

Reliability of Instruments in a Cooperative, Multisite Study: Employment Intervention Demonstration Program

Michelle P. Salyers,^{1,4} Gregory J. McHugo,² Judith A. Cook,³ Lisa A. Razzano,³ Robert E. Drake,² and Kim T. Mueser²

Reliability of well-known instruments was examined in 202 people with severe mental illness participating in a multisite vocational study. We examined interrater reliability of the Positive and Negative Syndrome Scale (PANSS) and the internal consistency and test-retest reliability of the PANSS, the Rosenberg Self-Esteem Scale, the Medical Outcomes Study Short Form-36 (SF-36), and the Quality of Life Interview. Most scales had good levels of reliability, with intraclass correlation coefficients (ICCs) and coefficient alphas above .70. However, the SF-36 scales were generally less stable over time, particularly Social Functioning (ICC = .55). Test-retest reliability was lower among less educated respondents and among ethnic minorities. We recommend close monitoring of psychometric issues in future multisite studies.

KEY WORDS: reliability; stability; severe mental illness; psychometrics.

The Employment Intervention Demonstration Program (EIDP) is a cooperative study of eight sites providing vocational rehabilitation services to people with severe mental illness (SMI). The EIDP is an effectiveness trial, using typical clients and field sites rather than carefully selected subjects or highly controlled treatment settings. The common research protocol includes standardized instruments that are frequently used in evaluations of services for people with SMI. In addition, the EIDP has built-in features to maximize reliability, including extensive interviewer training, assessment of test-retest and interrater reliability, and feedback to the individual sites on reliability performance. Nevertheless, it is important to establish reliability and to identify sources of

unreliability in instruments, sites, procedures, interviewers, or clients (e.g., Drake, McHugo, & Biesanz, 1995). Unreliability may affect ongoing training procedures and data analytic approaches within a study. Further, unreliability might affect choice of instruments or training procedures in future studies of similar populations.

A close examination of reliability is particularly important within the context of collaborative, multisite studies (Tracy et al., 1997). In such studies, several sites with similar research goals and objectives use a common set of instruments and data collection procedures. Ultimately, client characteristics and outcomes from the common protocol may be compared across sites, thereby enhancing the generalizability of the individual studies. However, the value of this approach depends upon uniformity of assessment across sites. That is, in order to compare outcomes across clients and sites, measures must demonstrate reliability across these domains (i.e., client characteristics and site). In multisite studies, however, reliability may be more difficult to establish and maintain. For example, even with centralized training of interviewers, once interviewers return to their sites, there may be differences in local supervision or other conditions that

¹Department of Psychology, Indiana University Purdue University Indianapolis, Indiana.

²Dartmouth Medical School and NH-Dartmouth Psychiatric Research Center, Lebanon, New Hampshire.

³Mental Health Services Research Program, Department of Psychiatry, University of Illinois at Chicago, Illinois.

⁴Correspondence should be directed to Michelle P. Salyers, Department of Psychology, Indiana University Purdue University Indianapolis, LD124, 402 N. Blackford St., Indianapolis, Indiana 46202-3275; e-mail: mpsalyer@iupui.edu.

could adversely affect consistency across sites. The resulting divergence would add considerable error to cross-site analyses. Thus, assessment of reliability in multisite studies is important for increasing the rigor of the project and enhancing the generalizability of the findings.

The EIDP provided a unique opportunity to examine reliability of widely used instruments in a multisite, real-world setting. The common research protocol included the Positive and Negative Syndrome Scale (PANSS; Kay, Fiszbein, & Opler, 1987), the Rosenberg Self-Esteem Scale (Rosenberg, 1965), the Medical Outcomes Study Short Form-36 (SF-36; Ware & Sherbourne, 1992), and the Quality of Life Interview (QOLI; Lehman, 1988). Each of these instruments has been used frequently in evaluations of programs for people with SMI, with varying degrees of psychometric validation.

The PANSS and QOLI were developed for the SMI population, and substantial evidence of reliability has been provided both for the PANSS (Kay, Opler, & Lindenmayer, 1989) as well as for the QOLI (Lehman, 1996). Although well studied in other populations, the Rosenberg Self-Esteem Scale and the SF-36 have less psychometric validation specifically in SMI populations. A few studies have examined the SF-36 in varying populations of people with psychiatric disorders. The SF-36 has been shown to be internally consistent in samples of people with schizophrenia (Russo et al., 1998), bipolar disorder (Leidy, Palmer, Murray, Robb, & Revicki, 1998), and in a broader sample of homeless people with mental illness (Wood, Hurlburt, Hough, & Hofstetter, 1997). Two studies have examined test-retest reliability and have found the SF-36 to be moderately stable in people with schizophrenia over a 1-week period (Russo et al., 1998) and in a larger sample of psychiatric outpatients over a 3 week period (Burke, Burke, Baker, & Hillis, 1995). However, the Burke et al. sample did not include people with psychotic disorders. Despite international use and multiple psychometric studies of the Rosenberg Self-Esteem Scale (e.g., Ferring & Filipp, 1996), we could not locate studies specifically examining the psychometric properties of this scale in a sample of people with SMI. However, generally good levels of internal consistency have been reported in the context of broader studies (Arns & Linney, 1993; Torrey, Mueser, McHugo, & Drake, 2000; Van Dongen, 1996). In addition, Torrey et al. (2000) reported good stability over a 2-week period ($ICC = .87$).

