THE RELIABILITY OF SEVERELY MENTALLY ILL PATIENTS' REPORT OF TREATMENT SATISFACTION

PATRICK W. CORRIGAN* and MARJORY R. JAKUS
University of Chicago Center for Psychiatric Rehabilitation, 7230 Arbor Drive, Tinley Park, IL 60477, USA

SUMMARY
Patient satisfaction has been identified as an important variable in the evaluation and development of programs for severely mentally ill adults. Measuring this variable assumes, however, that patients can reliably report satisfaction with components of treatment programs. Cognitive deficits characteristic of this population could diminish reliability of satisfaction report. To test this issue, the Patient Satisfaction Interview (PSI) was administered at three epochs during six months of treatment to 30 patients participating in a partial hospitalization program; at each epoch, the test was administered twice to determine its test–retest reliability. Test–retest reliability assessed as raw agreement was found to correlate significantly with levels of psychotic symptoms and intelligence. Level of symptoms was determined at each epoch by completion of the Brief Psychiatric Rating Scale and verbal intelligence was estimated from performance on the vocabulary subscale of the WAIS-R. Pattern analyses showed that earlier items on the PSI elicited more reports of dissatisfaction with treatment than later items. These results suggest that subject and test factors may interact with reliability of patient reports of satisfaction.

KEY WORDS—Patient satisfaction, reliability of tests, severe mental illness.

INTRODUCTION
Program administrators, treatment planners and regulatory agencies have become increasingly interested in programs that are satisfying to severely mentally ill patients (Elbeck and Fecteau, 1990). Patients are more likely to collaborate with treatment teams and to benefit from interventions they perceive as pleasing (Corrigan et al., 1990). Unfortunately, previous research suggests that patient satisfaction is a difficult construct to measure (LeBow, 1983). Typically, patients are asked to report on treatment programs and therapists with whom they may have been intimately involved for a period of time; yet it is not known whether patients can provide reliable and valid information regarding their satisfaction with treatment programs.

Aspects of the test situation may lead to inflated accounts of satisfaction (LeVois et al., 1981). Patients may report high satisfaction to please an interviewer perceived to be affiliated with the treatment milieu. Alternatively, the reliability of patient reports may be compromised by their cognitive deficits (Corrigan, 1990). Patients who have poor insight into their illness or who are conceptually disorganized may not be able to reliably report satisfaction with the treatment program (Amador et al., 1991; MacDonald et al., 1988; McEvoy et al., 1981) or consistently identify its satisfactory components. In addition, lowered intelligence common to this population (Aylward et al., 1984) may prevent patients from fully understanding survey questions, thereby impeding reliability of their scores.

Thus, stability of patients’ responses to a survey on satisfaction with a specific treatment program may be impeded by their deficits in cognition as well as by deficiencies in the test itself. Stability, or lack thereof, of patients’ responses may be represented by test–retest reliability. If our assumptions are correct, test–retest reliability is expected to be positively associated with patients' intelligence and inversely correlated with psychotic symptoms. Identifying variables that affect reliability of patient report of treatment satisfaction has important implications for interpreting program evaluations that include patient satisfaction as an indicator.

*Author to whom correspondence should be addressed

1049-8931/93/040215-05S07.50
©1993 by John Wiley & Sons, Ltd.

Received 15 October 1992
Accepted 1 March 1993
METHODS

Subjects

Patients attending the Partial Hospitalization Program (PHP) at the University of Chicago Center for Psychiatric Rehabilitation were invited to participate in a study of program satisfaction. The PHP is a highly structured psychoeducational program where patients are taught social and coping skills necessary for community survival and where they receive assertive case management. Treatment in this program is typically provided five days a week for six months. PHP patients were included in the study if they were between 18 and 65 years of age, did not have a primary DSM-III-R chart diagnosis of organic mental disorder and had 20/30 corrected vision as measured on the Snellen eye chart. Of subjects meeting inclusion criteria, 95% agreed to participate. Subjects who participated in the study provided information regarding gender, age, education and marital status. Psychiatric diagnosis was obtained from chart review. Subjects also completed three formal tests: the Patient Satisfaction Interview (PSI), the Brief Psychiatric Rating Scale (BPRS) and the vocabulary subscale of the Wechsler Adult Intelligence Scale – Revised (WAIS-R).

