Regular Article

Reliability of Self-Reported Health Service Use: Evidence from the Women with Co-occurring Disorders, and Violence Study

Sukyung Chung, PhD
Marisa Elena Domino, PhD
Elizabeth W. Jackson, PhD
Joseph P. Morrissey, PhD

Abstract

In behavioral health services research, self-reporting provides comprehensive information on service use, but may have limited reliability because of recall bias and misclassification. This study examines test–retest reliability of self-reported health service use, factors affecting reliability, and the impact of inconsistent reporting on the robustness of cost estimates using the test–retest data from the Women, Co-occurring Disorders, and Violence Study (n = 186). Reliability varies widely across service types: moderate to substantial (k = 0.65–0.94) for any use; slight to substantial (ICC = 0.12–0.93) for quantity of use; and none to moderate (k = −0.06–0.79) for service content, but is not affected by psychiatric symptom severity. Cost estimates do not differ according to the use of test or retest data. Findings suggest that self-reporting provides reliable data on service quantity and is adequate for economic evaluations. However, self-reporting of treatment content in highly specified service categories (e.g., individual counseling during residential treatment) may not be reliable.

Address correspondence to Sukyung Chung, PhD, Department of Psychiatry, University of California, San Francisco, 2727 Mariposa Street, Suite 100, San Francisco, CA 94110, USA. Phone: +1–415–4373049. Fax: +1–415–4373020. Email: sukyung.chung@ucsf.edu.

Marisa Elena Domino, PhD, Department of Health Policy and Administration, School of Public Health, University of North Carolina at Chapel Hill, 1104G McGavran-Greenberg Hall, CB#7411, Chapel Hill, NC 27599-7411, USA. Email: domino@unc.edu.

Elizabeth W. Jackson, PhD, Innovation, Research and Training, Inc., 1415 NC Highway 54 West, Building 300, Suite 121, Durham, NC 27707, USA. Email: ejackson@mindspring.com.

Joseph P. Morrissey, PhD, Cecil G. Sheps Center for Health Services Research, University of North Carolina at Chapel Hill, 725 Martin Luther King Jr. Blvd., CB#7590, Chapel Hill, NC 27599-7590, USA. Email: joe_morrissey@unc.edu.

Journal of Behavioral Health Services & Research, 2007.

Introduction

Self-reported service use data are used extensively in health services research because they provide comprehensive information on a variety of services, and yet, are relatively inexpensive to obtain.[1,2] Self-reported data are particularly useful in behavioral health services research where any single source of provider records cannot describe all of the services received because of the number and wide variety of provider types available for treatment. Persons with behavioral health problems use a variety of services not only in conventional clinical settings, but also in conjunction with welfare programs and often in the criminal justice systems.[3] Typically, administrative records can only provide information about services received within one agency or organization, whereas a person's self-report ideally would include services received in all settings, from all providers. In an economic evaluation study taking a societal perspective, self-reported service use data could be more comprehensive and valid than information available from provider or insurer records, which may only represent service use from the perspective of the administrative party who maintains the data.[4]

The usefulness of self-reported service use data depends on the validity and reliability of the measurements.[5] Much attention regarding psychometric properties of self-reported health service use has been given to the *validity* of these measures. Generally, validity studies compare self-reported service use against administrative records and show inconsistent findings. Some show favorable level of congruency between data from the two sources,[6,7] but others do not.[8–10] Disagreements are often ascribed to different ranges of services represented in the two data sources, errors in the administrative data such as incomplete recording, and recall bias in self-reporting.[7–10] To compensate for problems with both sources and to produce comprehensive and valid utilization data, hybrid method can be used by collecting self-reported data using a brief measure of provider contact, such as the Brief Health Services Questionnaire, and then retrieving provider records for detailed health care use information.[7,12] Even in this method, however, obtaining reliable answers from self-reporting is prerequisite for further collection of information from provider records.

Our study focuses on the reliability of self-reported services use. Assessment of the *reliability* (or consistency) of self-reported health service use requires repeated measures for each survey item (or test–retest data), preferably by the same rater, within a reasonably narrow time window. Partly because of the lack of appropriate data, there is a paucity of evidence on reliability of self-reported service use. Test–retest is a commonly used method in psychology and education research to assess the reliability of survey items. In the health services literature, test–retest has been frequently applied to examine the reliability of self-reported health status.[11,13]

We identified eight studies that examined reliability of self-reported health service use.[14–21] All but one study[21] were conducted on small samples of focused populations and in a particular geographic area: six studies tested similar instruments, which were designed to elicit responses from parents and children about which services children used[14–20]; and one study examined the use of typical medical services (inpatient, outpatient, and emergency room) among persons with schizophrenia.[19] All the previous studies, except two, used samples drawn from either medical or mental health service settings to ensure enough service use to make test–retest reliability results valid.[14–18] Two studies, which used community samples rather than clinical samples, used limited measures: one examined *any use* of health services with a high school sample,[20] and another examined *any use* of preventive screening services in a small subsample from a national survey.[21]

