In the Field With the Beck Depression Inventory

Julie Barroso
Margarete Sandelowski

Qualitative methods are typically and formally used only in the earliest phases of instrument development to generate items. Once these items are generated, instrument development usually then moves into the testing phases, where quantitative methods prevail. The achievement of psychometric credibility is presumed to depend largely on quantitative measures of reliability and validity. Or if qualitative methods are employed, their use is masked, unfocused, and/or unplanned. The planned use of qualitative methods is critical in every phase of instrument use and in all studies that depend for their results on instruments, and their use is critical in illuminating problems with existing instruments. The authors illustrate these points by drawing on the first author's experiences in the field with the Beck Depression Inventory in her research program on managing fatigue in persons with HIV/AIDS.

Although qualitative research has become increasingly popular in the practice disciplines, it remains underused as a means to enhance and even to "salvage" (Weinholtz, Kacer, & Rocklin, 1995) quantitative research. Indeed, qualitative research is more often still seen to require salvaging by quantitative research. Sutton (1997) even suggested that researchers might mask their use of qualitative research to improve their chances for publication.

An especially good illustration of the underuse of qualitative research is in the domain of instrument development (including initial creation and successive refinement and evaluation) and use. If qualitative methods are used at all, they are typically and formally used only in the earliest phases of instrument development to generate items. Qualitative methods are seen here to comprise preliminary means to enhance the content validity of these items. An instrument is said to have content validity when its items adequately represent a target content domain (Waltz, Strickland, & Lenz, 1991) such as anxiety (in an instrument designed to measure anxiety) and depression (in an instrument designed to measure depression). Open-ended, relatively unstructured interviews and a range of qualitative analysis techniques (including qualitative content, phenomenological thematic, or constant comparison analyses) are used to generate these items. Instrument development then usually moves into the testing phases, where quantitative methods prevail as the achievement of psychometric credibility is presumed to depend largely on quantitative measures of reliability and validity. If qualitative methods are

AUTHORS' NOTE: The HIV-related fatigue study was funded by two grants from the University of North Carolina at Chapel Hill, School of Nursing: A Faculty Research Opportunity Grant and a Summer Research Award.

QUALITATIVE HEALTH RESEARCH, Vol. 11 No. 4, July 2001 491-504

employed at all, their use is masked, unfocused, and/or unplanned. For example, Waltz et al. (1991) advised researchers to pay attention to participants' reactions and comments while responding to an instrument but not to describe the systematic and ongoing use of qualitative methods to appraise and refine instruments.

Yet, evaluating participants' reactions to and comments about instruments in the field ought to entail something more deliberate than merely paying attention. The planned use of qualitative methods is critical in every phase of instrument development and use and in all studies that depend for their results on instruments. The purposeful use of qualitative methods—entailing a clearly articulated sampling strategy and plans for the collection and analysis of data—is critical in illuminating problems with existing instruments that are unique to the varying circumstances in which they are used. We illustrate these points in this article by drawing on the first author's experiences in the field with the well-known and frequently used Beck Depression Inventory (BDI) in her research program on managing fatigue in persons with HIV/AIDS. She encountered problems with the instrument in the field and during data analysis that raised questions about its validity and utility. These problems drew our attention to the importance of qualitative methods not only in the earliest stages of instrument development but also for the ongoing evaluation of instruments in the field. In this article, we describe these problems as well as their research and clinical significance, highlighting how ongoing qualitative inquiry can be used to inform quantitative psychometric evaluations and the appropriate use of instruments. To set the scene for our discussion, we begin with a brief overview of fatigue in persons with HIV. We refer to the first author's research as the HIV-related fatigue study. We use *I* to refer to the first author and *we* to refer to both authors.

THE PROBLEM OF FATIGUE IN HIV-INFECTED PERSONS

Although mortality rates from HIV infection have dropped dramatically over the past 5 years, persons with this infection must still contend with symptoms that seriously interfere with their ability to lead full and productive lives. One of these symptoms is fatigue.

Fatigue is the most frequent and debilitating complaint of HIV-positive people, with an incidence ranging from 20% to 60% (Breitbart, McDonald, Rosenfeld, Monkman, & Passik, 1998; Cunningham et al., 1998; Walker, McGown, Jantos, & Anson, 1997). HIV-related fatigue is associated with significantly poorer physical functioning and is a strong predictor of both limitations of daily activity and disability days (Breitbart et al., 1998; Cleary et al., 1993; O'Dell, Hubert, Lubeck, & O'Driscoll, 1998). Fleishman and Crystal (1998) found that people with HIV infection exhibited increasing limitations in vigorous activities over a 1-year period. Declines in physical function were related to increasing disease severity, to poor self-rated health, and most important for the purpose of this article, to prior reports of fatigue. Several studies have found no correlation between fatigue and disease severity (Breitbart et al., 1998; De Boer, van Dam, Sprangers, Frissen, & Lange, 1993; Perkins et al., 1995).

