Development of a brief mental health impairment scale using a nationally representative sample in the USA

SCOTT P. NOVAK,¹ LISA J. COLPE,² PEGGY R. BARKER² & JOSEPH C. GFROERER²

¹ RT International, Research Triangle Park, NC, USA

² Substance Abuse and Mental Health Services Administration, Office of Applied Studies, Rockville, MD, USA

Key words

impairment, item response theory, mental health, World Health Organization Disability Assessment Schedule

Correspondence

Lisa J. Colpe, PhD, MPH;
National Institute of Mental Health, 6001 Executive Blvd,
Bethesda, MD 20892, USA.
Telephone +1 301-443-3815
Email: Lisa.Colpe@NIH.GOV

Received 8 March 2010;
revised 19 April 2010;
accepted 19 April 2010

Abstract

A psychometric analysis was conducted to reduce the number of items needed to assess the disability associated with mental disorders using the World Health Organization Disability Assessment Schedule (WHODAS). The WHODAS was to be used in the Substance Abuse and Mental Health Services Administration National Survey on Drug Use and Health (NSDUH), beginning in 2008, as part of a screening algorithm to produce estimates of the prevalence of serious mental illness (SMI) in the US adult population. The goal of the work presented in this paper was to create a parsimonious screening scale from the full 16-item WHODAS that was administered to 24 156 respondents (aged 18+) in the 2002 NSDUH. Exploratory factor analysis showed that WHODAS responses were unidimensional. A two-parameter polytomous Item Response Theory model showed that all 16 WHODAS items had good item discrimination (slopes greater than 1.0) for each response option. Analysis of item difficulties and differential item function across socio-demographic categories was then used to select a subset of eight items to create a short version of the WHODAS. The Pearson correlation between scores in the original 16-item and reduced eight-item WHODAS scales was 0.97, documenting that the vast majority of variation in total scale scores was retained in the reduced scale. Copyright © 2010 John Wiley & Sons, Ltd.

Introduction

It is well-known that accurate prevalence estimation is critical to assessing service delivery needs for people with mental disorders (Manderscheid *et al.*, 2005). Several community surveys are available to provide prevalence estimates of mental disorders in the USA, including the National Comorbidity Survey (Kessler *et al.*, 1994), the National Comorbidity Survey Replication (Kessler and Merikangas, 2004), the Epidemiological Catchment Area

Study (Robins and Regier, 1991), and the National Epidemiologic Survey on Alcohol and Related Conditions (Grant *et al.*, 2003). These surveys collected data using an in-depth, fully structured diagnostic interview, to generate prevalence estimates for a wide range of disorders using the diagnostic criteria in various editions of the American Psychiatric Association's Diagnostic and Statistical Manual of Mental Disorders (DSM) and the World Health Organization's International Statistical Classification of Diseases. Between 30 and 50% of the

respondents in these surveys were classified as meeting criteria for a mental disorder at some time in their life. A range of 20–30% of respondents was estimated to meet criteria for at least one such disorder at some time in the 12 months before the interview. The estimated population counts represented by these prevalence figures are much greater than the number who could realistically be treated under the current treatment system in the USA. This gap highlights the importance of triaging those based on levels of illness severity as a way to direct limited resources to those with the greatest level of impairment (Narrow *et al.*, 2002). This argument rests on the assumption that individuals who meet criteria for a mental disorder, but who do not exhibit significant impairment, represent a group with low need for treatment (Regier *et al.*, 1998). As many mental disorders are not seriously impairing, excluding those with mental disorders that are not seriously impairing would lead to a substantially lower estimate of treatment need than if all persons with a mental disorder were included (Kessler *et al.*, 2003b).

Recognizing the importance of impairment as a feature of mental illness, and as directed under Public Law (PL) 102-321, the Alcohol, Drug Abuse and Mental Health Administration Reorganization Act (Alcohol Drug Abuse and Mental Health Administration, 1992), the Substance Abuse and Mental Health Services Administration (SAMHSA) developed an operational definition of adult serious mental illness (SMI). This definition requires a DSM diagnosis other than the diagnosis of a substance use disorder or an organic mental disorder, that results in what the law refers to as “substantial functional impairment in one or more major life activities” (Substance Abuse and Mental Health Services Administration, 1993). This impairment requirement was subsequently operationalized by SAMHSA as a Global Assessment of Functioning (GAF) score of ≤ 50 (Substance Abuse and Mental Health Services Administration, 2005). It is of value to SAMHSA to track the population prevalence of SMI. The challenge in doing this is that large multipurpose surveys capable of monitoring the prevalence of SMI on an ongoing basis cannot devote the enormous amount of time developed in special-purpose surveys such as the National Comorbidity Survey, Epidemiological Catchment Area study, and National Epidemiologic Survey on Alcohol and Related Conditions to this task. Instead, large ongoing surveys use brief screening scales to estimate the prevalence of the disorders they monitor. These screening scales, while not providing the rich detail available in more in-depth assessments, are designed to provide good accuracy in estimating global

prevalence at a point in time as well as change in prevalence over time.