Taken together, the existing studies on reliability of self-reported services use consistently reported substantial agreement for *any use* (yes/no) of service for specific service types, and fair to moderate agreement for the quantity of service for specific service type. By service type, reliability of reporting was higher for inpatient care than outpatient-based services and higher for aggregate service categories than more specific service categories. However, self-reporting of more specific information, such as frequency of outpatient services provided by mental health professionals, tended to be less reliable.

Regarding determinants of reliability, studies document that question factors, such as sentence complexity, recall period (time between events and reporting), and service types are more important than individual characteristics, such as age, gender, and ethnicity.[16,18,20] None of the previous studies has examined reliability of specific content of services received or evaluated the impact of inconsistency in reporting on evaluation study outcomes, which were explored in the present study.

The present study examines the reliability of self-reported service use among women with behavioral health problems and extends the population studied and scope of analyses beyond that of previous studies. First, the study participants are from a population not studied before, and they were recruited from nine sites nationwide representing diverse geographic areas. Second, this study examines instruments measuring diverse dimensions of service use including *any use* as a binary variable, *quantity* of use for service users, and *content* of service in terms of the focus of treatment for each service type. Third, the survey instrument from this study captures a comprehensive range of services including typical inpatient and outpatient care, residential treatments, and jail or shelter use from which a significantly large fraction of participants with behavioral health problems received services.[3,22] The reliability of self-reporting for services received from these atypical sectors is unknown. Fourth, this study is the first to examine factors influencing consistent reporting and to explore the impact of inconsistency of reporting on the robustness of overall cost estimates.

The specific research questions this study seeks to answer are: (1) What is the test–retest reliability of self-reporting on quantity and content of service use in a variety of settings where health care services are provided? (2) What are the determinants of the consistency of self-reported service use considering such factors as service type, level of service use, severity of psychiatric symptoms, study site, and demographic characteristics? and (3) How sensitive are cost estimates to inconsistency in the quantity of self-reported service use with repeated measures?

Methods

Data

The test and retest data used in this study come from the baseline survey of the Women, Co-occurring Disorders, and Violence Study (WCDVS) conducted in nine sites nationwide from 2001 to 2003. The study participants were women with psychiatric and substance abuse disorders and histories of interpersonal violence. The WCDVS is a quasi-experimental study with an intervention arm that provided comprehensive, integrated, trauma-informed, and consumer/survivor/recovering person-involved care, and a comparison arm that provided usual care in each of the nine sites. Other details about the WCDVS study design have been reported previously.[23] Because retest data were collected at baseline, before the intervention was executed, no intervention effect is anticipated in the present study.

The retest sample included 8% (n = 186) of all study participants (n = 2,729) at baseline. There were approximately equal numbers of participants from each study site and in both intervention and comparison arms. The retest participants were randomly selected and hence display characteristics similar to the other WCDVS participants (Table 1). The retest interview (*retest* hereafter) was conducted during an average of 7 days (s.d. = 4.2; range = 2–35 days) after the initial interview (*test* hereafter) by the interviewer who conducted the test and used the same set of survey items as those used in the test. One exception was that the retest used the identical recall period that was used for the test (i.e., previous 3 months from the original baseline interview date) for the service use questions. All the survey questions were read and answers were recorded by interviewers during in-person interviews.

Table 1
Sample characteristics[a]

Characteristics[b]	Frequency/mean (s.d.) [range][c]
Age	37 (8.4) [19–59]
Race[d]	
White	56%
Black	32%
Other race[e]	13%
Education level	
Less than high school education	35%
High school graduates or equivalent	35%
Some college, technical school or more	30%
Marital status	
Married or partnered	32%
Divorced, separated or widowed	37%
Never married	31%
Insurance status[d]	
Medicaid	49%
Other insurance[f]	20%
No insurance	37%
General mental health symptoms index: global severity index (GSI)	1.4 (0.84) [0–4]
Number of days between test and retest	7.1 (4.2) [2–35]

[a]Statistics are based on test data (n=186)
[b]The retest subsample has similar sample characteristics to the overall WDCVS sample (n=2,729) with no statistical difference at 5% significance level. One exception is the proportion of people with Medicaid (56.7% in the overall sample; p=0.02).
[c]Standard deviations and range of values are presented for continuous variables.
[d]Respondents could choose two or more categories.
[e]Hispanic, Native American, Asian, and Pacific Islander were collapsed into the "other race" category.
[f]Medicare, CHAMPUS, other public insurance and private insurance were collapsed into the "other insurance" category.