The purpose of this study was to assess the reliability of commonly used instruments in a multisite sample of persons with SMI. We examined internal consistency of scales and their test-retest stability, as well as interrater reliability for clinically rated scales (i.e., PANSS). In addition, we were also interested in identifying client characteristics (e.g., education level) that might systematically affect the reliability of the instruments. Thus, the test-retest reliability sample was split on several client characteristics to examine reliability between subgroups of the overall sample. Finally, we identified unreliable clients, that is, clients who were “consistently inconsistent” upon retesting across instruments, and we contrasted them with highly stable responders as another means of determining factors influencing reliability.

METHOD

Background

The sample used in this analysis was a subgroup of respondents in a multisite research demonstration program to provide vocational rehabilitation services to people with severe and persistent mental illness. This program, known as the Employment Intervention Demonstration Program (EIDP), was funded by the Center for Mental Health Services of the Substance Abuse and Mental Health Services Administration, beginning in 1995. The eight project sites evaluated vocational rehabilitation programs for clients with SMI who were accessing public mental health services.

A Coordinating Center (CC), based at the University of Illinois at Chicago, in collaboration with Human Services Research Institute of Massachusetts, was responsible for working with the sites to develop a common protocol of research measures to be administered across sites. The CC organized a 2-day intensive train-the-trainer interviewer training for all sites consisting of three components. The first component was a half-day training on general interviewing techniques based on the University of Michigan Institute for Social Research's Personal Interviewer Training (Guenzel, Berckmans & Cannell, 1983). The second was a 1-day training on administration of the Positive and Negative Syndrome Scale (PANSS) (Kay et al., 1987), conducted by one of the scale's authors involving didactic and interactive (group rating and discussion of videotaped assessments) instructional techniques. The third component involved training on

using the Quality of Life Interview (QOLI) (Lehman, 1988), conducted by the scale's author and involving didactic and interactive (group rating and discussion of videotaped assessments) instruction. Manuals for each of the three training components were distributed to all sites. In addition, the CC prepared a common protocol interviewer manual to promote standardization of interviewer procedures based on pilot testing conducted at some sites and questions from the field. Moreover, each site conducted further training and developed procedures designed to ensure standardization of administration across the site's interviewers. Specifically for the PANSS, interviewers participated in telephone conferences with the CC and one of the PANSS developers (Dr. Opler) monthly for the first year, and then quarterly during the project.

Interviewers and interview settings varied across sites; while interviewers at all sites conducted interviews in service delivery settings, some also conducted interviews in clients' homes or in public settings of the respondent's choice. Interviewers were required to have at least a master's degree and/or research interviewing experience; however, interviewers differed in age, race, gender, prior experience, and educational background.

Procedures

The procedures for interrater and test-retest reliability were developed collaboratively by the sites, CC, and members of the EIDP Steering Committee. Each site agreed to conduct 20 interrater reliability assessments using the PANSS: 5 of these assessments were conducted using the same videotaped interview across sites, and an additional 15 assessments involved rating videotaped or "live" interviews that were unique to each site. That is, a single interview was conducted with two or more raters. In addition, each site readministered the remainder of the common protocol to a sample of respondents following an "ideal" 2-week interval.

Respondent Characteristics

The total sample for this analysis was 202; 57% were male and 43% female; 43% were Caucasian/non-Hispanic, 38% were African American, 12% were Hispanic/Latino, 5% were Native American, 0.5% were Asian, and 2% were of "other" or mixed

ethnicities. The mean age was 39.4 years ($SD = 10.1$). Twenty percent had less than a high school degree, 30% had completed high school or had a GED, 26% reported attending some college, and 16% reported an Associates Degree or better. The sample averaged a mean of 15.2 months of lifetime hospital use ($SD = 42.2$; median = 6.0). Regarding adherence to psychotropic medication regimens, 85% rated themselves as "almost always" compliant with medications; even higher proportions reported avoiding use of street drugs in the past month (91%) and no days high from alcohol (83%).

Although each of the sites targeted people with SMI who were interested in employment, specific characteristics of participants varied across sites. Sites differed on age, $F(7,193) = 2.33$, $p < .05$, race (white vs. other), $\chi^2 = 58.44$, $p < .001$, and education level, $F(7,194) = 4.52$, $p < .001$. However, sites did not differ on client characteristics of gender or lifetime hospital use.

Assessment Instruments

The common protocol for the EIDP included several instruments that are commonly used in evaluations of services for people with SMI. The protocol also included some instruments developed specifically for the EIDP, for example, scales assessing work motivation and misperceptions of social security benefits. For the current study, we focused on instruments that are widely used.

The Positive and Negative Syndrome Scale (PANSS)

The PANSS is a 30-item assessment of positive and negative psychiatric symptoms as well as general severity of mental disorder. The PANSS includes three scales: the positive scale consists of 7 items measuring symptoms such as hallucinations and delusions; the negative scale includes 7 items assessing symptoms such as blunted affect and apathetic social withdrawal; and a general psychopathology scale includes 16 items assessing severity of schizophrenic illness. Symptom severity is assessed on a 7-point scale ranging from 1 (*absent*) to 7 (*extreme*). Internal consistency reliability coefficients from a study of 101 individuals with schizophrenia were .73, .83, and .79, respectively, for the positive, negative, and general psychopathology scales (Kay et al., 1987). This scale was included in the common protocol in its entirety.

Rosenberg Self-Esteem Scale

The Rosenberg Self-Esteem Scale is a unidimensional measure of global self-esteem. As measured by this scale, high self-esteem involves the feeling that one is a person of worth and has self-respect, but does not consider oneself to be superior to others. Low self-esteem implies dissatisfaction with oneself and self-rejection in which the self-picture is disagreeable and one wishes it were otherwise. This 10-item measure elicits responses to positive and negative statements that are rated on a 4-point scale ranging from 1 (*strongly agree*) to 4 (*strongly disagree*). The internal consistency value for the scale in a study of 5,024 high school juniors and seniors from 10 randomly selected high schools in New York state was .93 (Rosenberg, 1965). This scale was included in the common protocol in its entirety.