In an effort to develop a system for tracking trends in SMI, SAMHSA's Office of Applied Studies began a program of methodological research designed to develop a screening scale for SMI that could be used on an ongoing basis in the National Survey on Drug Use and Health (NSDUH). A methodological pilot study carried out by the Office of Applied Studies in 2000 found that a six-item scale of non-specific psychological distress, the K6 scale, did a good job of screening for SMI (Kessler *et al.*, 2003a). The K6 was added to the NSDUH based on this result. However, subsequent analyses of the K6 scale in the NSDUH showed that responses are sensitive to scale placement in the survey and suggested that concordance of survey responses with clinical diagnoses of SMI could be improved by expanding the NSDUH assessment of SMI beyond the questions in the K6. A decision was made to address these issues by adding an ongoing clinical calibration study to the NSDUH in which a probability sub-sample of NSDUH respondents would be administered a gold-standard clinical psychiatric diagnostic interview shortly after completing the NSDUH interview (Colpe *et al.*, 2009). Calibration rules can then be developed from this sub-sample to generate a predicted probability of SMI for each NSDUH respondent from responses to the K6 combined with impairment related questions designed to operationalize SMI and use this information as the basis for tracking national trends in the prevalence of SMI.

This study was undertaken in preparation for the clinical calibration study described in this issue by Colpe *et al.* (2010). Its purpose was to determine if the 16 WHODAS questions contained in the NSDUH could be reduced in number without loss of important information. The high inter-correlations among responses to the WHODAS questions in previous studies led us to believe that some item reduction of this sort would be possible. WHODAS data from the 2002 NSDUH was used to investigate the implications of item reduction of the precision of the scale. We also examined the impact of changing the WHODAS response options from a polytomous format to a dichotomous format.

Methods

Sample

The NSDUH is an ongoing face-to-face survey of the civilian, non-institutionalized population of the USA aged 12 years or older. This population includes residents of households, college dormitories, group homes,

homeless shelters, and rooming houses, as well as civilians living on military installations. Because details of the NSDUH sampling design are described elsewhere (Substance Abuse and Mental Health Services Administration, 2003), only a brief summary is provided here. NSDUH uses a multistage-clustered area probability sampling design to select dwelling units in the 50 states and the District of Columbia. An in-person screen is used to identify eligible residents of the sampled dwelling units and survey participants are then randomly selected from this roster of residents. Respondents are paid \$30 for participation in the survey, which uses computer-assisted administration methods. The data used in the current analysis were gathered as part of the NSDUH during calendar year 2002. The 2002 NSDUH had a weighted screening response rate of 91% and a weighted interview response rate of 79%. The survey data were weighted to yield estimates that would be representative for the entire US population in the age range of sample on the cross-classification of numerous census socio-demographic and geographic variables. The focus of the current report is on NSDUH respondents aged 18 years and older.

The 2002 NSDUH survey contained multiple measures of psychiatric symptoms, impairments and mental health service use to include: a series of disorder-specific items that had been determined to predict with certain probability meeting 12-month criteria for DSM-IV disorders from the World Health Organization (WHO) Composite International Diagnostic Interview Short-Form (CIDI-SF) scales (Kessler *et al.*, 1998), the K6/K10 scales of non-specific psychological distress (Furukawa *et al.*, 2003), and the NSDUH Adult Mental Health Service Utilization module. These measures were used to identify and exclude the members of the study sample who would likely not meet the criteria for a DSM disorder, because associated mental health impairments would not apply to those respondents. We restricted our analyses to those respondents with the following characteristics:

- 1 endorsing a disorder-specific item from the CIDI-SF screener
- 2 endorsing 'most' or 'some of the time' for any one of the non-specific distress items on the K6/K10 scale
- 3 seeing a mental health professional in the past 12 months
- 4 staying in a hospital or facility overnight for mental health problems in the past 12 months
- 5 receiving outpatient treatment for mental health problems in the past 12 months
- 6 taking prescription medication for a mental health condition in the past 12 months.

The final unweighted analytic sample size for 2002 was 24 156 (Substance Abuse and Mental Health Services Administration, 2003).

Measures

The 16-item WHODAS used in the NSDUH included items assessing levels of difficulty in social and physical domains such as carrying out daily activities in the domains of self-care, mobility, productive activity, community participation, and social and family life. The reference period for each item was the 'one month in the past 12 months in which interference due to your emotions, nerves, or mental health was the most.' Responses were coded using a four-category scale (severe, moderate, mild, or no difficulty). For seven of the items, a fifth response option was provided: 'You didn't perform this task'. For these items, a yes–no follow-up question asked whether mental problems interfered with the respondent's ability to perform the task. Responses to the question were recoded to either 'no difficulty' or 'severe difficulty' based on respondents' answers to the follow-up questions.