The characteristics of the study sample, described in Table 1, are based on the test data. The study sample consists of women aged 19–59 with a mean age of 37 and represents diverse racial groups, education levels, marital status, and insurance status. General psychological distress level was measured using the Global Severity Index (GSI). GSI is the average score of Brief Symptom Inventory (BSI),[24] a 53-item self-report scale (ranging from 0 to 4; higher scores indicating greater severity) measuring nine psychiatric symptom dimensions. The average GSI, 1.4, was quite high, which reflects the fact that all the women in our study sample had complex behavioral health problems.

Variables

Study participants were asked to report frequency and content of services they received during the last 3 months in a variety of categories (Table 2). The survey instruments capture all the services received and are not limited to services received at the participating study site. For each service type, respondents were asked if they received any service. If positive, respondents were asked to answer questions on the frequency of service use. The frequency of emergency room visits and the

Table 2

Test–retest reliability of self-reported service on any use and quantity of service use

	Any use	Quantity of service use			
Servicer type	**Frequency[a]**	**Kappa**	**Unit**	**Mean[a] (s.d.)**	**ICC**
Hospitalization	19%	0.88	Day	1.94 (6.12)	0.89
Emergency room	33%	0.83	Visit	0.60 (1.22)	0.89
Detoxification	25%	0.87	Day	3.16 (9.83)	0.64
Homeless or domestic violence shelter	16%	0.80	Day	5.54 (17.1)	0.93
Jail	20%	0.94	Day	7.27 (21.0)	0.85
Residential individual counseling	35%	0.68	Session[b]	2.22 (4.82)	0.27
Residential group counseling	48%	0.80	Session[b]	9.32 (15.8)	0.81
Any residential counseling	50%	0.83	Session[b]	11.50 (19.3)	0.74
Outpatient individual counseling	55%	0.79	Session[b]	10.05 (11.3)	0.62
Outpatient group counseling	33%	0.74	Session[b]	9.29 (19.0)	0.82
Outpatient medical visit	61%	0.65	Session[b]	7.14 (18.9)	0.12
Any outpatient counseling or medical visit	83%	0.71	Session[b]	27.50 (33.7)	0.61

[a]Statistics based on test data (n=186)

[b]Service quantity is calculated based on answers on frequency category: 1=daily; 2=a few (2–4) times a week; 3=about once a week; 4=2–3 times a month; 5=once a month; 6=only once

number of days in inpatient (or overnight stay) facilities were requested with open-ended questions. For counseling sessions or outpatient visits, frequency categories were used instead of open-ended numeric responses. Each frequency was converted into the total number of visits or sessions during 3 months for the analysis. "Daily" was converted into five times a week or 65 times for the 3 months; "a few or two to four times a week" into three times a week or 39 times for the 3 months; and "two to three times a month" into 2.5 times a month or 7.5 times for the 3 months. We also used the original categorical scale and found similar results, and thus presented results based on the continuous scale throughout this paper. For all service types, respondents were also asked about the content of services received for each stay, visit or session and could choose one or more of the relevant categories. Figure 1 demonstrates how quantity and content of services were measured and coded.

For the reliability of *quantity* of service use, the ten service types examined are hospital, emergency room, detoxification, individual and group counseling at residential and outpatient facilities, outpatient medical visits, homeless or domestic violence shelters, and jail. The frequency of service use was generally high, ranging from 19% for hospital to 61% for outpatient medical visit during the last 3 months. The average number of hospital days was 1.9 and the average number of individual counseling sessions in a residential facility was 2.2. See Table 2 for the frequency and intensity of other types of service use.

For the reliability of *content* of service, the four content areas examined are physical health, mental health, substance abuse, and trauma in each of the five service types: individual and group counseling during residential stay and outpatient visits and outpatient medical visits.

Analysis

Agreement between test and retest data on service use is indexed for dichotomously coded services by Cohen's kappa statistic (k)[25] and for continuous-scaled measures of service use by the

Figure 1

Examples of responses to WCDVS interview questions on service use

1. In the past three months have you been hospitalized overnight or did you receive any type of inpatient care for any reason? †

___ No *[skip to (2) (emergency room section)]*

_V_ Yes

⇓

a. Where were you hospitalized?	b. Which of the following types of treatment or services did you receive when you were hospitalized? ‡ 1 = Physical complaint, injury or medical treatment 2= Violence/Abuse/Trauma treatment 3= Emotions, nerves, or psychological problem 4= Psychiatric medication check 5= Alcohol/other drug abuse treatment 6= Parenting support services 7= Legal assistance 8= Hosing or assistance in finding a place to live 9= Educational or vocational/job assistance	c. How many days were you in the hospital or inpatient facility?
A hospital, NY city, NY	*1, 3*	*2*
A hospital, NY city, NY	*2*	*4*
B hospital, NY city, NY	*1, 2*	*3*

....