Emphasizing the high rate of depression among people seropositive for HIV, some researchers have speculated that the fatigue also exhibited in these persons is psychological. For example, Perkins and her colleagues (1995) found that increased levels of fatigue were associated with increased depression and dysphoric mood but not associated with HIV disease progression. They therefore concluded that fatigue in otherwise asymptomatic persons was probably related to psychological rather than to physiological disturbances. Walker and colleagues (1997) also found a strong positive correlation between fatigue and depression in HIV-positive persons. Breitbart et al. (1998) concluded that HIV-positive patients with fatigue had significantly greater psychological distress, significantly more depressive symptoms, and greater hopelessness than those without fatigue. Yet they too surmised that fatigue was a symptom of depression in some but not all fatigued patients with AIDS and, therefore, that fatigue was not merely a reflection of underlying depression.

Indeed, sorting out the relationship between fatigue and depression has been difficult (Hoover et al., 1993; Walker et al., 1997). Depression may cause fatigue, but fatigue may also cause depression (Wessely, Hotopf, & Sharpe, 1998). Complicating the problem is the fact that many of the tools used to measure depression include somatic complaints that are similar to the symptoms of fatigue. The somatic symptoms of depression included in the BDI have been found to be closely associated with several AIDS-related diagnoses (Kalichman, Rompa, & Cage, 2000; Kalichman, Sikkema, & Somlai, 1995). These researchers recommended that persons who experience HIV-related symptoms be appraised for depression using instruments that do not include somatic symptoms.

THE BDI

The BDI is one of the most commonly used tools in research. The BDI has been used in more than 2,000 studies and translated into many languages (Richter, Werner, Heerlein, Kraus, & Sauer, 1998), and it is the most commonly used measure of depression in studies involving subjects with HIV. The current form of the tool is the BDI-II,[1] a 21-item, self-report instrument for measuring the severity of depression in adults and adolescents age 13 and older (Beck, Steer, & Brown, 1996). Richter and his colleagues (1998) listed as advantages of the BDI its high internal consistency, high content validity, its specificity in differentiating between depressed and nondepressed subjects, its sensitivity to change, and its international popularity.

The BDI-II was developed for the assessment of symptoms corresponding to criteria for diagnosing depressive disorders listed in the *Diagnostic and Statistical Manual of Mental Disorders*, 4th edition (*DSM-IV*) (American Psychiatric Association, 1994). For the diagnosis to be made, five (or more) of the symptoms listed in Table 1 have to be present during the same 2-week period and to represent a change from previous functioning; at least one of the symptoms has to be either depressed mood or loss of interest or pleasure. As shown in Table 1, virtually all of these criteria are based on self-report. Only one of them is based on observations by others. Indeed, researchers outside of the psychiatric/mental health field will most likely not be familiar with the *DSM-IV* criteria for depression, relying instead on well-known instruments used to measure depression, such as the BDI.

This version of the BDI-II is based on typical descriptive statements regarding symptoms that depressed psychiatric patients had frequently reported. Instrument items were not selected to reflect any particular theory of depression (Beck et al., 1996). Indeed, no definition of depression is included in the BDI-II manual. Each

TABLE 1: Symptoms Captured in the Beck Depression Inventory

1. Depressed mood most of the day, nearly every day, as indicated by either subjective report (e.g., feels sad or empty) or observation made by others (e.g., appears tearful).
2. Markedly diminished interest or pleasure in all, or almost all, activities most of the day, nearly every day (as indicated by either subjective account or observation made by others).
3. Significant weight loss when not dieting or weight gain (e.g., a change of more than 5% of body weight in a month) or decrease or increase in appetite nearly every day.
4. Insomnia or hypersomnia nearly every day.
5. Psychomotor agitation or retardation nearly every day (observable by others, not merely subjective feelings of restlessness or being slowed down).
6. Fatigue or loss of energy nearly every day.
7. Feelings of worthlessness or excessive or inappropriate guilt (which may be delusional) nearly every day (not merely self-reproach or guilt about being sick).
8. Diminished ability to think or concentrate, or indecisiveness, nearly every day (either by subjective account or as observed by others).
9. Recurrent thoughts of death (not just fear of dying), recurrent suicidal ideation without a specific plan, or a suicide attempt or a specific plan for committing suicide.

SOURCE: American Psychiatric Association (1994).

TABLE 2: Scoring the Beck Depression Inventory

Total Score	*Depression*
0-13	Minimal
14-19	Mild
20-28	Moderate
29-63	Severe

item is rated on a 4-point scale, ranging from 0 to 3 to indicate severity. The total score for an individual can range from 0 to 63. The last revision of the BDI, completed in 1979, resulted in the BDI-IA. Moran and Lambert (1983) compared the content of this version of the BDI against the *DSM-III* affective disorders criteria and found that the BDI adequately met only six of the nine criteria. In addition, depression is now diagnosed and treated at an earlier stage. These factors led to the revision that is now the BDI-II. Four items, including weight loss, body image change, somatic preoccupation, and work difficulty, were dropped and replaced by four new items, including agitation, worthlessness, concentration difficulty, and loss of energy. This change was made to capture symptoms typical of severe depression. Two items were changed to allow for increases as well as decreases in appetite and sleep (Beck et al., 1996).