A study conducted by Kessler *et al.* (2003a) used an exploratory factor analysis to show that the 16 WHODAS items loaded on a primary principal factor of global impairment. However, there is a possibility that the level of impairment associated with particular response patterns might differ across socio-demographic sub-groups. Based on this possibility, we examined whether reports of impairment are affected by age, gender, race/ethnicity, and education. The socio-demographic definitions in this study are similar to those used in other NSDUH tables and reports. Age was coded as a categorical variable representing three levels (18–25, 26–49, 50+). Race/ethnicity was divided into non-Hispanic White, non-Hispanic Black, Hispanic, and other. Education was coded as less than high school, high school graduate, some college, and college graduate.

Analysis methods

After exploratory factor analysis documented the unidimensionality of the 16-item WHODAS scale, a two-parameter Item Response Theory (IRT) model (Steinberg and Thissen, 1995) was used to determine the extent to which a reduced set of items would retain all the information in the full scale. The information would then be used in tandem with a mental health screening scale to discriminate clinical cases of SMI from non-cases. Sub-group comparisons within the IRT framework were also used to examine the possibility of differential item functioning (DIF) in segments of the population defined by

age, gender, race/ethnicity, and education. The main advantage of IRT over traditional psychometric approaches is that IRT allows scale precision to be assessed easily across a specific range of the underlying latent trait (Emberson and Reise, 2000). In the current application, we are interested in distinguishing impairments caused by mental illness equivalent to a GAF score ≤ 50 . IRT is also useful in that it makes it possible to evaluate changes in overall scale precision at a given point on the scale range as the items included in the scale are changed (Reeve and Mâsse, 2004). This facilitates the selection of an optimal short set of items in our study.

We worked with a two-parameter IRT model. In these types of models, each scale item has a difficulty parameter (b) that represents the point on the latent true score scale (θ) where a person has a 50% chance of endorsing the level of response for the item. This point is known as the item 'inflection point'. It is important to note that the difficulty parameter is a feature of an item (much as we would say that different questions in a test differ in difficulty). For items that are dichotomous, only a single difficulty parameter is estimated. For polytomous items, estimation requires the calculation of multiple thresholds to account for multiple response options. We employed the Graded Response Model (Samejima, 1997) to estimate the polytomous IRT models. In this parameterization, b_i represents the threshold on the latent scale at which respondents have a 50% probability of endorsing level k or higher on item i . Because the highest response option is not estimated ($k - 1$), the polytomous IRT models of the WHODAS scale yielded three distinct thresholds for the response options of no difficulty (b_1), mild difficulty (b_2), and moderate difficulty (b_3).

We also estimated models that dichotomized response options into two categories: (1) severe and (2) moderate, mild, or no difficulty. For these models, the inflection point represents the level on the latent trait at which the respondent has a 50% probability of endorsing 'severe disability' on a given item. The second parameter in the two-parameter IRT model, the discrimination parameter (a_i), represents the strength of an item's discrimination between respondents above versus below the item inflection point. It is conventional in the IRT literature to focus on items that have slope parameters of 1.0 or greater and to select items of this sort for the final scale that have an inflection point near the point on the scale of primary importance. As noted above, this point in our case is a GAF of 50, which we estimate to be equivalent to a score at about the 95th percentile on the standardized true score in the general population (i.e. a score that would yield a prevalence of SMI of about 5%). In selecting the

optimal short version of the WHODAS for our purposes, we made use of the fact that each IRT scale yields a test information curve (TIC), a graphic representation of the precision of measurement of the scale at each point on the scale range. The shape of the TIC is dependent on the difficulty and discrimination parameters of the items included in the scale. We sought a short scale that had good precision near the 95th percentile of the true score distribution. In addition, we wanted the items in this short scale to have IRT parameters that were relatively invariant across the socio-demographic sub-samples considered in the analysis.

The analysis proceeded in five steps. First, exploratory factor analysis of the 16-item WHODAS scale was used to confirm a strong unidimensional structure. Second, once it was established that the scale was unidimensional, we estimated a single polytomous two-parameter IRT model to the full 16 items. Third, we identified a subset of WHODAS items based on their IRT parameters that had inflection points in the scale range of interest along with good discrimination. We then re-estimated the model with this reduced set of items and compared the TIC for the reduced scale versus the full scale. Fourth, we examined measurement invariance of the reduced scale across major socio-demographic sub-categories. Finally, we considered whether the reduced scale could be further simplified by reducing the number of response options from four to two. These activities resulted in three primary scales that are further described below: the full 16-item WHODAS, scored on a range of 0 to 48 with four response options; a reduced WHODAS scale with four response options; and a reduced WHODAS scale with dichotomous response options (severe versus moderate, mild, or no difficulty). All descriptive analyses were conducted using SUDAAN 9.0.1 (Research Triangle Institute, 2006), while the factor analysis and IRT analysis were conducted using MPLUS 4.4 (Muthén and Muthén, 1998–2007). Both packages can correctly account for the multistage sampling design of NSDUH using sample weights and a Taylor series linearization method to account for the design effect. Evaluation of statistical significance was carried out consistently at the 0.05 level using two-sided design-based tests.