The Medical Outcomes Study Short Form-36 (SF-36)

The SF-36 assesses eight health concepts: physical limitations, role limitations due to physical health, social functioning, bodily pain, general health perceptions, general mental health, vitality, and role limitations due to emotional problems. It is designed for use in basic research and clinical practice, as well as health policy and general population surveys (Ware & Sherbourne, 1992). For the purposes of the EIDP, only four of the SF-36 scales were included in the common protocol: general health, consisting of five items; role limitations from physical health issues, consisting of four items; role limitations due to emotional problems, consisting of three items; and social functioning, consisting of two items. In a study of over 2,471 patients visiting medical practice settings that participated in the Medical Outcomes Study (Hays, Sherbourne, & Mazel, 1993) these scales had internal consistency reliabilities of .78, .84, .83, and .85, respectively.

The Quality of Life Interview (QOLI)

The QOLI consists of items assessing life domains including: general life satisfaction, daily activities and functioning, housing, family relations, social relations, leisure activities, work and school, legal and safety issues, finances, and health. Items comprising each life domain elicit both objective and subjective (i.e., satisfaction) information regarding quality

of life for people with severe mental illness. Internal consistency reliabilities for the QOLI in studies of 278 mentally ill residents of 30 large board and care homes in Los Angeles, 99 chronically mentally ill inpatients in Rochester, NY, and 92 chronically mentally ill residents in supervised community residences in Rochester, NY ranged from .79 to .88 for the subjective scales, while objective quality of life scale reliability coefficients ranged from .42 to .82 (Lehman, 1988). The EIDP common protocol included the following QOLI domains: general life satisfaction, housing, family relations, social relations, leisure activities, legal and safety issues, finances, and health. Our analyses included the subjective satisfaction items rather than the objective items, because of relatively low internal consistency of the objective scales (Lehman, 1988) and conceptually objective items may not form a consistent scale (e.g., participation in a sport may not be expected to be related to reading a book).

Data Analysis

We examined both interrater and test-retest reliability with the intraclass correlation coefficient (ICC). As described by Shrout and Fleiss (1979) and by McGraw and Wong (1996), there are several models for the calculation of the ICC, which vary on the basis of whether raters are considered random or fixed effects, whether each rater rates each subject, and whether the interest is in consistency or agreement. In order to examine interrater reliability at the site level, we used a model in which all raters rate each subject, but the rater effect is treated as a random rather than a fixed factor. This was considered most appropriate, because we were interested in rater agreement, that is, the extent to which raters could be considered interchangeable, rather than rater consistency. For sites in which more than two raters were used, we randomly selected two raters, and we calculated the ICC for all available ratings, that is, study clients and the five common videotaped interviews. For the total sample, however, we used only actual client interviews so there would not be redundancy in the data. Consequently, ICCs for the total sample were based on a model in which not all raters rated each subject.

To assess test-retest reliability, we again used the ICC, conceptualizing each administration of the interview as a separate "judge." Again, we chose a model in which raters are a random factor and not all clients are rated by the same two raters. As with interrater

Table 1. Interrater Reliability (ICCs) for PANSS Scales for the Total Sample ($N = 115$) and Mean Across Sites ($N = 8$)

Scale	Total sample ICC	95% CI ICC	Mean (<i>SD</i>) across sites	No. of sites ICC < .70
Positive symptoms	.87	.82–.91	.92 (0.05)	0
Negative symptoms	.74	.65–.82	.81 (0.16)	1
General symptoms	.71	.60–.79	.83 (0.15)	1

Note. Total ICC and 95% CI are based on formula 1, and site-level ICCs are based on formula 2 (Shrout & Fleiss, 1979).

reliability, we calculated test-retest ICCs both within each site as well as for the total sample.

To examine factors associated with reliability, we formed subgroups of clients on the basis of client characteristics (e.g., gender). We compared the stability of scales for subgroups by directly comparing ICCs, using the formula described by Alsawalmeh and Feldt (1992). We also compared subgroups on internal consistency (Cronbach's alpha) using a similar formula (Charter & Feldt, 1996; Feldt, 1969). Because of the large number of comparisons and the small sample sizes within most subgroups, we focused on patterns of results, rather than individual levels of statistical significance.

Clear guidelines for the assessment of acceptable levels of ICC values have not been identified. Further, the acceptable level of reliability will depend upon the type of reliability (e.g., interrater vs. test-retest) and the purposes of the assessment. However, we chose a minimum criterion of .70 as a lower bound for good reliability, with lower values representing questionable reliability. We recognize that while the absolute level is somewhat arbitrary, some value is needed in order to apply a consistent standard of evaluation across scales.

RESULTS

Interrater Reliability

Overall, raters demonstrated acceptable to good agreement on each of the PANSS scales for the total sample. As shown in Table 1, Positive Symptoms had the strongest overall reliability; Negative Symptoms and General Symptoms were somewhat lower, but still above .70. When averaged across sites, ICCs were higher for each of the scales. Raters had the greatest agreement for Positive Symptoms, mean ICC = .92 ($SD = 0.05$). On this scale, no sites were below the minimum criterion of .70. The mean ICCs for Negative Symptoms and General Symptoms also were strong. Only one site was below .70 on both of these scales. No other sites had consistently low scores across scales. Two sites had ICCs above .90 on each of the scales.