Procedure

Subjects were administered the PSI (Corrigan and Jakus, 1993) at three points during their treatment in the PHP: one week after admission, at the three-month midpoint and at discharge. Subjects were administered the test twice within two days at each epoch to measure test–retest reliability; they completed the test six times in all.

The PSI is comprised of 42 items that describe components of day treatment programs. When completing the measure, subjects are instructed to compare the index treatment program (in this study, the University of Chicago PHP) to another program in which they had participated in the past, using a seven point better–worse scale. The instrument was administered by part-time graduate students because students were thought less likely to elicit flattering evaluations than PHP staff members with whom patients had an on-going relationship. Items were combined to yield an Overall Satisfaction score plus four subscale scores: (1) satisfaction with the treatment environment, (2) satisfaction with the therapists, (3) satisfaction with specific treatment strategies and (4) satisfaction with preparation for autonomous living. Previous research has shown PSI scales to have good internal consistency, test–retest reliability, six month stability and content validity (Corrigan and Jakus, 1993).

During the first week of PHP participation, subjects also completed the vocabulary subtest of the WAIS-R. Depending on the age of the subject, correlations of this subtest with verbal IQ range from 0.82 to 0.91 and with overall IQ, of 0.82 to 0.87 (Wechsler, 1981). Psychiatric symptoms were assessed at one week, three months and six months using the BPRS (Lukoff et al., 1986). Interviewers who administered the BPRS had been trained to 85% of criteria established in our laboratory. Two scores were calculated from the BPRS: a psychotic symptoms factor, which is a sum of scores on items covering unusual thought content, hallucinations and conceptual disorganization (Hedlund and Vieweg, 1980), and overall symptomatology, the sum of all items.

RESULTS

Twenty-eight patients participated in the study. Their mean age was 37.8 years (S.D. = 13.5) and the mean number of years of education was 12.8 (S.D. = 2.5). Subjects earned a mean Vocabulary score on the WAIS-R of only 31.9 (S.D. = 19.6); this score is one standard deviation below the norm for adults this age (Wechsler, 1981). The sample was 63.3% female. Two-thirds of the sample was single, 13.3% were married and 20% were divorced or widowed. They were 77% African American and 23% Caucasian. For their primary psychiatric diagnosis, 51.9% had received a DSM-III-R diagnosis of schizophrenia or schizoaffective disorder, 25.9% of affective disorder and 22.2% other diagnoses, including anxiety disorder, substance abuse and Axis II disorders. All patients had received inpatient care. Their mean age at first hospitalization was 25.2 years (S.D. = 15.4), on average 12.7 years earlier (S.D. = 12.6).

Change in psychotic and overall symptoms on the BPRS during participation in the PHP is presented in Table 1. Note the decline in scores consistent with improvement over time. Unfortunately, this decline was not significant for psychotic ($F(2,24) = 0.43$) or overall symptoms ($F(2,24) = 0.45$).

Test–retest reliability

Test–retest reliabilities for the PSI, readministered after a two-day interval, were determined at one
scores, showing that the significant relationship seemed to be specific to psychotic symptoms. Moreover, raw agreements for the total and subscale scores were not found to differ significantly across diagnostic groups; i.e. although psychotic symptoms are most common in schizophrenic patients, it is symptoms and not diagnosis that matters. Verbal intelligence was also shown to be significantly correlated with reliability of the Environment subscale, with low intelligence predicting low reliability. Results of Kruskal-Wallis tests showed no significant differences in raw agreement values across gender or marital status.