Table 2 shows how scores are now interpreted. Beck and his colleagues (1996) noted that users of the BDI could vary the ranges somewhat, depending on what was more important to them: either sensitivity (or the probability of an instrument correctly classifying a respondent as depressed) or specificity (or the probability of an instrument correctly classifying a respondent as not depressed). In general, sensitivity is more important than specificity, meaning that users would likely rather detect persons who are depressed, even if that means falsely identifying as depressed some persons who are not depressed (Beck et al., 1996). A score of 14 to 19 indicates mild depression, 20 to 28 indicates moderate depression, and 29 to 63 indicates severe depression.

The BDI is a norm-referenced measure. Beck and his colleagues (1996) used samples from four different psychiatric outpatient clinics and one college student sample to establish its norms. All four of the outpatient samples were from urban or suburban locations. Of 500 participants, 63% were women, and 91% were Caucasian. Coefficient alpha (a measure of the internal consistency of an instrument) for the BDI was high: 0.92 in the outpatients and 0.93 in the college students. Test-retest reliability was 0.93 when the tool was readministered after 1 week. Beck and his colleagues provided evidence of both convergent (with other depression instruments) and discriminant (relative to instruments measuring other potentially related concepts, such as anxiety) validity.

THE HIV-RELATED FATIGUE STUDY

Knowing from clinical practice how serious fatigue can be to HIV-infected persons and also being aware of the lack of clarity about this symptom, I decided to direct my research program toward the study of fatigue to determine its intensity, characteristics, and patterns as a basis for developing interventions to minimize fatigue (Barroso, 2000). These data were not in the published research about HIV-related fatigue. Researchers had conducted correlational and intervention studies without having an answer to the basic question: What is HIV-related fatigue? I sought to answer that question through a qualitative descriptive study using open-ended and largely unstructured interviews—with people who were experiencing HIV-related fatigue—and qualitative content analysis (Morgan, 1993). To have a fuller description of the target phenomenon, I included standardized measurements of fatigue and depression. I chose the General Fatigue Scale (GFS) (Meek & Nail, 1997) as a measure of fatigue intensity and the BDI (Beck et al., 1996), which is the most widely used standardized measure of depression. The HIV-related fatigue study was a pilot study in preparation for a program of research in this area. Institutional Review Board approval for the protection of human subjects was obtained.

Thirty-one people with HIV-related fatigue participated in the HIV-related fatigue study. They called me in response to flyers I had distributed at health care facilities and community-based organizations for people with HIV disease. To determine potential participants' eligibility for the study, I asked them briefly about their experience with fatigue since becoming HIV-positive. If they stated that they were more fatigued than they were before becoming infected and that it was interfering with some aspect of their lives, I invited them to participate in the study. The sample included 1 Caucasian female, 9 African American females, 7 Caucasian males, 13 African American males, and 1 African male from Uganda. At the time of interview, these participants were an average of 4.74 years since diagnosis. Their median income was $495 a month, and 84% were unemployed. Many of them were no longer able to work, in part because of the fatigue they were experiencing. As a group, the participants were relatively healthy immunologically, with a self-reported mean CD4 count of 575 and mean HIV viral load of 147. (A normal CD4, or T helper cell count, can range from 500 to 1200. HIV-RNA viral load, a measure of the amount of virus in the blood, can range from undetectable to in the millions.) Thirty-one participants were recruited from two large university infectious disease clinics, a county health department early intervention program for persons with

HIV, and an AIDS service agency. All of the recruitment sites were in a southeastern state with a large rural population.

I asked a broad question about HIV-related fatigue—"Describe your fatigue for me"—and used prompts that would allow me to obtain an in-depth description of the symptom of fatigue. I also asked about experiences that might be relevant to fatigue, such as depression and problems with sleeping. I then administered the GFS and the BDI. I chose this order because I wanted the participants to determine the content of the interviews and did not want the instruments to sensitize the participants to any particular content. The interviews were audiotaped and then transcribed verbatim.