Internal Consistency

As shown in Table 2, the majority of the scales were internally consistent. All but two scales had

Table 2. Internal Consistency (Cronbach's Alpha) for Total Sample ($N = 202$) and Mean (*SD*) Across Sites ($N = 8$)

	No. of items	Total sample	Mean (<i>SD</i>) across sites	Range	No. of sites alpha < .70
QOLI					
General life satisfaction	2	.80	.78 (0.08)	.64–.89	1
Housing satisfaction	3	.76	.75 (0.09)	.61–.85	2
Family relations satisfaction	2	.89	.88 (0.08)	.70–.97	0
Social relations satisfaction	3	.78	.74 (0.12)	.58–.94	2
Leisure satisfaction	4	.77	.73 (0.15)	.48–.89	4
Safety satisfaction	3	.86	.84 (0.06)	.76–.94	0
Financial satisfaction	3	.85	.82 (0.10)	.63–.94	1
Health satisfaction	3	.74	.74 (0.08)	.64–.84	3
Rosenberg Self-Esteem	10	.84	.80 (0.08)	.67–.89	1
MOS-SF 36					
General health	5	.79	.77 (0.05)	.70–.86	0
Physical role impairment	4	.86	.82 (0.10)	.67–.96	1
Emotional role impairment	3	.83	.81 (0.14)	.56–.96	2
Social function	2	.67	.65 (0.23)	.22–.91	5
PANSS					
Positive symptoms	7	.65	.64 (0.11)	.42–.72	4
Negative symptoms	7	.78	.76 (0.11)	.55–.87	2
General symptoms	16	.75	.73 (0.12)	.54–.87	2

Cronbach's alphas above .70. Social Functioning (SF-36) had an overall alpha of .67, with five sites not meeting minimum criterion of .70. Similarly, Positive Symptoms (PANSS) had an overall alpha of .65, with four sites not meeting the minimum criterion. Although Leisure Satisfaction (QOL) had acceptable overall internal consistency ($\alpha = .77$), half of the sites did not meet minimum criterion of .70. It should be noted that Social Functioning consists of only two items, which can lower alpha levels. However, two of the QOLI scales (General Life Satisfaction and Family Relations Satisfaction) also consisted of only two items each, but they still attained strong levels of internal consistency. Two sites had generally high levels of internal consistency across scales; only one scale for each site had an alpha below .70. Three sites had generally low internal consistencies, with alphas less than .70 on six of 16 scales.

Test-Retest Reliability

Although the goal for each site was to achieve a 2-week interval between administrations of the test and retest interviews, there was substantial variation, ranging from interviews administered the same day ($n = 2$) to one interview that took place 118 days after the initial interview. In order to narrow the sample to similar test-retest assessments, we excluded six outliers in length of time between administrations: two clients interviewed on the same day, and

four clients where the test-retest interval exceeded 60 days. The remainder of the analyses were based on this restricted sample ($n = 202$), with a mean length of time between interviews of 12.5 days ($SD = 5.2$) and a range from 4 to 33 days. As a check, we compared subgroups of clients with short intervals (<11 days), medium intervals (11–17 days), or long intervals (>17 days) between interviews and found no consistent relationship between length of time between administrations and assessments and stability.

Prior to examining the test-retest reliability of the instruments, we examined the stability of several responses that would be unlikely to change within a period of a month (gender, race, marital status, and education level) in order to provide a reference point for expected levels of stability. Following Bartko and Carpenter (1976), we used Kappa to measure reliability of the categorical variables. Although none of the four variables we examined had perfect reliability over time, all had high levels: gender ($\kappa = .96$), race ($\kappa = .91$), marital status ($\kappa = .90$), education level (ICC = .91).

As shown in Table 3, the majority of scales were stable over time. Six of the eight QOL satisfaction scales had ICCs greater than or equal to .70; Family Relations and Social Relations were below .70. The Self-Esteem scale and PANSS scales had good stability overall. However, three of the four SF-36 scales had relatively poor stability: Physical Role Impairment, Emotional Role Impairment, and Social Functioning.

Table 3. Test-Retest Reliability (ICCs) for Total Sample ($N = 202$) and Mean Across Sites ($N = 8$)

	Total sample ICC	95% CI	Mean across sites ICC (<i>SD</i>)	No. of sites ICC < .70
QOLI				
General life satisfaction	.71	.61–.77	.68 (0.16)	3
Housing satisfaction	.74	.67–.79	.72 (0.16)	4
Family relations satisfaction	.66	.58–.73	.63 (0.25)	5
Social relations satisfaction	.65	.56–.72	.61 (0.15)	6
Leisure satisfaction	.71	.63–.77	.70 (0.10)	3
Safety satisfaction	.70	.62–.76	.68 (0.11)	2
Financial satisfaction	.80	.74–.84	.77 (0.09)	2
Health satisfaction	.74	.66–.79	.73 (0.10)	3
Rosenberg Self-Esteem	.80	.75–.85	.74 (0.14)	2
MOS-SF 36				
General health	.76	.70–.82	.76 (0.09)	2
Physical role impairment	.67	.58–.74	.60 (0.15)	5
Emotional role impairment	.63	.54–.71	.60 (0.11)	7
Social function	.55	.44–.64	.50 (0.22)	6
PANSS				
Positive symptoms	.80	.74–.85	.79 (0.12)	1
Negative symptoms	.83	.77–.87	.78 (0.18)	3
General symptoms	.78	.71–.83	.76 (0.11)	2

Note. Total and site-level ICCs based on formula 2 (Shrout & Fleiss, 1979).

Sites were variable in their test-retest performance. Two sites had high levels of stability; both of these sites were below the criterion of .70 on only three scales. In contrast, three sites were below criterion on 10 of the 16 scales.