Results

Factor structure of the WHODAS

Exploratory factor analysis found only one unrotated principal factor with an eigenvalue greater than 1.0. This first factor accounted for nearly 60% of the variance in the WHODAS items (Table 1). Factor loadings were in the

Table 1 Descriptive statistics for WHO-DAS items: National Survey on Drug Use and Health

8-item Brief		No difficulty % (SE)	Mild difficulty % (SE)	Moderate difficulty % (SE)	Severe difficulty % (SE)	16-item* factor loading	8-item** factor loading
1: REMBR	Remembering to do things you needed to do	23.3 (0.7)	45.6 (0.8)	25.5 (0.7)	5.3 (0.4)	0.76	0.63
2: CONC	Concentrating or doing something important when other things were going on	19.9 (0.7)	47.1 (0.8)	27.4 (0.7)	5.5 (0.4)	0.80	0.68
3: GO	Going out of the house and getting around on your own	67.0 (0.7)	20.0 (0.6)	8.7 (0.4)	4.1 (0.3)	0.77	0.58
4: NOTKNOW	Dealing with people you did you know well	44.2 (0.8)	32.6 (0.8)	16.9 (0.6)	6.1 (0.3)	0.75	0.65
5: SOCIAL	Participating in social activities	32.0 (0.7)	31.8 (0.8)	24.8 (0.7)	11.2 (0.5)	0.78	0.68
6: HOUSE	Taking care of household responsibilities	32.0 (0.8)	35.8 (0.8)	23.9 (0.7)	8.2 (0.4)	0.80	0.74
7: DAILY	Taking care of your daily responsibilities at work or school	39.1 (0.8)	39.3 (0.8)	16.3 (0.6)	5.2 (0.4)	0.85	0.78
8: W ORK	Getting your daily work done as quickly as needed	27.6 (0.7)	43.9 (0.8)	22.5 (0.7)	5.9 (0.4)	0.83	0.79
16-item Full							
9: THINK	Thinking clearly	26.6 (0.7)	42.0 (0.8)	26.7 (0.7)	4.9 (0.4)	0.74	
10: UNDERSTAND	Understanding what was going on around you	59.7 (0.8)	27.6 (0.7)	10.7 (0.5)	1.9 (0.2)	0.72	
11: MONEY	Managing your own money	42.8 (0.8)	27.3 (0.7)	21.6 (0.7)	8.2 (0.4)	0.68	
12: LEARN	Learning a new task	43.4 (0.8)	33.5 (0.7)	17.6 (0.6)	5.3 (0.4)	0.73	
13: DRESS	Washing, dressing, or feeding yourself	80.5 (0.7)	12.7 (0.5)	4.9 (0.4)	1.4 (0.2)	0.78	
14: GETTINGALNG	Getting along with people in your personal life	33.9 (0.8)	41.4 (0.8)	19.9 (0.6)	4.7 (0.3)	0.75	
15: EMOTIONS	Controlling your emotions when you were around people	38.1 (0.8)	40.8 (0.8)	17.2 (0.5)	3.7 (0.2)	0.77	
16: RELAX	Doing things to relax at home	33.7 (0.8)	35.8 (0.8)	24.2 (0.7)	6.1 (0.3)	0.72	

Estimates (per cent, standard error, factor loadings) adjusted for sampling design of 2002 NSDUH (N = 11 178).

*Eigen values for 16-item scale: factor 1 = 9.6, factors 2 to 16, all less than 1.

**Eigen values for 8-item scale: factor 1 = 4.3, factors 2 to 8, all less than 1.

range 0.68–0.85. Based on this result, we considered only a single factor in subsequent IRT analysis.

Full scoring for the 16-item WHODAS

As noted above in the section on analysis methods, we estimated a two-parameter IRT model in the full 16-item WHODAS scale (Table 2). The correlations between item residuals from this model were generally small (i.e. less than 0.20), which indicated minimal local dependence among items. Two exceptions were a correlation of 0.31 between the residuals of Item 2 ('Concentrating on something important') and Item 9 ('Thinking clearly') and a correlation of 0.28 between the residuals of Item 15 ('Controlling your emotions when you were around people') and Item 2.

The item discrimination parameters in this model are all greater than 1.0, indicating that all the WHODAS items have good discrimination between higher and lower disability at their inflection points. This being the case, we focused next on the location of the item's inflection points. The inflection point associated with the top five percentiles of a standardized normal distribution is in the range 1.5–1.6, so we sought items that had at least one threshold as close as possible to this range (i.e. in the range 1.2–2.0). All but one of the 16 items (difficulty understanding) met that criterion. Finally, we looked for items that had consistent IRT parameters across socio-demographic sub-samples. All 16 items exhibited evidence of significant socio-demographic variation in odds ratios (ORs) across at least one socio-demographic contrast. We decided to retain any item with less than three statistically significant differences, given that most of these were generally small effect sizes (i.e. OR ranged between 0.50 and 2.00). Of the items with four statistically significant differences, we then re-estimated several nested models by dropping each of the items and inspecting difficulty, discrimination, DIF, and the total scale information function. We also evaluated Item 9 ('Thinking clearly') and Item 15 ('Controlling your emotions when you were around people') because they had the highest residual correlations with Item 2 ('Concentrating or doing something important when other things were going on'). Through these comparisons, we selected the first eight items in Table 2 as those with optimal reproducibility out of the full 16-item scale.