10. **In the past three months did you receive any type of medical clinic or doctor's office services?**

___ No *[skip to (11) (medication section)]*

_V_ Yes

⇓

a. What is the name of the clinic or doctor and where did you receive the services?	b. Which of the following types of treatment or services did you receive during the clinic or doctor's visits? *[same choice set as 1.b]*	c. How often did you receive the clinic or doctor's services? 1= daily (5-7 days per week) 2= a few (2-4) times a week 3= about once a week 4= 2-3 times a month 5= once a month 6= only once	d. On average, how long was each visit? 1= Less than 15 minutes 2= 15 – 29 minutes 3= 30 – 59 minutes 4= 1 – 2 hours 5= More than 2 hours
Dr. Smith, NY city, NY	*1, 4*	*3*	*2*
Dr. Johnson, NY city, NY	*2*	*6*	*3*

....

†Interviewers can also record either 98=Refused or 99=Don't Know. In the present study, we recode 98 or 99 as zero to create an aggregate frequency and a representative content of service use variable for each service type. ‡In the interview, this question was assisted by a showcard that lists nine choices; options 98 and 99 were not listed on the card. In the present study, we classify answers for "1=physical complaints, injury or medical treatment" as *physical health*, either "3=emotions, nerves, or 4=psychological problem or psychiatric medication check" as *mental health*, "5=alcohol/other drug abuse treatment" as *substance abuse*, and "6=violence/abuse/trauma treatment" as *trauma.*

intraclass correlation coefficient (ICC).[26] Following the method proposed by Shrout,[27] we interpret kappa and ICC below 0.1 to represent no agreement; 0.1 to 0.39 as slight agreement; 0.4 to 0.59 as fair agreement; 0.6 to 0.79 as moderate agreement; and above 0.8 as substantial agreement. The reliability of self-reporting is assessed with the magnitude rather than statistical significance of these indices. Although *k* and ICC are by far the most widely used indices, they are not free from limitations. These indices are influenced by the variability of event frequency and could be upwardly biased for seldom used services.[28,29] However, this potential bias might not cause serious

problems given that frequencies of events in our sample were relatively high even for less frequently used services such as hospitalization (19%) and jail use (20%).

Multivariate regressions are used in the analysis of determinants of consistency of reporting. A logit model is used for the dependent variable indicating consistent reporting of *any use* of services in test and retest data (1 = agreement), and a linear regression model is used for the agreement rate in *quantity of use* among any users. This study defines agreement rate as $1-|(N_T - N_R) / (N_T + N_R)|$, following a method similar to that used in the literature,[6,30] where N is the total number of visits or stays, T is test, and R is retest. The agreement rate ranges from 0 (none) to 1 (perfect). The number of observations for the *any use* model (n = 1,820) is the number of respondents with valid answers for all covariates (n = 182) multiplied by the number of service types (n = 10). Those who reported non-zero response in either test or retest for each service type are used for the analysis of *quantity of use* (n = 725). Factors examined are service type, time interval between test and retest, study sites, total number of visits or days of 10 service types as a proxy of utilization level, GSI at the test interview, age, race, marital status, education level, and insurance type. Service type is coded with dummy variables with a hospital day as the reference category. Huber-White cluster-adjusted robust standard errors are used to correct for individual clustering across service types.

Finally, we estimate cost for each service type and overall service cost using test and retest data to examine whether the cost estimates drawn from two datasets differ because of the potential inconsistency in reporting. The estimate of unit cost of each type of service is from diverse sources and approximates the societal perspective as is described elsewhere.[31] To draw statistical inferences between estimates from test and retest data, standard errors are calculated by bootstrapping with 500 replications with replacement.

Results

Test–retest reliability of quantity of service use

The test–retest reliability of self-reporting on *any use* of each category of service is generally good. The levels of agreement are moderate to substantial across all service categories (k = 0.65–0.94), highest for jail days and lowest for outpatient medical visits (Table 2). Agreement on the *quantity of service use* is lower than agreement on any use, ranging from slight (ICC = 0.12) for outpatient medical visits to substantial (ICC = 0.93) for shelter days.

The reliability of the total number of days in inpatient facilities is substantial for all the services except for detoxification days (ICC = 0.64). Individuals may not distinguish services received through detoxification from those received in residential facilities. When these two categories are combined, the reliability of service quantity improves (ICC = 0.79; result not in the table), but remains at a moderate level.