Client Characteristics and Reliability

We examined the impact of client characteristics by forming subgroups of clients based on age, gender, race, education, and chronicity. Age groups were formed by the lower and upper quartiles, resulting in a younger group (less than 34 years; *n* = 51) and an older group (greater than 44 years, *n* = 51). Males (*n* = 114) were compared to females (*n* = 88). Caucasians (*n* = 86) were compared to other ethnicities (*n* = 116). A low-education group (less than high school, *n* = 57) was compared to a high-education group (some college or more, *n* = 85). Finally, chronicity was determined by the number of months of lifetime hospital use: low chronicity (less than 3 months, *n* = 57) versus high chronicity (12 months or more, *n* = 57).

Although we also were interested in symptomatology, medication compliance, and substance use, large enough groups could not be formed to make meaningful comparisons. For example, 85% were rated as “almost always” compliant with medications,

83% reported no days high from alcohol, and 91% reported no use of drugs in the past month. Similarly, we tried to develop groups of clients with consistently few or consistently frequent symptoms based on the distribution of PANSS scores. However, each of the groups had fewer than 40 clients, which is problematic for comparing ICCs (Alsawalmeh & Feldt, 1994).

Internal consistencies of the measures did not differ as a function of client characteristics. Excluding the scales that had weak overall internal consistency (Social Functioning, Positive Symptoms, and Leisure Satisfaction), there were only two subgroups that had more than one scale below alpha of .70. Clients who were highly educated had low internal consistency on Negative Symptoms and General Symptoms on the PANSS; additionally, alphas for the highly educated group were significantly lower than the alphas for the low education group on these two scales (*p* < .01). Clients who had high levels of chronicity (12 or more months hospitalized) had low internal consistency for General Life Satisfaction and Social Relations Satisfaction; however, these alphas were not significantly different than those for the low chronicity subgroup.

As shown in Table 4, client characteristics were related to test-retest reliability. Omitting the scales with low overall test-retest reliability (Family Satisfaction, Social Relations Satisfaction, Physical Role Impairment, Emotional Role Impairment, and Social

Table 4. Subgroups with Questionable Levels (ICC < .70) of Test-Retest Reliability

	Age		Gender		Race		Education (Years)		Chronicity (Months)	
	<34	45+	M	F	W	Minority	<12	13+	<3	12+
QOLI										
General life satisfaction		X				X	X			X
Housing satisfaction			X			X	X			
Family relations satisfaction		X	X	X		X	X		X	
Social relations satisfaction	X	X	X	X	X	X	X		X	X
Leisure satisfaction		X		X		X	X			
Safety satisfaction	X		X			X	X			X
Financial satisfaction										X
Health satisfaction					X					
Rosenberg Self-Esteem										
MOS-SF 36										
General health										
Physical role impairment	X	X	X	X		X	X		X	
Emotional role impairment	X		X	X		X		X	X	X
Social function	X	X	X	X	X	X	X	X	X	X
PANSS										
Positive symptoms										
Negative symptoms										
General symptoms		X						X		

Note. “X” indicates ICC < .70.

Functioning), there were five subgroups with at least two of the remaining scales with ICCs below .70: older age, males, minority, low education, and high chronicity. Direct comparisons of reliability coefficients between subgroups showed consistent differences for race, education, and chronicity. Scores for minority clients were significantly less stable than for Caucasians on five of the remaining 11 scales (Housing, Leisure, Safety, and Financial Satisfaction, as well as Self-Esteem). Scores for the low education group were significantly less stable than scores for college-educated clients on General Life, Housing, and Leisure Satisfaction, but the ratings of General Symptoms for clients in the low education group were more stable. Scores for clients in the high chronicity group were significantly less stable than clients in the low chronicity group on Housing, Safety, and Financial Satisfaction, as well as on Positive Symptoms.

Because sites differed in test-retest reliability, we examined the relationship between site and client characteristics. Clients from three sites with consistently low test-retest reliability ($n = 76$) were grouped together and compared to clients from the remaining sites ($n = 126$) on the client characteristics. The two groups did not differ significantly on any of the variables examined (age, gender, race, education, and chronicity). Second, because there were three client characteristics related to stability (race, education, and chronicity), we examined the overlap between these groups. Racial and educational differences were strongly associated; in the low education group, 74% were minority clients and 26% were Caucasian, and in the highly educated group, 46% were minorities and 54% were Caucasian, $\chi^2 = 10.76$, $p < .001$. Thus, it is not clear what factors are accounting for the observed differences (e.g., reading level of the forms, ethnic bias in the wording). Chronicity, however, was unrelated to race or to education level.

Finally, we examined client characteristics in a slightly different way, trying to identify inconsistent clients rather than scales. That is, we formed groups of clients based on the consistency of test-retest differences on the client-rated scales (QOL, Self-Esteem, and SF-36). Clients who were at or above the median of absolute change scores on the scales of all three instruments were in the unreliable group ($n = 34$). Clients whose change scores were below the median for each of the three scales were in the reliable group ($n = 28$). These two groups were compared on a variety of measures (demographics, symptoms, medication compliance, substance use, and hospitalization history). The groups differed significantly on three

variables. Unreliable clients were more likely to be minorities: 59% were minority whereas only 32% of the reliable group were minority clients, $\chi^2 = 4.39$, $p < .05$. Unreliable clients had lower mean levels of education: 3.9 (1.4) versus 4.6 (1.3), $t = 2.26$, $p < .05$ (on a scale ranging from 1 to 10). Unreliable clients also were more likely to report using drugs in the month prior to the first interview, 16% versus 0%, $\chi^2 = 3.85$, $p < .05$, but not prior to the second interview, 12% versus 4%, $\chi^2 = 0.92$.