Full scoring for the eight-item reduced WHODAS

The psychometric properties of the reduced eight-item scale are highly consistent with those of the same items in the original scale (Table 3). However, significant DIF

remained in the reduced scale, with the main difficulties associated with differences by gender and race/ethnicity. The distribution of the information in the TIC across the range of disability is similar for the reduced scale compared with the full scale, although the reduced scale inevitably provides less information than the full scale (Figure 1). The loss in information, though, described by the height of the peak, is only about 25% in the optimal information range despite the fact that the eight-item scale has only half as many items as the 16-item scale. Scale scores for individuals on the two scales are also very highly correlated (Pearson's correlation = 0.966, $P < 0.001$), suggesting that as a practical matter the loss of information introduced by reducing the scale from 16 to eight items is relatively modest.

Reduced scoring for eight-item scale

Finally, we compared the psychometric properties of the reduced eight-item scale using the four-category response options with those of a version of the eight-item scale that truncated the response options into a dichotomous distinction between a response of 'severe' difficulty and all other responses (i.e. moderate, mild, and no difficulty). In comparing the full (Table 3) and reduced (Table 4) scoring options, the discrimination parameters were predictably higher when we used reduced scoring. This finding suggests that the truncated response options effectively differentiate in the range we want; that is, between high severity and all lower levels of severity on the latent trait. Further, the severity estimates were also higher on the latent trait, given that the scale is capturing only those with severe levels of impairment. The amount of DIF observed in the reduced eight-item scale was significantly greater than the full scoring scale. For instance, many of the ORs were less than 0.50 or more than 2.00 compared with the more modest ORs in the eight-item full scoring. DIF appeared to be concentrated in the college graduate population, relative to those with a high school education or less. Gender was also a significant source of DIF for about half of the items.

Nonetheless, the eight-item reduced scale with binary scoring for the response options was significantly correlated with the full scoring eight-item scale (Pearson's correlation = 0.69, $P < 0.001$), and the 16-item scale with full scoring was correlated with the eight-item reduced scoring scale (Pearson's correlation = 0.677, $P < 0.001$). Inspection of the scale TIC (Figure 1) shows that scale information in the eight-item scale with the truncated response options is more highly skewed to the higher end of the disability continuum and actually has more information

Table 2 IRT 2-PL and differential item functioning for 16-item WHO-DAS (full scoring)

Item	Item parameters						Differential item functioning (DIF)					
	a	b1	b2	b3	Male OR	Black OR	Hisp OR	Other OR	HSG rad OR	College Grad OR	Age 26–49 OR	Age 50+ OR
1: REMBR	1.21	–0.07	1.47	2.69	0.74	0.72	1.17	1.31	0.99	0.82	1.11	1.12
2: CONC	1.37	–0.13	1.37	2.55	0.67	0.73	1.27	1.20	1.16	1.21	1.05	0.79
3: GO	1.22	1.22	2.11	2.84	0.68	1.02	1.52	2.01	0.76	0.58	1.15	1.14
4: NOTKNOW	1.22	0.56	1.68	2.59	0.91	0.74	0.84	1.43	0.78	0.63	0.88	0.55
5: SOCIAL	1.30	0.25	1.32	2.16	0.66	0.64	0.82	0.94	1.19	1.18	1.37	1.16
6: HOUSE	1.48	0.27	1.34	2.24	0.66	0.61	0.80	1.02	1.20	1.21	2.11	1.19
7: DAILY	1.74	0.45	1.54	2.37	0.86	0.72	0.86	1.42	1.01	1.28	0.99	0.42
8: WORK	1.58	0.09	1.40	2.38	0.67	0.57	0.65	1.17	0.95	1.28	1.30	0.97
9: THINK	1.20	0.14	1.47	2.74	0.66	0.83	1.08	1.24	0.88	0.81	1.29	1.69
10: UNDERSTAND	1.11	1.04	2.22	3.42	1.17	1.53	2.68	1.99	0.72	0.45	0.98	0.91
11: MONEY	1.01	0.55	1.64	2.65	1.09	1.31	1.33	1.73	1.27	1.02	1.13	0.70
12: LEARN	1.20	0.50	1.70	2.71	1.03	0.71	1.27	2.25	1.15	0.95	1.54	1.79
13: DRESS	1.28	1.64	2.52	3.32	0.88	1.42	1.71	2.84	0.67	0.48	1.28	0.97
14: GETTINGALNG	1.28	0.27	1.62	2.70	0.70	0.67	0.53	1.02	1.10	1.32	0.99	0.37
15: EMOTIONS	1.25	0.38	1.75	2.87	0.53	0.72	0.75	1.17	1.08	1.02	1.13	0.62
16: RELAX	1.13	0.22	1.53	2.70	0.76	0.72	0.61	0.89	1.17	1.34	1.61	0.81

Estimates adjusted for sampling design of NSDUH.