The number of residential or outpatient counseling sessions and outpatient medical visits show slight to substantial agreement (ICC = 0.12–0.82). The reliability of reporting is higher for counseling services received in outpatient settings than for counseling services received during residential treatment, and is lowest for outpatient medical visits. When aggregated, the reliability of any outpatient visit and any residential counseling is moderate (ICC = 0.61, 0.74, respectively). We repeated all the analysis with the two subgroups above and below the median level of GSI, 0.76, and found no noticeable difference between the two groups.

Test–retest reliability of content of service use

The reliability of self-reported service *content* during residential or outpatient counseling and outpatient medical visits ranges from none to moderate (k = −0.06–0.79) (Table 3). Generally, the reliability of reporting on the content of services received during counseling sessions is higher for mental health (k = 0.56–0.77) or substance abuse (k = 0.52–0.75) than for trauma (k = 0.45–0.60)

Table 3

Patients perceptions on service contents—mental health/ substance abuse/ trauma

	Physical health		Mental health		Substance abuse		Trauma	
Service type	**Frequency[a]**	**Kappa**	**Frequency[a]**	**Kappa**	**Frequency[a]**	**Kappa**	**Frequency[a]**	**Kappa**
Residential individual counseling	4%	−0.06	26%	0.58	25%	0.52	13%	0.45
Residential group counseling	5%	0.19	31%	0.56	46%	0.75	26%	0.52
Any residential counseling	8%	0.24	37%	0.64	47%	0.79	31%	0.55
Outpatient individual counseling	5%	0.13	48%	0.77	28%	0.53	24%	0.49
Outpatient group counseling	7%	0.40	25%	0.66	26%	0.72	16%	0.60
Outpatient medical visit	56%	0.63	18%	0.59	4%	0.12	2%	0.32
Any outpatient counseling or medical visit	60%	0.64	61%	0.78	39%	0.67	31%	0.55

Number of observations: 186

[a]Statistics based on test data

or physical health (k = −0.06–0.40). An exception is that for outpatient medical visits, physical health is more consistently reported (k = 0.63) than other content areas (k = 0.12–0.59). The reliability of reporting of service content is higher for services received during outpatient visits (k = 0.13–0.77) than for those received during residential treatment (k = −0.06–0.75). Again, the reliability increases only slightly when aggregate categories (i.e., any residential counseling, any outpatient visit) are used.

Determinants of the consistency of self-reporting

We find few observable factors that are associated with consistent reporting. For *any use*, counseling services and outpatient medical visits are less likely to be consistently reported than hospital use, after controlling for other relevant factors (Table 4). For *quantity of use*, only the number of outpatient medical visits is less consistently reported than hospital days. Consistency of reporting also varies across study sites, which may reflect differences in the research staff who conducted the interviews. White race (vs. other race) and some college education (vs. less than high school) are associated with more consistent reporting in quantity of use. None of the following factors affects consistency of reporting: level of mental distress (GSI), level of service use in aggregate, and time interval between test and retest.

Robustness of cost estimates from test and retest data

The average total cost estimates from test and retest data are $9,168 (s.d.: 10,128) and $8,883 (s.d.: 10,243), respectively (Table 5). Note that the variances in costs are quite large, which is typical for cost data particularly among high-end users of health services. Rather than excluding extremely high cost users using an arbitrary cut-point, we used standard errors from bootstrapping to address the issues of relatively skewed distribution and small sample size. The mean difference in total cost ($285) is only 3.2% of the average total cost and is not statistically different from zero. By service type, hospital costs ($2,772; 30%) and residential treatment costs ($1,800; 20%) comprise a majority of total costs in the test data.

Discussion

This study adds to the literature on the reliability of self-reported service use by extending the population studied and the scope of analyses. Studies on children in the community[14–18,20] and on persons with schizophrenia[19] have shown substantial agreement in reporting any use of service and fair-to-moderate agreement in reporting quantity of services. Consistent with the previous findings, this study shows moderate to substantial agreement for any use and slight-to-substantial agreement for quantity of services, among women with behavioral health problems.

The wide variation in reliability by service types is notable. Quantity of service is more consistently reported for inpatient days than for outpatient visits, maybe because inpatient stay is a more salient episode and thus easier to remember than outpatient visits. On the other hand, quantity of counseling services is more consistently reported for services received during outpatient visits than for services received during residential treatment. The treatments received during residential stay are so complex that service recipients may be difficult to discern specific treatment elements. It is also likely that the frequency of receiving specific services while staying in residential facilities might be harder to remember than the frequency of visits to outpatient facilities that require more effort and time to attend. We also found that reliability improves by aggregation of service categories, which suggests that a lower level of details would be easier to remember and answer consistently.

