Observations outside of ± 1 SEM indicate that a covariate has salient individual-level DIF impact (first three plots) or that the covariates evaluated for multiple sources of DIF considered together have salient individual-level DIF impact (last plot).

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

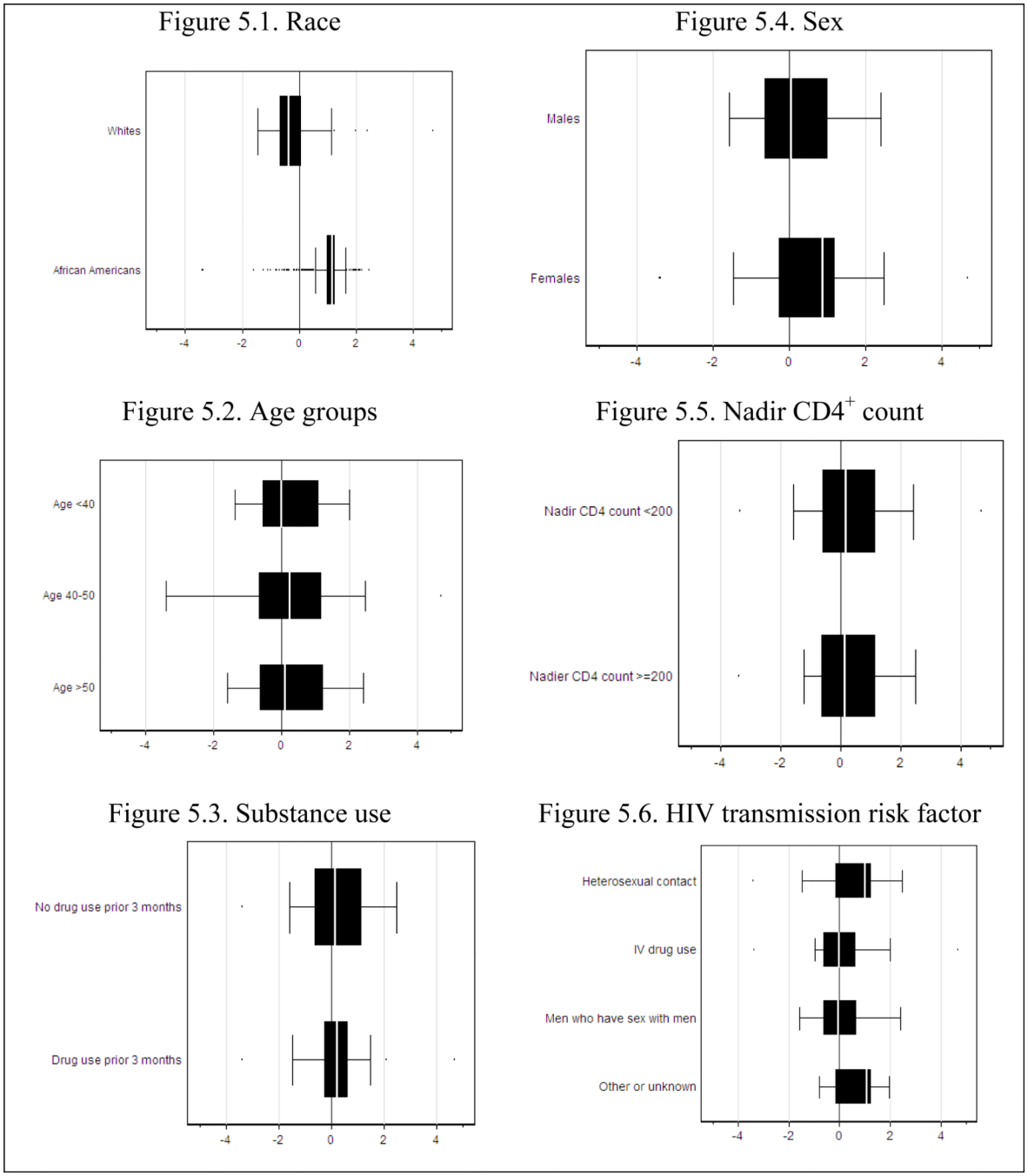

Figure 5.

Group level DIF impact across groups defined by each of the covariates. DIF=differential functioning. In these plots, differences between scores accounting for DIF related to all covariates and the unadjusted IRT scores (depicted in the bottom plot of Figure 3) are plotted across groups defined by each of the covariates. The median effect for African-Americans is in the opposite direction of the median effect for whites (Figure 5.1).