Note: 16-item and eight-item full scoring estimated using four-item polytomous IRT. Eight-item reduced scoring estimated using 2-PL IRT.

Full scoring includes 4 response options. Reduced scoring coded 1 = Severe difficulty versus all others.

a = discrimination, b = difficulty for endorsing response option k or higher. Response option 4 dropped.

categories: 1 = no difficulty; 2 = mild difficulty; 3 = moderate difficulty; 4 = severe difficulty (reference).

O.R. = Odds ratio (reference categories are female, white, <high school grad, and age 18–21).

Items in **BOLD** indicate significant item differential.

Table 3 IRT 2-PL and differential item functioning for eight-item WHO-DAS (full scoring)

Item	Item parameters						Differential item functioning (DIF)					
	a	b1	b2	b3	Male OR	Black OR	Hisp OR	Other OR	HS Grad OR	College Grad OR	Age 26–49 OR	Age 50+ OR
1: REMBR	1.10	−0.08	1.53	2.82	0.75	0.76	1.14	1.28	1.00	0.82	1.12	1.11
2: CONC	1.27	−0.14	1.48	2.61	0.70	0.76	1.19	1.15	1.13	1.15	1.06	0.81
3: GO	1.20	1.23	2.13	2.87	0.70	0.97	1.42	1.96	0.77	0.58	1.15	1.13
4: NOTKNOW	1.17	0.56	1.72	2.65	0.91	0.72	0.83	1.38	0.77	0.62	0.88	0.57
5: SOCIAL	1.23	0.25	1.34	2.22	0.67	0.64	0.80	0.95	1.20	1.19	1.36	1.16
6: HOUSE	1.56	0.26	1.33	2.21	0.65	0.60	0.77	1.01	1.16	1.22	2.21	1.21
7: DAILY	1.95	0.44	1.50	2.00	0.81	0.66	0.81	1.44	1.01	1.33	0.99	0.38
8: WORK	1.86	0.08	1.34	2.29	0.62	0.50	0.58	1.18	0.96	1.39	1.38	0.99

Estimates adjusted for sampling design of NSDUH.

Note: 16-item and eight-item full scoring estimated using four-item polytomous IRT. Eight-item reduced scoring estimated using 2-PL IRT.

Full scoring includes four response options. Reduced scoring coded 1 = Severe difficulty versus all others.

a = discrimination, b = difficulty for endorsing response option k or higher. Response option 4 dropped.

categories: 1 = no difficulty; 2 = mild difficulty; 3 = moderate difficulty; 4 = severe difficulty (reference).

O.R. = Odds ratio (reference categories are female, white, <high school grad, and age 18–21).

Items in **BOLD** indicate statistically significant item differential ($P < 0.05$).

Table 4 IRT 2-PL and differential item functioning for 8-item WHO-DAS (reduced scoring)

Item	Parameters									
	a	b1	Male OR	Black OR	Hisp OR	Other OR	HS Grad OR	College Grad OR	Age 26–49 OR	Age 50+ OR
1: REMBR	1.70	2.37	0.46	1.09	2.57	4.63	0.73	0.44	0.79	0.62
2: CONC	1.77	2.32	0.30	1.41	2.01	3.63	0.70	0.54	0.87	0.63
3: GO	1.69	2.50	0.61	2.37	3.33	4.46	0.32	0.18	1.18	2.23
4: NOTKNOW	1.85	2.25	0.76	1.60	1.78	2.60	0.62	0.30	0.81	0.64
5: SOCIAL	1.56	2.04	0.63	0.98	1.19	1.39	0.79	0.59	1.27	1.21
6: HOUSE	1.71	2.16	0.44	1.00	1.27	1.55	0.91	0.55	1.78	1.34
7: DAILY	2.27	2.37	0.76	2.04	2.74	2.60	0.86	0.67	0.77	0.83
8: WORK	2.28	2.18	0.65	1.27	1.96	2.43	0.53	0.52	1.14	1.75

Estimates adjusted for sampling design of NSDUH.

Note: 16 and 8 item full scoring estimated using 4 item polytomous IRT. 8 item reduced scoring estimated using 2-PL IRT.

Full scoring includes four response options. Reduced scoring coded 1 = Severe difficulty versus all others.

a = discrimination, b = difficulty for endorsing response option k or higher. Response option 4 dropped.

categories: 1 = no difficulty; 2 = mild difficulty; 3 = moderate difficulty; 4 = severe difficulty (reference).

O.R. = Odds ratio (reference categories are female, white, <high school grad, and age 18–21).

Items in **BOLD** indicate significant item differential.