DISCUSSION

For the overall sample, interrater agreement, temporal stability, and internal consistency were satisfactory for the interviewer-rated and client-rated scales that we examined. However, there were some subgroups and some scales that had questionable levels of reliability, particularly when examining test-retest reliability. Prior to discussing findings for the individual scales and client subgroups, some general discussion of test-retest issues is warranted.

On one hand, test-retest reliability is appealing because it most closely represents the notion of reliability as repeatability of measurement (Pedhazur & Schmelkin, 1991). That is, when the same measure of a stable construct is given to the same group of people at two points close in time, we would expect to get very similar responses. On the other hand, test-retest reliability poses several problems. One general problem is that reliability estimates can be overinflated because subjects may remember how they responded the first time and therefore respond similarly the second time (Nunnally, 1978). Thus, the time frame between assessments should be long enough to minimize effects due to memory. In the current study, the majority of subjects were interviewed 2 weeks apart, and the interviews included a large number of instruments. Thus, inflation of test-retest coefficients due to memory for specific responses is not likely to be strong in the current study.

In addition to possible overinflation, there are several factors that can contribute to change in reporting, which lowers reliability estimates. As Kelly and McGrath (1988) note, there are at least four sources of change in test-retest situations: real change in the construct being measured, fluctuations in the construct that represent stable patterns, other change in the subjects or instrument even when the construct is stable, and random errors of measurement. Although it is only the last category that reflects true unreliability,

the sources of change cannot be completely separated. Thus, test-retest reliability should always be interpreted with caution.

One concern in the current study is that because test-retest reliability was examined in the context of an intervention, we might expect clients to change in response to that intervention. Indeed, *poor test-retest* coefficients could actually represent *good sensitivity* of the instruments to detect change. While we agree that this is a possibility, we do not believe that this type of change played a large role in our findings. First, the interventions were targeted at people with SMI, who often have long-standing difficulties that would not be expected to change much within a 2-week time frame. Second, the interventions were vocational in nature, and other studies of vocational services have shown minimal direct impact on nonvocational outcomes such as those examined here (Bond, Drake, Mueser, & Becker, 1997). When improvements in nonvocational outcomes have been found, they are associated with *extended* work experiences, not just brief exposure to work (Bond et al., 2000). Thus, even if some clients began working during the test-retest period, the impact in other areas such as self-esteem and quality of life may not be evident in such a short time frame.

Beyond issues of test-retest reliability, some broader caveats to reliability also exist. Reliability is truly within the eye of the beholder. There are several different types of reliability that can be examined, depending on the sources of error one is interested in minimizing. Moreover, there are no clear guidelines as to what represents an acceptable level of reliability. For example, Nunnally (1978) provides several different minimal levels, ranging from .70 to .90, depending on the purpose of assessment and cost of error. In the context of the current EIDP study, we selected the lower end of this range as our standard. In the EIDP study, unreliability would impact the statistical power to detect differences between groups or to detect differences over time, but would not directly impact the clients involved (such as when placement decisions are based on a test score). Further, power analyses can be conducted to determine the impact of low reliability on the study findings (DeVellis, 1991), and data analysis can adjust for low reliability. Thus, even lower levels of reliability may be acceptable in this context. We have attempted to provide a variety of estimates of reliability, as well as some possible explanations of low reliability coefficients, so that readers can make their own judgements as to the acceptability of a particular scale.

Regarding individual scales examined, there was good agreement between raters on the PANSS, with only one site falling below criterion levels on two of the PANSS scales. Internal consistency was low for positive symptoms, with several sites showing questionable levels. However, stability was fairly good, across sites and in the overall sample. For the PANSS, the reliability coefficients in our study are similar, but slightly lower than those reported by Kay et al. (1989) for interrater and internal consistency, and the stability coefficients in our sample (.78 to .83) were somewhat higher. We examined ratings over a shorter time interval than did Kay et al., so stronger stability in the current study may be expected.

The QOLI satisfaction scales had good internal consistency, with alpha values similar to those reported in other large samples of people with SMI (Lehman, 1988; Russo et al., 1997). Moreover, six of the eight scales were stable over a 2-week period. The Rosenberg Self-Esteem Scale also was internally consistent ($\alpha = .84$) and fairly stable (ICC = .80) in our sample, similar to recent findings in other samples of people with SMI (Arns & Linney, 1993; Torrey et al., 2000; Van Dongen, 1996).

The scales from the SF-36 were more variable in their performance. The internal consistency of most of these scales was good ($\alpha = .79$ or above), except for Social Functioning, which had an alpha of .67. The two items that comprise this scale ask for perceptions about the frequency and extent to which health interferes with social activities. Although internal consistency is somewhat limited by the number of items on a scale, other two-item scales in this study fared much better (with α of .80 and .89). Our findings are similar to Russo et al. (1998) who found lower levels of internal consistency for Social Functioning (.65) when compared to the other SF-36 scales (.75 or above) in their sample of people with schizophrenia. Interestingly, Social Functioning was as reliable or better than other scales in the Medical Outcomes sample (McHorney, Ware, Lu, & Sherbourne, 1994). Thus, for people with severe mental illness, these two items may tap somewhat different aspects of social functioning. It is also possible that these items may be differently interpreted in the SMI population, leading to unstable responses.