Table 4

Predictors of consistency in reporting: any use and frequency of service use[a,b]

	Consistent answer for any use: Logit Odds ratio (z-stat)	**Agreement rate in the quantity of use: OLS[c]** Coefficients (t-stat)
Service type (referent category: hospital)		
Emergency room	0.438 (1.76)	0.050 (0.69)
Detoxification facility	0.766 (0.50)	0.061 (0.81)
Homeless or domestic violence shelter	0.684 (0.72)	0.057 (0.61)
Jail	1.792 (0.89)	0.117 (1.88)
Residential individual counseling	0.201** (3.46)	0.075 (1.14)
Residential group counseling	0.334* (2.28)	0.070 (1.07)
Outpatient individual counseling	0.315** (2.76)	−0.075 (1.15)
Outpatient group counseling	0.282** (2.90)	−0.098 (1.18)
Outpatient medical visit	0.193** (3.80)	−0.139* (2.00)
Total number of days or visits	0.999 (0.54)	−0.000 (1.50)
GSI	1.032 (0.28)	0.033 (1.78)
Days between test and retest	1.004 (0.16)	−0.001 (0.50)
Age	1.004 (0.36)	0.0005 (0.24)
White	1.111 (0.46)	0.096** (2.63)
Married	1.182 (0.86)	0.004 (0.15)
Education (referent category: <high school)		
High school graduate	1.270 (1.12)	0.066 (1.78)
Some college education	0.906 (0.38)	0.113** (3.02)
Insurance type (referent category: no insurance)		
Medicaid	0.943 (0.33)	0.025 (0.85)
Other insurance	1.297 (0.85)	−0.028 (0.70)
Constant		0.570** (5.11)
No. Observations	1820	725
Log likelihood	−501.15	
R-squared		0.158

*Significant at 5%

**Significant at 1%

[a]Robust *z* (logit) and *t* (OLS) statistics based on the standard errors adjusted for the individual clustering are in parentheses.

[b]Eight site dummies are not reported in the table. One site is significant in the logit model and another one site is significant ($p<0.01$) in the OLS model.

[c]Agreement rate is defined as $1-|(N_T-N_R) / (N_T+N_R)|$, where *N*: frequency; *T*: test data; *R*: retest data

A confounding factor that might have influenced the reliability of counseling services and medical visits is the wording of the question. Frequencies of these services were elicited by *fixed* categories for the *average* frequency of service use per week or per month during the previous 3 months versus *open-ended* questions about the *total* frequency during the previous 3 months for other service types (See Fig. 1 for an example of each). Therefore, the difference in reliability may partly come from the difference in question format (i.e., categorical vs. open-ended). Furthermore, because counseling and outpatient visits are high-frequency events, an inconsistency in the answer

Table 5
Cost estimates in test and retest data

	Mean (s.d.)[a]	
	Test data (n=186)	**Retest data (n=186)**
Average total cost ($)	9,168 (10,128)	8,883 (10,243)
Average cost by service type ($)		
Hospital	2,772 (8,921)	2,802 (8,957)
Emergency room	264 (538)	250 (552)
Detoxification	641 (1,995)	517 (1,518)
Homeless or domestic violence shelter	349 (1,076)	368 (1,093)
Jail	462 (1,393)	422 (1,231)
Residential care	1,800 (2,807)	1,955 (2,847)
Outpatient counseling	1,327 (2,035)	1,237 (1,874)
Outpatient medical visit	721 (1,987)	547 (1,284)
Other services[b]	833 (1,261)	783 (1,053)

[a]For any category, difference in costs between test and retest data is not statistically different from zero based on standard errors calculated by bootstrapping with 500 replications.
[b]Included are prescription drug, peer support group and outpatient case management.

for one category may result in a large difference in the total frequency over the 3-month period. With this survey design, the variation in reliability ascribed to different question formats could not be teased apart from the variation ascribed to different service types.

This study provides novel evidence on the reliability of self-reported *content* of care received during counseling services and medical visits. The reliability of service content is generally lower than the reliability of service quantity, and is below the acceptable level (k<0.4) for some categories. Particularly of concern is the lowest reliability of reporting on service content during medical visits, which is the most common type of service relying on self-reporting in health services research. People with behavioral health problems receive a variety of services and therefore may have difficulty in differentiating services focusing on behavioral health from those addressing comorbid physical health problems during medical visits.

We find no evidence of an association between severity of psychiatric symptoms, measured by the GSI, and the consistency in reporting. This is consistent with the findings of other studies,[6,33,34] which reported validity or reliability of self-reporting was not influenced by the severity of psychiatric conditions. On the other hand, type of illness or symptomatology may influence reliable reporting because of cognitive deficits associated with some psychiatric conditions. We were not able to investigate the variation across different symptomatology because of the limited sample and data on diagnostic information. Previous studies have shown that self-reported health behavior or services use among persons with severe mental illness or substance abuse problems are also reliable and valid.[19,35–39] This suggests that there would be little influence on reliability of reporting because of cognitive deficit associated with psychiatric conditions.