Figure 1 Total information curves for three WHODAS scales.

than the other scales at true score values greater than 2.0. However, this is a higher threshold than we need to distinguish GAF of greater than or equal to 50. As a result, this scoring approach is suboptimal for our purposes. However, the fact that the truncated scale out-performs even the 16-item scale at the top end of the scale suggests that some intermediate scoring rule (e.g. combining 'severe' and 'moderate' difficulties into a single category versus all lower levels of difficulty or perhaps distinguishing 'severe' and 'moderate' as two separate thresholds but not distinguishing any lower levels of difficulty, resulting in a three-category response scale) might result in a reduced scale to capture all, or perhaps even more, of the information in the full 16-item scale and the eight-item scale with full scoring in the severity range of most importance in distinguishing respondents with GAF scores of 50 or less from those with GAF scores greater than 50.

Discussion

This investigation identified eight items that captured much of the same information (Pearson correlation = 0.97) as the full 16-item WHODAS scale using data in the NSDUH. Our finding that observed differential item function was more pronounced in the eight-item scale than the full 16-item scale is consistent with a study by Rehm *et al.* (1999), although that earlier study primarily found differential item functioning across countries rather than within a single country. Much of the DIF

observed in this study was concentrated in gender and race. The presence of DIF within a large study conducted within a single country points to a greater need to understand the underlying latent trait of disability and how it varies across important segments of the populations (e.g. race, gender, age). Yet, it is also important to note that significant DIF does not imply a poorly measured latent trait in the overall population. Rather, it suggests that sub-analyses are needed, perhaps by race and gender as indicated by findings from this study, to calibrate comparisons appropriately to determine impairment.

There are several considerations that should be kept in mind when evaluating this work. First, the analyses described here sought to explore whether it was feasible to generate an abbreviated version of the WHODAS for use in the NSDUH. Given that we did not compare specific versions of the WHODAS, based on length and administration (self-administered versus interviewer-administered methods); we are unable to comment on the comparability of our findings to other validation studies of the WHODAS. Second, the period of the WHODAS in this study was limited to the month in the past year in which the respondents' mental health problems caused the most serious impairments. This extended time frame may exhibit greater bias than the 30-day window that is used on most other WHODAS studies, as more error might be introduced in recall errors. However, it is also possible that increases in the valence of impairment for an individual increase the likelihood of respondent recall. Third, the evidence found here of differential item functioning does not mean that the scale is less capable of distinguishing between GAF of 0–50 and 51–100 in some sub-samples than others, as DIF was evaluated across the entire scale range. It might mean, though, that the optimal scoring rules for defining SMI with the K6 in the NSDUH will use different cut-points for women than men or for non-Hispanic Whites than non-Whites. These are issues, though, that can only be adjudicated by studying the joint predictive associations of K6 scores and WHODAS scores in predicting clinical diagnoses of SMI, which is a line of analysis that goes beyond the scope of the current report. A companion paper by Aldworth *et al.* (2010) in this issue presents results of such an analysis. The critical finding of the current report for purposes of the Aldworth *et al.* analyses is that the reduced eight-item version of WHODAS has a Pearson correlation of 0.97 with the full 16-item WHODAS, justifying the use of the shorter version of the scale in those analyses. Because of this extremely high correlation, we can be assured that the extent to which interactions of WHODAS scores with gender or race/ethnicity are needed to optimize

prediction of SMI from joint K6 and WHODAS scores can be studied with as much precision using the eight-item version of the WHODAS as the 16-item version.

Acknowledgments

This work was prepared by the Division of Population Surveys, Office of Applied Studies, Substance Abuse and Mental Health Services Administration, US Department of Health and Human Services, and by RTI International (a trade name of Research Triangle Institute). Work by RTI was performed under Contract No. 283-2004-00022.

Declaration of interest

The authors have no competing interests.