On the GFS, 25 of the 31 participants scored above the mean score of cancer patients (most of them breast cancer patients) undergoing treatment and at their most fatigued (Meek & Nail, 1997) or at a level of fatigue that significantly interferes with daily functioning. On the BDI-II, 13 of the 31 participants scored in the moderately to severely depressed range, which is a score of ≥ 20. If a score of ≥ 14 had been used, as recommended in the BDI-II manual as the cutoff score for mild depression, the number of participants scoring as depressed would have been much higher. However, I chose the cutoff score of 20 to maximize the sensitivity of the tool: to obtain as pure a group of persons with depression as possible with regard to symptomatology (Beck et al., 1996). As expected, GFS and BDI scores were positively correlated ($r = 0.46$, $p = 0.01$). Because the somatic items on the BDI-II are closely related to the symptoms of fatigue, I removed them for data analysis purposes after administering the test. With the somatic items removed, depression and fatigue remained significantly correlated ($r = 0.45$, $p = 0.03$). Controlling for the use of antidepressant drugs (10 of the participants were taking such drugs), there was still a significant correlation between fatigue and depression (partial $r = 0.42$, $p = 0.05$). For statistical purposes, the interview data were coded as yes/no, depending on participants' responses to the question, "Do you feel you are having a problem with depression?" asked toward the end of the interview. These data were analyzed by me and a graduate nursing student/clinician working as a research assistant on the project; the responses to this question were analyzed independently, then the analyses were compared. There was 100% agreement between us on the coding of the interview data regarding depression. No one in the study asked to have the term *depression* defined nor were any of the participants unable to answer the question about being depressed. Indeed, as described later in this article, many of the participants spoke eloquently on the subject of depression.

PSYCHOMETRIC PROBLEMS WITH THE BDI

Although psychometrically sound in many respects, researchers have found significant problems with the BDI. For example, the results of factor analysis show important differences between psychiatric outpatient and college samples. For the psychiatric outpatients, the factor analysis placed items into two categories: (a) somatic/affective and (b) cognitive. For the college students, the two categories were labeled cognitive/affective and somatic. There were some items that loaded on one factor in one group and on another in another group. For example, sadness loaded on the cognitive category for the psychiatric outpatients but in the cognitive/affective category for the college students. There were several items that did not load in either category; this was done using promax-rotated, iterated-principal-factor standardized regression coefficients (Beck et al., 1996). Other researchers, using a confirmatory factor analysis, found a three-factor oblique model consisting of negative attitude, performance difficulty, and somatic elements (Osman et al., 1997). More recent work supports a model in which the BDI-II reflects one underlying second-order dimension of self-reported depression composed of two first-order factors representing cognitive and noncognitive symptoms (Steer, Ball, Ranieri, & Beck, 1999). Richter et al. (1998) referred to this problem with factor analysis as one of the disadvantages of the BDI.

Another problem has been in the area of social desirability. Patients with a greater social desirability response bias may report less depression because they do not want to appear depressed (Deshields, Tait, Gfeller, & Chibnall, 1995; Langevin & Stancer, 1979). Langevin and Stancer (1979) contended that depression scales such as the BDI really measure self-esteem, not depression. Other researchers have described subjects who fit a "fake-good" profile. Beebe, Finer, and Holmbeck (1996) proposed a low-end specificity problem whereby high scorers on measures such as the BDI tended to be broadly distressed rather than specifically depressed and very low scorers were likely denying depressive symptomatology. These researchers found that college students who scored less than 9 on the BDI, and particularly those scoring 0 to 1, showed a fake-good profile on the Minnesota Multiphasic Personality Inventory (MMPI). According to these researchers, the MMPI fake-good profile is suggestive of symptom denial coupled with a naive assertion of personal virtue and psychological resourcefulness. They therefore interpreted these findings as evidence that the BDI has poor low-end specificity, meaning that using the BDI to classify persons may result in some dysphoric patients being erroneously classified as nondysphoric. If a disproportionate number of very low scorers are faking good, then any findings from the use of such extreme groups may be due more to a fake-good response style than to depression per se.

The BDI is seen to be especially vulnerable to faking because of its high face validity and the likelihood of a response set. Dahlstrom, Brooks, and Peterson (1990) argued that the lack of subtle content in the item groupings of the BDI, in addition to the consistency in the ordering of these items from least to most pathological, make this instrument unduly susceptible to either defensive or malingering response sets. After subjects have examined the first two or three items, they can answer the remaining items without even reading through them. These researchers conducted a study in which the descriptors in each set were in random sequence. The random-order BDI resulted in a significantly higher depression score than did the original version.

THE BDI IN THE HIV-RELATED FATIGUE STUDY

My use of the BDI with HIV-positive patients also raised questions about the validity of this instrument. To begin, many of the participants were puzzled by the options for many of the items. If the items were read to them (as some participants preferred),[2] I was often asked to repeat what I had said. Or if participants were reading the items themselves, they often reread them repeatedly and out loud. Most of the items contain some combination of positively and negatively worded statements, which seemed to interfere with participants' comprehension of the options.

Researchers have noted item difficulty as one of the disadvantages of the BDI (Faravelli, Albanesi, & Poli, 1986; Richter et al., 1998).

For example, the choices offered on the item about worthlessness are as follows:

0. I do not feel I am worthless.
1. I don't consider myself as worthwhile and useful as I used to.
2. I feel more worthless as compared to other people.
3. I feel utterly worthless.