Three of the four SF-36 scales also were relatively unstable over time. Social Functioning, Physical Role Impairment, and Emotional Role Impairment all had coefficients less than .70, and more than half of the sites had low stability coefficients on each of these scales as well. General Health, however, was more

stable ($ICC = .76$). In the Russo et al. (1998) sample, Pearson correlations between these scales over a 1-week interval were considered "good," ranging from .73 to .83, with the exception of Social Functioning, which was correlated .42 with itself over time. In a broader sample of psychiatric outpatients (excluding those with psychotic disorders), Burke et al. (1995) found strong stability over a very short time interval (duration of psychiatric appointment), with ICCs ranging from .72 to .97. Over a month, however, stability was much lower; General Health ($ICC = .88$) and Social Functioning ($ICC = .69$) were relatively stable, but Physical Role Impairment and Emotional Role Impairment fared much worse, with ICCs of .39 and .51, respectively. Leidy et al. (1998) reported even lower stability in their sample of euthymic patients with Bipolar disorder, ranging from .18 to .71 on these four scales. However, they used an 8-week time period between assessments, and patients were in treatment as well. Thus, change in scores would be expected.

Across each of these studies, including our own, General Health was the most stable of the four scales studied, exceeding the minimum criterion of .70, despite differing time intervals and patient samples. The other scales were more variable in their performance. It may be that global estimates of health are more stable than specific areas, such as social functioning or physical role impairments. Others have begun to use general mental and physical health components from the SF-36, which include items from each of the health domains, rather than individual scales (Ware et al., 1995). The psychometric properties of these components have yet to be examined among people with SMI, but they may provide more stable estimates of overall physical and mental health. For example, in a study of the SF-12 (a briefer version of the SF-36), component summary scores showed good stability ($ICC = .79$) over a 2-week period in a sample of people with SMI (Salyers, Bosworth, Swanson, Lamb-Pagone, & Osher, 2000).

We identified several demographic factors associated with unreliability in our sample. Notably, racial minority status and lower education (less than a high school degree) were associated with low stability on the majority of scales examined, independent of site differences. However, race and education were confounded. Thus, it is not clear whether difficulties were primarily racial (e.g., cultural bias in the wording) or educational (e.g., reading level of the forms). Nevertheless, these findings suggest that the wording of these instruments, the QOLI in particular, may need

attention to improve reliability in some subgroups of clients.

Although no site was below standard on the majority of scales, there were three sites that had relatively low internal consistency and stability coefficients on several of the scales. Some of this inconsistency may have been due to the small sample sizes upon which the within-site reliability coefficients were based. However, most sites had acceptable reliability on most scales. Further, these sites did not differ significantly from the other sites on the client variables predictive of unreliability. Thus, the source of unreliability is unclear, and further exploration would be necessary to determine the factors at work. Although a few client factors were identified as sources of unreliability, future work might look at interviewer characteristics such as level of training, experience, gender, or race as well as the interaction of these characteristics with client characteristics. For example, interviewer gender may influence amount of disclosure during psychiatric interviews (Pollner, 1998). Additional factors such as interview setting (Drake et al., 1995) also might contribute to reliability.

This study provides additional support for the reliability of several scales commonly used in evaluations of rehabilitation programs for persons with SMI. Within a context of close attention to interviewer training and feedback, a multisite study, like the EIDP, can reliably implement these measures. Given the focus on interviewer training and efforts to achieve reliability and standardization in the EIDP, the variability in the study sites supports the need for continued close attention to these issues in future multisite studies. In addition, further work needs to be done for certain subgroups of clients, for example, those with the lowest levels of education, and for certain scales, for example, Social Functioning on the SF-36, to improve reliability.

ACKNOWLEDGMENTS

This study is part of the Employment Intervention Demonstration Program (EIDP), a multisite collaboration among eight research demonstration sites, a coordinating center, and the Center for Mental Health Services, Substance Abuse and Mental Health Services Administration. This research was funded, in part, by the Center for Mental Health Services, Substance Abuse and Mental Health Services Administration (Cooperative Agreement # 5 UD7 SM51820). The opinions expressed herein

do not reflect the official position or policy of the funding agency. We thank members of the EIDP Steering Committee and Publications Committee for comments on an earlier draft of this paper.