Despite these correlations, severe psychotic symptoms and low verbal intelligence do not irreparably diminish reliability. A case-by-case examination of the raw agreements showed that no more than three subjects (10% of the sample) ever exhibited agreements below 60% on the subscales or overall scale at each epoch.

Within-group pattern of satisfaction

The mean of the PSI subscale scores over the three epochs is summarized in Fig. 1; the distributions of PSI scores were approximately normal. Results of a $3 \times 4$ (epoch by subscale) repeated measures ANOVA found significant main effects for subscale [$F(3,33) = 7.70, p < 0.001$], but non-significant effects for epoch [$F(2,22) = 1.32, \text{N.S.}$] and no significant interaction between epoch and subscale [$F(6,66) = 0.47, \text{N.S.}$]. Post hoc tests showed that subjects were less satisfied with the environment than with program content at each test period ($p < 0.05$).

While this finding suggests a dissatisfaction with the therapeutic environment compared to other PHP components, this pattern may alternatively suggest a response bias. The PSI presented environment items first, followed by, in order, therapist, treatment and autonomy subscales. Subjects may have scored significantly lower on the environment scale because they were more critical at the beginning of the interview and assumed their dissatisfaction as the test progressed. This explanation is suggested by the stepwise increase in satisfaction in successive scales seen in Fig. 1.

To learn whether an order effect was the correct explanation, a second version, in which subscale items were randomly dispersed throughout the PSI rather than grouped by subscale, was administered to 15 different severely mentally ill adults in the PHP. This version of the PSI also had high test--retest reliability (intra-class correlations ranging from 0.64

---

**Table 1. BPRS scores for psychotic symptoms and overall symptoms at the first week, the third month and the sixth month**

<table>
<thead>
<tr>
<th></th>
<th>Week 1</th>
<th>Month 3</th>
<th>Month 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>BPRS psychotic</td>
<td>6.75</td>
<td>5.40</td>
<td>4.60</td>
</tr>
<tr>
<td>symptoms</td>
<td>(4.33)</td>
<td>(3.84)</td>
<td>(2.58)</td>
</tr>
<tr>
<td>BPRS overall</td>
<td>46.25</td>
<td>42.20</td>
<td>40.00</td>
</tr>
<tr>
<td>symptoms</td>
<td>(12.06)</td>
<td>(2.59)</td>
<td>(2.60)</td>
</tr>
</tbody>
</table>

Standard deviations of BPRS scores are in parentheses.

---

**Table 2. Test--retest reliability (ICCs) for the Patient Satisfaction Interview (PSI) at the first week, the third month and the sixth month**

<table>
<thead>
<tr>
<th></th>
<th>Week 1</th>
<th>Month 3</th>
<th>Month 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total PSI</td>
<td>0.65</td>
<td>0.74</td>
<td>0.85</td>
</tr>
<tr>
<td>Environment</td>
<td>0.67</td>
<td>0.82</td>
<td>0.75</td>
</tr>
<tr>
<td>Therapists</td>
<td>0.71</td>
<td>0.69</td>
<td>0.82</td>
</tr>
<tr>
<td>Treatment</td>
<td>0.74</td>
<td>0.63</td>
<td>0.48</td>
</tr>
<tr>
<td>Anatomy</td>
<td>0.79</td>
<td>0.58</td>
<td>0.91</td>
</tr>
</tbody>
</table>
Table 3. Percentage raw agreement between responses to items from two test administrations of the Patient Satisfaction Interview (PSI) during week 1. Spearman correlations of percentage raw agreement with symptoms and intelligence