Participants would try to emphasize the key words in each item to determine which choice best represented how they felt. It was clear that they were not simply reading out loud but rather struggling to comprehend many of the items. After reading this item once, they would say things such as the following:

"OK, I do not feel I am worthless."
"I don't consider myself as *worthwhile* and useful as I used to."
"I feel more *worthless* as compared to other people. . . . wait a minute, OK . . . "
"I feel utterly *worthless*."

This struggle for comprehension was not evident in the completion of the GFS nor was it present in the interviews.

Casting doubt on the internal consistency of the BDI were the multiple interpretations some items elicited. To function properly psychometrically, any one item can only mean one thing to anyone responding to that item at any one time. Although researchers generating items from qualitative data make every effort to preserve the meaning of that data as they transform them into instrument items (Fleury, 1993), they are engaging in a contradictory project. This transformation process requires that the meaningfulness of data be reduced. The qualitative project of extracting all the meanings from a portion of data is at odds with the quantitative project of "encapsulating . . . meaning" (Masse, 2000, p. 411). As Masse demonstrated, the "quest for meaning and the quest for measurement are incommensurable research objectives" (2000, p. 411). In the HIV-related fatigue study, one man noted that to respond to the item "I cry more than I used to" by saying yes would not mean that he was depressed. As he explained it, although he did cry more than he used to, it was not because he was depressed but rather because he had become more in touch with his feelings as a consequence of his encounter with HIV infection.

This man's comment also raised issues concerning the use of the BDI with people very different from those on whom it was normed: namely, psychiatric outpatients and college students. The people in the HIV-related fatigue study were largely members of minority groups, of a low socioeconomic status, and living in rural areas. They were of a very different culture from those on whom the BDI was normed. Even those who may have been more similar to the groups on which it was normed found themselves, once infected with HIV, part of a new culture: namely, a group of highly stigmatized, disenfranchised individuals. Richter et al. (1998) and the developers of the tool themselves (Beck, Steer, Ball, & Ranieri, 1996) viewed the homogeneity of the population on which the tool was normed as an important disadvantage of the BDI.

This disadvantage was evident in the HIV-related fatigue study. For example, most of the items on the BDI-II refer to how things "used to be"; however, the instructions advise the respondent to circle the statement that best describes how they have been feeling during the past 2 weeks, including today or the day on which they are responding to the test. For people with HIV infection, life is often divided into before and after infection and/or diagnosis with HIV; how things used to be is how they were before becoming infected or learning of their infection. Time is a complex phenomenon, and the way things used to be may not correspond to the past 2 weeks. In response to the question about depression in the HIV-related fatigue study, participants emphasized the frequency and duration of depression. The narrow temporal orientation of the BDI seemed incongruent with the lived temporal orientation of these persons living with HIV infection.

Moreover, items addressing guilt and punishment elicited responses different from those Beck and his colleagues likely intended. Persons with HIV disease understand that many people believe the disease itself to be a punishment for persons guilty of engaging in behaviors deemed immoral and/or designated as illegal, such as homosexuality and drug use. To live with the disease, affected persons have to manage these beliefs. Indeed, in the HIV-related fatigue study, only 8 participants responded to the third or fourth options on the guilt and punishment items (indicating that they felt guilty most or all of the time or that they expected to be punished or actually are being punished). I found, from my own clinical and research experience (Barroso, 1997; Barroso & Powell-Cope, 2000), that persons with HIV infection often used cognitive reframing to offset such negative feelings. Accordingly, the context for guilt and punishment in persons with a culturally stigmatizing disease such as HIV/AIDS is different from that for the depressed persons Beck envisioned. The items regarding guilt and punishment, in particular, may be misleading measures of depression in such persons. Researchers must have not only a clinical understanding of a disease but also a historical and cultural understanding of the multiple meanings diseases such as HIV/AIDS generate. Indeed, HIV/AIDS is itself a discourse as well as a medically defined disease (Treichler, 1990).

With regard to the issue of social desirability, it is possible that the participants did not tell the truth about being depressed. If participants can fake good or try to present themselves in a more appealing way on an instrument, they can also fake good in an interview. There is no method of data collection that is immune to faking good. Yet these participants were not likely faking good about depression. They had already been frank in admitting to having HIV/AIDS, using drugs, engaging in risky sexual practices, and/or being gay. Against these potentially stigmatizing admissions, depression is hardly an experience about which these participants were likely to lie.