References

- Alcohol Drug Abuse and Mental Health Administration. *Alcohol, Drug Abuse, and Mental Health Administration (ADAMHA) Reorganization Act, Public Law. 102-321, S. 1306*, 1992.
- Aldworth J., Colpe L.J., Gfroerer J.C., Novak S.P., Chromy J.R., Barker P.G., Barnett-Walker K., Karg R.S., Morton K.B., Spagnola K. (2010). National Survey on Drug Use and Health's Mental Health Surveillance Study: calibration analysis. *Int J Methods Psychiatr Res*, **19**(Suppl. 1), 61–87.
- Colpe L.J., Barker P.G., Karg R.S., Batts K.R., Morton K.B., Gfroerer J.C., Stolzenberg S.J., Cunningham D.B., First M.B., Aldworth J. (2010). The National Survey on Drug Use and Health Mental Health Surveillance Study: calibration study design and field procedures. *Int J Methods Psychiatr Res*, **19**(Suppl. 1), 36–48.
- Colpe L.J., Epstein J.F., Barker P.R., Gfroerer J.C. (2009). Screening for serious mental illness in the National Survey on Drug Use and Health (NSDUH). *Ann Epidemiol*, **19**, 210–211, DOI: 10.1016/j.annepidem.2008.09.005
- Emberson S.E., Reise S.P. (2000). *Item Response Theory for Psychologists*, Lawrence Earlbaum Associates.
- Furukawa T.A., Kessler R.C., Slade T., Andrews G. (2003). The performance of the K6 and K10 screening scales for psychological distress in the Australian National Survey of Mental Health and Well-Being. *Psychol Med*, **33**, 357–362, DOI: 10.1017/S0033291702006700
- Grant B.F., Kaplan K., Shepard J., Moore T.C. (2003). *Source and Accuracy Statement for Wave 1 of the 2001–2002 National Epidemiologic Survey on Alcohol and Related Conditions*, National Institute on Alcohol Abuse & Alcoholism.
- Kessler R.C., Andrews G., Mroczek D., Üstün T.B., Wittchen H.-U. (1998). The World Health Organization Composite International Diagnostic Interview Short Form (CIDI-SF). *Int J Methods Psychiatr Res*, **7**, 171–185, DOI: 10.1002/mpr.47
- Kessler R.C., Barker P.R., Colpe L.J., Epstein J.F., Gfroerer J.C., Hiripi E., Howes M.J., Normand S.L., Manderscheid R.W., Walters E.E., Zaslavsky A.M. (2003a). Screening for serious mental illness in the general population. *Arch Gen Psychiatry*, **60**, 184–189.
- Kessler R.C., McGonagle K.A., Zhao S., Nelson C.B., Hughes M., Eshleman S., Wittchen H.U., Kendler K.S. (1994). Lifetime and 12-month prevalence of DSM-III-R psychiatric disorders in the United States. Results from the National Comorbidity Survey. *Arch Gen Psychiatry*, **51**, 8–19.
- Kessler R.C., Merikangas K.R. (2004). The National Comorbidity Survey Replication (NCS-R): background and aims. *Int J Methods Psychiatr Res*, **13**, 60–68, DOI: 10.1002/mpr.166
- Kessler R.C., Merikangas K.R., Berglund P., Eaton W.W., Koretz D.S., Walters E.E. (2003b). Mild disorders should not be eliminated from the DSM-V. *Arch Gen Psychiatry*, **60**, 1117–1122.
- Manderscheid R., Kessler R.C., Gfroerer J., Larson S. (2005). The field must capitalize on prevalence data. *Behav Healthcare Tomorr*, **14**, 7–8.
- Muthén L.K., Muthén B.O. (1998–2007). *Mplus User's Guide*. fifth Edition, Muthén & Muthén.
- Narrow W.E., Rae D.S., Robins L.N., Regier D.A. (2002). Revised prevalence estimates of mental disorders in the United States: using a clinical significance criterion to reconcile 2 surveys' estimates. *Arch Gen Psychiatry*, **59**, 115–123.
- Reeve B.B., Masse L.C. (2004). Item response theory modeling for questionnaire evaluation. In: *Methods for Testing and Evaluation Survey Questionnaires* (eds Presser S, Rothgeb JM, Couper MP, Lessler JT, Martin E, Martin J, Sinter E), pp. 247–273, Wiley.
- Regier D.A., Kaelber C.T., Rae D.S., Farmer M.E., Knauper B., Kessler R.C., Norquist G.S. (1998). Limitations of diagnostic criteria and assessment instruments for mental disorders. Implications for research and policy. *Arch Gen Psychiatry*, **55**, 109–115.
- Rehm J., Üstün T.B., Saxena S., Nelson C.B., Chatterji S., Ivis F., Adlaf E. (1999). On the development and psychometric testing of the WHO screening instrument to assess disablement in the general population. *Int J Methods Psychiatr Res*, **8**, 110–123, DOI: 10.1002/mpr.61
- Research Triangle Institute. (2006). *SUDAAN User's Manual, Release 9.0*, Research Triangle Institute.
- Robins L.N., Regier D.A. *et al.* (1991). *Psychiatric Disorders in America: The Epidemiologic Catchment Area Study*. New York: The Free Press.
- Samejima F. (1997). The graded response model. In: *Handbook of Modern Item Response Theory* (eds van der Linden WJ, Hambleton RK), pp. 85–100, Springer.
- Steinberg L., Thissen D. (1995). Item response theory methods in personality research. In: *Personality Research, Methods, and Theory: A Festschrift Honoring Donald W*

- Fiske (eds Shrout PT, Fiske ST), pp. 161–181, Lawrence Erlbaum Associates.
- Substance Abuse and Mental Health Services Administration. (1993). Final notice [Final definitions for: (1) Children with a serious emotional disturbance, and (2) adults with a serious mental illness]. *Fed Register*, **58**, 29422–29425.
- Substance Abuse and Mental Health Services Administration. *Results from the 2002 National Survey on Drug Use and Health: National Findings* (Office of Applied Studies, NHSDA Series H-22, DHHS Publication No. SMA 03–3836). Rockville, MD, 2003.
- Substance Abuse and Mental Health Services Administration. *Results from the 2004 National Survey on Drug Use and Health: National Findings* (NSDUH Series H-28, DHHS Publication No. SMA 05-4062). Washington, DC: Substance Abuse and Mental Health Services Administration, 2005.