<table>
<thead>
<tr>
<th></th>
<th>Environment</th>
<th>Therapist</th>
<th>Treatment</th>
<th>Autonomy</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean raw agreement</td>
<td>0.82 (0.18)</td>
<td>0.88 (0.13)</td>
<td>0.84 (0.14)</td>
<td>0.88 (0.14)</td>
<td>0.82 (0.12)</td>
</tr>
<tr>
<td>Correlation with BPRS psychotic symptoms</td>
<td>-0.48**</td>
<td>-0.02</td>
<td>-0.39*</td>
<td>-0.23</td>
<td>-0.19</td>
</tr>
<tr>
<td>BPRS overall symptoms</td>
<td>-0.22</td>
<td>-0.16</td>
<td>-0.11</td>
<td>0.09</td>
<td>-0.04</td>
</tr>
<tr>
<td>WAIS-R vocabulary</td>
<td>0.51**</td>
<td>0.03</td>
<td>0.20</td>
<td>0.21</td>
<td>0.37</td>
</tr>
</tbody>
</table>

Standard deviations of percentage raw agreements are in parentheses.

*p < 0.10

**p < 0.05

to 0.84). The second sample was similar to the first in mean age (35.2 years; S.D. = 12.1), years of education (10.9 years; S.D. = 3.1), gender (50% male), marital status (62% single) and ethnicity (60% African American).

Results of a repeated measures ANOVA across subscales did not show significant differences when this randomly ordered form of the PSI was used \[ F(3,21) = 1.42, N.S. \]. This finding suggests that significantly lower environment subscale scores found in the earlier version may indeed be attributable to order effects. Alternatively, this finding may represent no more than statistical regression to the mean. Competing interpretations of the data can only be resolved by replications of this design.

**DISCUSSION**

Reliability of patient report of treatment satisfaction is not a static psychometric quality of the test alone. Rather, it seems to be a dynamic variable that is affected by test construction and characteristics of the person tested. In terms of test construction, patients were likely to be more critical of program components asked early in the satisfaction interview. This outcome may support conclusions of other studies about the interviewing situation; namely, demand characteristics lead to inflated reports of treatment satisfaction (LeVois et al., 1981).

Some patient characteristics also affected the reliability of reports of satisfaction. In particular,
level of psychosis was found to reduce raw agreement on some of the PSI subscales. This finding seemed to specifically reflect the deleterious effects of psychosis, rather than psychiatric symptoms or diagnosis in general, on reporting dissatisfaction. Also, patients with higher intelligence scores were able to more reliably report satisfaction with the treatment environment.

Unfortunately, relationships between level of cognitive function and reliability of satisfaction report were not found consistently across PSI subscales. These effects may be test artifacts, however. Effects of patient characteristics were seen only on the environment subscale of the PSI. The environment subscale had the greatest range and variance of scores; as a result, it was more likely to produce significant correlations. It also came first in the PSI. These characteristics may affect initial ability to understand test questions.

Altogether, these findings may suggest strategies in developing tests that assess patient satisfaction. Subscales with relatively high variance are desirable because they better discriminate patients who are satisfied with program components from those who are not. However, subscales with high variance are also likely to lead to relatively unreliable reports of satisfaction from more cognitively disorganized patients. Tests should be designed and selected depending on the program development questions and the patient population.

Methodological limitations prevented drawing any solid conclusions about the manner in which diagnosis interacts with cognitive deficits and reliability. Diagnoses of subjects in this study were determined from chart review and were not validated by structured clinical interview. Future research needs to adopt more rigorous classification methods to determine whether the relationships of cognitive deficits and reliability of satisfaction report are specific to diagnostic subgroups. Issues of comorbidity should also be investigated. Do severely mentally ill patients who also suffer personality disorders or substance abuse show greater decrements in reliability?

Findings from this study are not meant to discount collection of data that describe patient satisfaction or dissatisfaction. Even though reliability of PSI scores were confounded by test and patient variables, most test–retest reliabilities were still high. Rather, our conclusions are meant to serve as caveats for program developers and evaluators when interpreting satisfaction data.

ACKNOWLEDGEMENTS

The authors wish to thank Tracey Ziegler for assistance in data gathering, Stan McCracken for a cogent review of an earlier draft and Michelle Zimowski for statistical consultation.

REFERENCES