Perhaps most important, what the HIV-related fatigue study participants conveyed in interview data was incongruent with their BDI scores. Table 3 shows this problem with convergent validity. Of 31 participants, 12 persons, or almost 39% of the sample, showed discrepant data. Although there is no one rule about what the level of agreement should be between two different sources of data about the same target phenomenon, a researcher might reasonably expect the congruence between interview and instrument responses here to be much higher. If the somatic items are deleted from the BDI—and there is precedent for doing this (Kalichman et al., 1995, 2000)—the discrepancy is even greater. Participants were asked, "Do you feel you are having a problem with depression?" All of the participants appeared to understand the question. Examples of some of the things those persons said who scored as *not* depressed on the BDI include the following:

TABLE 3: Congruence of Interview Data With Beck Depression Inventory (BDI) score

	Depressed on BDI	*Not depressed on BDI*
Depressed on interview	12	5
Not depressed on interview	7	7

> The depression is there since childhood. . . . When I'm analyzing myself, I say, yes, it probably has something to do with my homosexuality. Because I am a very flamboyant homosexual, and it's not like I can walk the street and people, say, don't know. They automatically know. I mean, I've had to live with that through the years, but I guess deep down inside it really probably does bother me. (African American man, living in HIV-designated housing)

> I feel depressed . . . a lot of times when I'm by myself, I just think about everything, about everybody in the house [referring to her husband and two children] being negative. And it's hard to sleep. (African American woman, living in an apartment complex with her husband and her two children, ages 13 and 3 months)

> Oh, most of the time I do [feel depressed]. Sometimes I just be crying. I just, ohhh, ohhh. (African American woman with a flat affect, living in a run-down, filthy boarding house in a high-crime area)

> I've been suffering depression since I was a child. It's something that's been plaguing me for years. (African American man living by himself in a small, tidy house)

> What makes me depressed? The whole bag. You know, the diagnosis, the being knocked out of all the—I mean, I can't travel, run the road like I used to. I mean, that's a 24-hour job. And animals—not being allowed to be around animals anymore. So really, all my forte was taken away from me. You know, I couldn't do those either one of these things if I wanted to. (Caucasian man, living in a duplex, who used to be a manager for a rock and roll band and who had also worked in a pet shop)

Finally, one man from Uganda, who scored as not depressed on the BDI, who spoke fluent English, and who showed no sign of not comprehending the items on that instrument, spoke at length about being profoundly depressed; he had lost his father and many friends in the civil war in his country. He showed me a picture of himself and his five closest friends, taken during the 1970s. The photo showed five young people dressed nicely and smiling for the camera. On the back, he had listed the years since the picture had been taken. Next to five of those years were the names of his friends, indicating the year each of them had died.

In a study of 192 HIV-seropositive patients, 95 scored ≥ 14 on the BDI, but only one third of these patients had a depressive disorder (Judd, Mijch, McCausland, & Cockram, 1997). The authors were unable to explain the reasons for this discrepancy. Finally, in their study of depression in the terminally ill, Chochinov, Wilson, Enns, and Lander (1997) found that the single question—"Are you depressed?"—correctly identified depressed and nondepressed patients in every case and outperformed the visual analog measures and questionnaires they had also used, including the BDI. These investigators had conducted semistructured diagnostic interviews for depression with 197 patients receiving palliative care for advanced cancer. The diagnoses from interviews served as the standards against which the screening performance of the four brief screening methods was assessed. They concluded that for diagnostic purposes, brief screening measures do not approach the validity of asking the single question, "Are you depressed?"

CONCLUSION

My and other researchers' experiences with the BDI show the importance of incorporating qualitative techniques of interviewing and observation in all phases of the process of instrument utilization. They also suggest the "ontological incommensurability" (Masse, 2000) of efforts both to preserve the fullness of and to reduce meaning. We have several recommendations that derive from this work.

The informational content and representational form of an instrument should always be appraised by persons sharing the same kind of life experience as the ones who will be evaluated using the instrument. (Some would argue that no life experiences are ever the same and thus against the ability of instruments to capture them.) This evaluation should occur not only in pilot studies but throughout the life of a project. Researchers can appraise the content validity and form of an instrument with a random purposeful sample of participants at purposefully selected times. A stratified purposeful sampling stategy may be useful for studies involving more heterogeneous samples, where such variables as gender, social class, or mode of infection—in the case of studies with persons with HIV—can be surmised to influence instrument reception. In addition, because they are most intimately involved with patients, a purposefully selected sample of clinicians is also vital to the development and selection of appropriate instruments for research and to the credible interpretation of participants' responses to them. Instrument validity is not just a matter of content, convergent, and/or construct validity but also of communicative, pragmatic, and ethical validity (Kvale, 1995; Maxwell, 1992). Qualitative research with carefully selected samples is uniquely suited to enhancing these validities.

Specifically, researchers should plan systematically to collect interview and observation data from participants on the content and form of every item in an instrument and their embodied response to it. Participants should be asked what meaning or meanings each item had for them, the degree of difficulty they had in ascertaining what each item was supposed to mean, and their reasons for responding to each item as they did. Researchers should also systematically collect observational data on such parameters as nonverbal behaviors (including facial expression and nonverbal vocalizations), the time it takes for participants to respond to each item, and the actual order in which they respond to them. Researchers must understand the judgments participants make and the behaviors they exhibit when they encounter items on an instrument. Interviewing participants about the items will also help to address social desirability issues in that researchers can better ascertain how individuals desire to represent themselves. Depression is both a "language and a lived experience" and therefore cannot be treated as an "empirical entity" outside of the "individual discourses" in which it appears (Masse, 2000, p. 413).