REFERENCES

- Alsawalmeh, Y. M., & Feldt, L. S. (1992). Test of the hypothesis that the intraclass reliability coefficient is the same for two measurement procedures. *Applied Psychological Measurement, 16*, 195–205.
- Alsawalmeh, Y. M., & Feldt, L. S. (1994). Testing the equality of two related intraclass reliability coefficients. *Applied Psychological Measurement, 18*, 183–190.
- Arns, P. G., & Linney, J. A. (1993). Work, self, and life satisfaction for persons with severe and persistent mental disorders. *Psychosocial Rehabilitation Journal, 17*, 63–79.
- Bartko, J. J., & Carpenter, W. T. (1976). On the methods and theory of reliability. *Journal of Nervous and Mental Disease, 163*, 307–317.
- Bond, G. R., Drake, R. E., Mueser, K. T., & Becker, D. R. (1997). An update on supported employment for people with severe mental illness. *Psychiatric Services, 48*, 335–346.
- Bond, G. R., Resnick, S. R., Drake, R. E., Xie, H., McHugo, G. J., & Bebout, R. R. (2001). Does competitive employment improve nonvocational outcomes for people with severe mental illness? *Journal of Consulting and Clinical Psychology, 69*.
- Burke, J. D., Burke, K. C., Baker, J. H., and Hillis, A. (1995). Test-retest reliability in psychiatric patients of the SF-36 health survey. *International Journal of Methods in Psychiatric Research, 5*, 189–194.
- Charter, R. A., & Feldt, L. S. (1996). Testing the equality of two alpha coefficients. *Perceptual and Motor Skills, 82*, 763–768.
- De Vellis, R. F. (1991). *Applied social research methods series: Vol. 26. Scale development: Theory and applications*. Newbury Park, CA: Sage.
- Drake, R. E., McHugo, G. J., & Biesanz, J. C. (1995). The Test-retest reliability of standardized instruments among homeless persons with substance use disorders. *Journal of Studies on Alcohol, 56*, 161–167.
- Feldt, L. S. (1969). A test of the hypothesis that Cronbach's alpha or Kuder-Richardson coefficient twenty is the same for two tests. *Psychometrika, 34*, 363–373.
- Ferring, D., & Filipp, S. (1996). Measurement of self-esteem: Findings on reliability, validity, and stability of the Rosenberg Scale. *Diagnostica, 42*, 284–292.
- Guenzel, P. J., Berckmans, T. R., & Cannell, C. F. (1983). *General Interviewing Techniques: A Self-Instructional Workbook for Telephone and Personal Interviewer Training*. Ann Arbor, MI: Survey Research Center of the Institute for Social Research, The University of Michigan.
- Hays, R. D., Sherbourne, C. D., & Mazel, R. M. (1993). The RAND 36-item health survey 1.0. *Health Economics, 2*, 217–227.
- Kay, S. R., Fiszbein, A., & Opler, L. A. (1987). The Positive and Negative Syndrome Scale (PANSS) for schizophrenia. *Schizophrenia Bulletin, 13*, 261–276.
- Kay, S. R., Opler, L. A., & Lindenmayer, J. P. (1989). The positive and negative syndrome scale (PANSS): Rationale and standardisation. *British Journal of Psychiatry, 155*(Suppl. 7), 59–65.
- Kelly, J. R., & McGrath, J. E. (1988). *Applied social research methods series: Vol. 13. On time and method*. Newbury Park, CA: Sage.
- Lehman, A. F. (1988). A quality of life interview for the chronically mentally ill. *Evaluation and Program Planning, 11*, 51–62.
- Lehman, A. F. (1996). Quality of life interview. In L. I. Sederer & B. Dickey (Eds.), *Outcomes assessment in clinical practice* (pp. 117–119). Baltimore, MD: Williams & Wilkins.
- Leidy, N. K., Palmer, C., Murray, M., Robb, J., & Revicki, D. A. (1998). Health-related quality of life assessment in euthymic and depressed patients with bipolar disorder: Psychometric performance of four self-report measures. *Journal of Affective Disorders, 48*, 207–214.
- McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods, 1*, 30–46.
- McHorney, C. A., Ware, J. E., Lu, R., & Sherbourne, C. D. (1994). The MOS 36-item short-form health survey (SF-36): III. Tests of data quality, scaling assumptions, and reliability across diverse patients groups. *Medical Care, 32*, 40–66.
- Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). NY: McGraw-Hill.
- Pedhazur, E. J., & Schmelkin, L. P. (1991). *Measurement, design, and analysis: An integrated approach*. Hillsdale, NJ: Erlbaum.
- Pollner, M. (1998). The effects of interviewer gender in mental health interviews. *The Journal of Nervous and Mental Disease, 186*, 369–373.
- Rosenberg, M. (1965). The measurement of self-esteem. In *Society and the adolescent self-image* (pp. 16–36). Princeton, NJ: Princeton University Press.
- Russo, J., Roy-Byrne, P., Reeder, D., Alexander, M., Dwyer-O'Connor, E., Dagadakis, C., Ries, R., & Patrick, D. (1997). Longitudinal assessment of quality of life in acute psychiatric inpatients: Reliability and validity. *The Journal of Nervous and Mental Disease, 185*, 166–175.
- Russo, J., Trujillo, C. A., Wingerson, D., Decker, K., Ries, R., Wetzler, H., & Roy-Byrne, P. (1998). The MOS 36-item short form health survey: Reliability, validity, and preliminary findings in schizophrenic outpatients. *Medical Care, 36*, 752–756.
- Salyers, M. P., Bosworth, H. B., Swanson, J. W., Lamb-Pagone, J., & Osher, F. C. (2000). Reliability and validity of the SF-12 Health Survey among people with severe mental illness. *Medical Care, 38*, 1141–1150.
- Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*, 420–428.
- Torrey, W. C., Mueser, K. T., McHugo, G. J., & Drake, R. E. (2000). Self-esteem as an outcome measure in vocational rehabilitation studies of adults with severe mental illness. *Psychiatric Services, 51*, 229–233.
- Tracy, K., Adler, L. A., Rotrosen, J., Edson, R., Lavori, P., & the Veterans Affairs Cooperative Study #394 Study Group. (1997). Interrater reliability issues in multicenter trials, Part I: Theoretical concepts and operational procedures used in department of veterans affairs cooperative study #394. *Psychopharmacology Bulletin, 33*, 53–57.
- VanDongen, C. J. (1996). Quality of life and self-esteem in working and nonworking persons with mental illness. *Community Mental Health Journal, 32*, 535–548.
- Ware, J. E., Kosinski, M., Bayliss, M. S., McHorney, C. A., Rogers, W. H., & Raczek, A. (1995). Comparison of methods for the scoring and statistical analysis of SF-36 health profile and summary measures: Summary of results from the Medical Outcomes Study. *Medical Care, 33*, AS264–279.
- Ware, J. E., & Sherbourne, C. D. (1992). The MOS 36-item short health survey (SF-36): I. Conceptual framework and item selection. *Medical Care, 30*, 473–481.
- Wood, P. A., Hurlburt, M. S., Hough, R. L., & Hofstetter, C. R. (1997). Health status and functioning among the homeless mentally ill: An assessment of the Medical Outcomes Study SF-36 scales. *Evaluation and Program Planning, 20*, 151–161.