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

Table 1

Demographic and clinical characteristics of study participants stratified by race/ethnicity (N=1452)

	White		African- American		Other*		Total	
	N=825	%	N=552	%	N=75	%	N=1452	%
Sex								
Male	725	88	384	70	63	84	1172	81
Female	100	12	168	30	12	16	280	19
Age in years								
< 40	249	30	197	36	31	41	477	33
40–49	355	43	211	38	30	40	596	41
≥50	221	27	144	26	14	19	379	26
CD4+ nadir (cells/mm^3)†								
< 200	478	58	335	61	49	65	862	59
200–350	214	26	135	24	17	23	366	25
>350	133	16	82	15	9	12	224	15
Depression severity as defined by the PHQ-9								
None (0–4)	439	53	388	70	40	53	867	60
Mild (5–9)	158	19	85	15	17	23	260	18
Moderate (10–19)	178	22	64	12	14	19	256	18
Severe (≥20)	50	6	15	3	4	5	69	5
Recent substance use (prior 3 months)								
No	704	85	492	89	56	75	1252	86
Yes	121	15	60	11	19	25	200	14
HIV transmission risk factor‡								
MSM	510	62	195	35	37	49	742	51
IVDU	154	19	52	9	14	19	220	15
Heterosexual contact	120	15	248	45	18	24	386	27
Other/Unknown	41	5	57	10	6	8	104	7

*American Indian (n=15), Asian-American (n=14), Hawaiian (n=7), Multiracial (n=7), Other or unknown (n=32).

†Due to the skewed distribution, we used categories of <200, 200–299, and ≥300 for DIF analyses.

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

[‡]We analyzed this covariate as MSM vs. other to identify items with DIF, though we plot values of cumulative DIF impact for each of these groups separately in Figure 5.

PHQ-9 = Patient Health Questionnaire. MSM = men who have sex with men. IVDU = intravenous drug use. Percentages may add up to more than 100% due to rounding.

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

Table 2

Item level DIF findings.

	Race		Age		Substance use		Sex		Nadir CD4± count		Transmission risk factor	
Item	**N**	**U**	**N**	**U**	**N**	**U**	**N**	**U**	**N**	**U**	**N**	**U**
1. Little interest or pleasure in doing things	0.11	0.44	0.83	0.22	0.52	0.47	0.75	0.08	0.80	0.86	0.95	0.56
2. Feeling down, depressed, or hopeless	0.04	0.82	0.21	0.92	0.17	0.61	0.78	1.00	0.42	0.87	0.57	0.84
3. Trouble falling asleep or staying asleep, or sleeping too much	**0.03**	**<0.01**	0.20	0.78	0.29	0.38	0.04	0.73	0.27	0.50	0.07	0.76
4. Feeling tired or having little energy	**0.03**	**<0.01**	0.50	0.59	0.53	0.14	0.52	0.11	0.77	0.62	0.89	0.88
5. Poor appetite or overeating	0.13	0.87	0.19	0.91	**<0.01**	0.08	0.11	**<0.01**	0.50	0.90	0.79	0.88
6. Feeling bad about yourself, or that you are a failure or have let yourself or your family down	0.94	0.20	1.00	**<0.01**	0.96	0.08	0.85	0.46	0.82	0.98	0.82	0.85
7. Trouble concentrating on things, such as reading the newspaper or watching television	0.08	0.46	0.08	0.61	0.62	0.91	0.67	0.91	0.71	0.59	0.69	0.58
8. Moving or speaking so slowly that people could have noticed. Or the opposite: being so fidgety or restless you have been moving around a lot more than usual	**<0.01**	0.42	0.94	0.41	0.85	0.36	0.46	0.52	0.54	0.39	0.76	0.64
9. Thoughts that you would be better off dead, or of hurting yourself in some way	0.34	0.85	0.68	0.04	0.53	0.19	0.53	**<0.01**	0.41	0.30	0.23	0.53

DIF= differential item functioning. N=non-uniform DIF. U=uniform DIF.

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

AIDS Care. Author manuscript; available in PMC 2012 August 02.

Table 3

Mean unadjusted scores and scores accounting for all sources of DIF stratified by covariates

Covariate and score type	Group 1	Group 2	Group 3	Difference	Amount of difference	Percent of difference due to DIF
Race	African-Americans	Whites				
Unadjusted	95.97	102.70		6.73	1.32	20%
Accounting for all sources of DIF	96.76	102.17		5.41		
Age	<40	40–49.9	50+			
Unadjusted	100.10	100.74	98.73	0.64	−0.05	−8%
Accounting for all sources of DIF	100.06	100.75	98.76	0.69		
Substance use	No	Yes				
Unadjusted	98.43	110.40		11.97	0.03	0%
Accounting for all sources of DIF	98.43	110.37		11.94		
Sex	Male	Female				
Unadjusted	100.35	98.55		−1.80	−0.36	20%
Accounting for all sources of DIF	100.28	98.84		−1.44		
$CD4^{+}$ nadir category	≤200	200+				
Unadjusted	100.50	99.28		−1.22	0.05	−4%
Accounting for all sources of DIF	100.52	99.25		−1.27		
HIV transmission risk factor	MSM	Other				
Unadjusted	99.27	100.69		1.42	0.36	25%
Accounting for all sources of DIF	99.46	100.52		1.06		

DIF=differential item functioning