Indeed, researchers should explicitly attend to the historical and cultural context for the content domain they wish to evaluate and the biographical context for the persons who will be evaluated with that content domain. Instruments are ideally supposed to transcend biography, culture, and history, but like any other human fabrication, they cannot. Qualitative approaches that emphasize these larger contexts must be used in every application of an instrument to ascertain its cultural, historical, and biographical sensitivity and specificity. Indeed, the use of qualitative methods is important not just to strengthen their mathematized validity but also because they surface features of experience obscured by virtue of the "denaturalization, objectification, and encapsulation of meaning . . . and decontextualization of lived experience" (Masse, 2000, pp. 411-412) that characterize instruments.

In summary, instruments require continuous qualitative, in addition to quantitative, assessment. Qualitative data gathered during the use and evaluation of quantitative instruments as well as insights gained from qualitative research related to the target population can illuminate and partly close the gaps between meaning and measurement. The ideal of the instrument-as-researcher can only be achieved by acknowledging and using the researcher-as-instrument.

NOTES

1. The Beck Depression Inventory is copyright protected and therefore cannot be reproduced here.
2. There is no mention in the manual of any differences between the scores of subjects who read the items themselves and of those who have the items read to them.

REFERENCES

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: Author.

Barroso, J. (1997). Reconstructing my life: Becoming a long-term survivor of AIDS. *Qualitative Health Research*, *7*, 57-74.

Barroso, J. (2000). "Just worn out": A qualitative study of HIV-related fatigue. In S. G. Funk, E. M. Tornquist, J. Leeman, M. S. Miles, & J. S. Harrell (Eds.), *Key aspects of preventing and managing chronic illness* (pp. 183-194). New York: Springer.

Barroso, J., & Powell-Cope, G. M. (2000). Metasynthesis of qualitative research on living with HIV. *Qualitative Health Research*, *10*, 340-353.

Beck, A. T., Steer, R. A., Ball, R., & Ranieri, W. F. (1996). Comparison of Beck Depression Inventories–IA and –II in psychiatric outpatients. *Journal of Personality Assessment*, *67*, 588-597.

Beck, A. T., Steer, R. A., & Brown, G. K. (1996). *BDI-II manual*. San Antonio, TX: The Psychological Corporation.

Beebe, D. W., Finer, E., & Holmbeck, G. N. (1996). Low-end specificity of four depression measures: Findings and suggestions for the research use of depression tests. *Journal of Personality Assessment*, *67*, 272-284.

Breitbart, W., McDonald, M. V., Rosenfeld, B., Monkman, N. D., & Passik, S. (1998). Fatigue in ambulatory AIDS patients. *Journal of Pain and Symptom Management*, *15*, 159-167.

Chochinov, H. M., Wilson, K. G., Enns, M., & Lander, S. (1997). "Are you depressed?" Screening for depression in the terminally ill. *American Journal of Psychiatry*, *154*, 674-676.

Cleary, P. D., Fowler, F. J., Weissman, J., Massagli, M. P., Wilson, I., Seage, G. R., Gatsonis, C., & Epstein, A. (1993). Health-related quality of life in persons with acquired immune deficiency syndrome. *Medical Care*, *31*, 569-580.

Cunningham, W. E., Shapiro, M. F., Hays, R. D., Dixon, W. J., Visscher, B. R., George, L., Ettl, M. K., & Beck, K. (1998). Constitutional symptoms and health-related quality of life in patients with symptomatic HIV disease. *American Journal of Medicine*, *104*, 129-136.

Dahlstrom, W. G., Brooks, J. D., & Peterson, C. D. (1990). The Beck Depression Inventory: Item order and the impact of response sets. *Journal of Personality Assessment*, *55*, 224-233.

De Boer, J. B., van Dam, F.S.A.M., Sprangers, M.A.G., Frissen, P.H.J., & Lange, J.M.A. (1993). Longitudinal study on the quality of life of symptomatic HIV-infected patients on a trial of zidovudine versus zidovudine and interferon-a. *AIDS*, *7*, 947-953.

Deshields, T. L., Tait, R. C., Gfeller, J. D., & Chibnall, J. T. (1995). Relationship between social desirability and self-report in chronic pain patients. *Clinical Journal of Pain*, *11*, 189-193.

Faravelli, C., Albanesi, G., & Poli, E. (1986). Assessment of depression: A comparison of rating scales. *Journal of Affective Disorders*, *11*, 245-253.

Fleishman, J. A., & Crystal, S. (1998). Functional status transitions and survival in HIV disease. *Medical Care*, *36*, 533-543.

Fleury, J. (1993). Preserving qualitative meaning in instrument development. *Journal of Nursing Measurement*, *1*, 135-144.