For both any use and quantity of service, the total number of outpatient medical visits is significantly less likely to be consistently reported than hospital days, which cautions against the wide use of self-reporting in measuring the frequency of outpatient medical visits. These results are consistent with the literature on determinants of the *validity* of self-reporting, which indicates that the saliency of events and well-defined (vs. ambiguous) events accounts for more of the variance in response accuracy than any other class of variables.[5]

The overall findings on the determinants of consistency of reporting suggest that factors associated with the survey administration are more important than those representing subject characteristics. Similar findings were reported in previous studies on health services use among children.[18,20] These findings are also consistent with the literature that indicates task factors, such as question form, wording, and mode of administration, account for more of the variance in response accuracy than any other class of variables.[32] On the other hand, it is noteworthy that a large proportion of the variation across repeated measures was not explained by the variables in the model, as indicated by relatively low *R*-square (0.16). More detailed information on individual and service and availability of different question forms would help increase our understanding of determinants of reliable measures of services use.

One of the important applications of self-reported service data is in economic evaluation research. Our results show that although reliability varies across service types, the aggregated cost estimate for overall service use is robust across repeated measures. This robustness is partly because reliability of reports of quantity of use is higher for the more intensive and costly services, such as hospital use. The less consistently reported services, such as outpatient medical visits, constitute a small proportion of total cost for the population of this study.

In interpreting our results, one should be careful in generalizing our study findings to population or settings different from ours. Reliability of self-reporting among our study participants (women with behavioral health problems) could be different from the reliability among other groups of behavioral health service users. Furthermore, the results indicated by kappa and ICC indices measured for different populations might not be directly comparable.

Based on the findings and the limitations of this study, we suggest several areas for further research to better understand the reliability of self-reporting of health service use. First, future research should explore the relationship between question phrasing (e.g., open-ended vs. close-ended) and response reliability of quantity of service use and between service type specification (e.g., residential treatment vs. specific type of services during residential treatment) and reliability of treatment content during the service use. Such research would help in developing survey instruments that induce more reliable data on service quantity and content during health care services. Second, future study may also consider aided recall to stimulate the memory for specific events. For example, providing a motivation to remember or contextual cues may considerably improve reliability and validity of recall because the vicissitudes of memory are common to both the test and retest data, particularly among clients with complex treatments or with cognitive deficits. Similarly, a shorter recall time frame than the 3 months in this study may improve the reliability of reporting. Third, the reliability of self-reported service use in other populations with different ranges or levels of services, such as clients with physical health problems or clients of primary care, would help in assessing generalizability of our findings. Finally, further study should examine the validity of self-reported service use data in populations similar to ours. A review of provider records and other objective and unobtrusive measures would be valuable in checking the validity of client's self-reporting. Such evidence is essential to understand psychometric properties of self-reported data in populations similar to ours and will allow for the comparison of validity of self-reported services use among diverse populations.

Implications for Behavioral Health

Although self-reported data are widely used in assessing health service use, evidence on the quality of the data, particularly on the reliability of reporting, is very limited. Findings of our study suggest that among individuals with behavioral health problems, self-reported health service use data are reliable in capturing the quantity of services received in a variety of service areas. However, self-reporting of treatment content in highly specified service categories (e.g., individual counseling during residential treatment) may not be reliable. Similarly, the low level of reliability for the quantity

of service use and content of service during outpatient medical visits, the most common medical events, needs attention. To determine the quantity and content of service use during general medical visits, physician records may be a better alternative than participant responses.

Despite some lack of agreement in reports of quantity and content of services, cost estimates did not vary with repeated measures and were unaffected by the inconsistency in reporting. Self-reported service use data produce robust cost estimates in aggregate and have the unique advantage of encompassing comprehensive types of service use. Therefore, self-reported service use data can serve as a useful source of information for the economic evaluation of behavioral health service programs.

Our findings on determinants of consistent reporting suggest that reliability of reporting varies widely by service types and may be improved with better measurements or administration methods, but may not be sensitive to respondent characteristics such as demographics and disease severity. However, more evidence from using different survey instruments, study populations, and study settings is needed to generalize our study findings to a broader behavioral health service context.