Hoover, D. R., Saah, A. J., Bacellar, H., Murphy, R., Visscher, B., Anderson, R., & Kaslow, R. A. (1993). Signs and symptoms of "asymptomatic" HIV-1 infection in homosexual men. *Journal of Acquired Immune Deficiency Syndromes*, *6*, 66-71.

Judd, F., Mijch, A., McCausland, J., & Cockram, A. (1997). Depressive symptoms in patients with HIV infection: A further exploration. *Australian and New Zealand Journal of Psychiatry*, *31*, 862-868.

Kalichman, S. C., Rompa, D., & Cage, M. (2000). Distinguishing between overlapping somatic symptoms of depression and HIV disease in people living with HIV-AIDS. *Journal of Nervous and Mental Disease*, *188*, 662-670.

Kalichman, S. C., Sikkema, K. J., & Somlai, A. (1995). Assessing persons with human immunodeficiency virus (HIV) infection using the Beck Depression Inventory: Disease processes and other potential confounds. *Journal of Personality Assessment*, *64*(1), 86-100.

Kvale, S. (1995). The social construction of validity. *Qualitative Inquiry*, *1*, 19-40.

Langevin, R., & Stancer, H. (1979). Evidence that depression rating scales primarily measure a social undesirability response set. *Acta Psychiatrica Scandinavia*, *59*, 70-79.

Masse, R. (2000). Qualitative and quantitative analyses of psychological distress: Methodological complementarity and ontological incommensurability. *Qualitative Health Research*, *10*, 411-423.

Maxwell, J. A. (1992). Understanding and validity in qualitative research. *Harvard Educational Review*, *62*, 279-300.

Meek, P., & Nail, L. M. (1997). *Internal consistency reliability and construct validity of a new measure of cancer treatment related fatigue: The General Fatigue Scale (GFS)*. Paper presented at the Oncology Nursing Society Annual Congress, New Orleans, LA.

Moran, P. W., & Lambert, M. J. (1983). A review of current assessment tools for monitoring changes in depression. In M. S. Lambert, E. R. Christensen, & S. DeJulio (Eds.), *The assessment of psychotherapy outcome* (pp. 263-303). New York: Wiley.

Morgan, D. L. (1993). Qualitative content analysis: A guide to paths not taken. *Qualitative Health Research*, *3*, 112-121.

O'Dell, M. W., Hubert, H. B., Lubeck, D. P., & O'Driscoll, P. (1998). Pre-AIDS physical disability: Data from the AIDS time-oriented health outcome study. *Archives of Physical Medicine and Rehabilitation*, *79*, 1200-1205.

Osman, A., Downs, W. R., Barrios, F. X., Kopper, B. A., Gutierrez, P. M., & Chiros, C. E. (1997). Factor structure and psychometric characteristics of the Beck Depression Inventory–II. *Journal of Psychopathology and Behavioral Assessment*, *19*, 359-376.

Perkins, D. O., Leserman, J., Stern, R. A., Baum, S. F., Liao, D., Golden, R. N., & Evans, D. L. (1995). Somatic symptoms and HIV infection: Relationship to depressive symptoms and indicators of HIV disease. *American Journal of Psychiatry*, *152*, 1776-1781.

Richter, P., Werner, J., Heerlein, A., Kraus, A., & Sauer, H. (1998). On the validity of the Beck Depression Inventory. *Psychopathology*, *31*, 160-168.

Steer, R. A., Ball, R., Ranieri, W. F., & Beck, A. T. (1999). Dimensions of the Beck Depression Inventory-II in clinically depressed outpatients. *Journal of Clinical Psychology*, *55*(1), 117-128.

Sutton, R. I. (1997). The virtues of closet qualitative research. *Organization Science*, *8*, 97-106.

Treichler, P. A. (1990). AIDS, homophobia and biomedical discourse: An epidemic of signification. *Cultural Studies*, *1*, 263-305.

Walker, K., McGown, A., Jantos, M., & Anson, J. (1997). Fatigue, depression, and quality of life in HIV-positive men. *Journal of Psychosocial Nursing*, *35*(9), 32-40.

Waltz, C. F., Strickland, O. L., & Lenz, E. R. (1991). *Measurement in nursing research* (2nd ed.). Philadelphia: F. A. Davis.

Weinholtz, D., Kacer, B., & Rocklin, T. (1995). Salvaging quantitative research with qualitative data. *Qualitative Health Research*, *5*, 388-397.

Wessely, S., Hotopf, M., & Sharpe, M. (1998). *Chronic fatigue and its syndromes*. Oxford, UK: Oxford University Press.

Julie Barroso, Ph.D., ANP, C.S., is an assistant professor in the School of Nursing at the University of North Carolina at Chapel Hill.

Margarete Sandelowski, Ph.D., R.N., FAAN, is a professor in the School of Nursing at the University of North Carolina at Chapel Hill.