Acknowledgment

This study was funded by the grant, number TI-00-003, from Substance Abuse and Mental Health Services Administration's three centers: the Center for Substance Abuse Treatment, the Center for Mental Health Services, and the Center for Substance Abuse Prevention. This grant was entitled "Cooperative Agreement to Study Women with Alcohol, Drug Abuse and Mental Health Disorders who have Histories of Violence: Phase II". The abstract of this paper was presented at the Academy Health's Annual Research Meeting held June 26–28, 2005, at Boston, MA. Additional support was received by Dr. Chung and Dr. Domino from the National Institute of Mental Health: T32-MH-0182-61 and K01-MH-0656-39, respectively.

The assistance of project staff at the following participating sites (listed in alphabetical order by state) is gratefully acknowledged: Los Angeles, California: PROTOTYPES Systems Change Center, Vivian Brown, Principal Investigator; Stockton, California: Allies: An Integrated System of Care, Jennie Heckman, Principal Investigator; Thornton, Colorado: Arapahoe House – New Directions for Families, Nancy Van DeMark, Principal Investigator; Washington, DC: District of Columbia Trauma Collaboration Study, Roger Fallot, Principal Investigator; Avon Park, Florida: Triad Women's Project, Margo Fleisher-Bond, Co-Principal Investigator, Colleen Clark, Co-Principal Investigator; Boston, Massachusetts: Boston Consortium of Services for Families in Recovery, Hortensia Amaro, Principal Investigator; Cambridge, Massachusetts: Women Embracing Life and Living (WELL) Project, Norma Finkelstein, Principal Investigator; Greenfield, Massachusetts: Franklin County Women's Research Project, Rene Andersen, Principal Investigator; New York, New York: Portal Project, Sharon Cadiz, Principal Investigator. The Coordinating Center is operated by Policy Research Associates (PRA), located in Delmar, New York, in coordination with the National Center on Family Homelessness of Newton, Massachusetts and the Cecil G. Sheps Center for Health Services Research at the University of North Carolina at Chapel Hill, North Carolina. The interpretations and conclusions contained in this publication do not necessarily represent the position of the WCDVS Coordinating Center, participating study sites, participating Consumer/Survivor/Recovering persons, or the Substance Abuse and Mental Health Services Administration.

References

1. Ganiats TG, Sieber WJ, Weisman M. Health-related quality of life. *Best Practices and Benchmarking in Healthcare.* 1997;2:57–62.
2. Brown JB, Adams ME. Patients as reliable reporters of medical care process: recall of ambulatory encounter events. *Medical Care.* 1992;30:400–411.
3. Veysey BM, Steadman HJ, Morrissey JP, et al. In search of the missing linkages: continuity of care in U.S. jails. *Behavioral Sciences & the Law.* 1997;15(4):383–397.

4. Hargreaves W, Shumway M, Hu TW, et al. (eds). *Cost-outcome Methods for Mental Health*. San Diego: Academic Press, 1998.
5. Del Boca FK, Noll JA. 2000 Truth or consequences: the validity of self-report in health services research on addictions. *Addiction*. 2000;95:347–360.
6. Rozario PA, Morrow-Howell N, Proctor E. Comparing the congruency of self-report and provider records of depressed elder's service use by provider type. *Medical Care*. 2004;42:952–959.
7. Booth BM, Kirchner JE, Fortney SM, et al. Measuring use of health services for at-risk drinkers: how brief can you get. *The Journal of Behavioral Health Services & Research*. 2006;33:254–264.
8. Fendrich M, Johnson T, Wislar JS, et al. Accuracy of parent mental health service reporting: results from a reverse record-check study. *Journal of the American Academy of Child and Adolescent Psychiatry*. 1999;38(2):147–155.
9. Rhodes AE, Fung K. Self-reported use of mental health services versus administrative records: care to recall. *International Journal of Methods in Psychiatric Research*. 2004;13(3):165–175.
10. Beebe TJ, McRae JA, Barnes SA. A comparison of self-reported use of behavioral health services with Medicaid agency records in Minnesota. *Psychiatric Services*. 2006;57(11):1652–1654.
11. Klabunde CN, Reeve BB, Harlan LC, et al. Do patients consistently report comorbid conditions over time? Results from the Prostate Cancer Outcomes Study. *Medical Care*. 2005;43:391–400.
12. Netert PJ, French MT, Kirchner J, et al. Health services utilization and cost for at-risk drinkers: rural and urban comparisons. *Journal of Studies on Alcohol*. 2004;65(3):352–362.
13. Beckett M, Weinstein M, Goldman N, et al. Do health interview surveys yield reliable data on chronic illness among older respondents. *American Journal of Epidemiology*. 2000;151:315–323.
14. Horwitz SM, Hoagwood K, Stiffman AR, et al. Reliability of the services assessment for children and adolescents. *Psychiatric Services*. 2001;52(8):1088–1094.
15. Hoagwood EK, Jensen PS, Arnold LE, et al. Reliability of the services for children and adolescents-parent interview. *Journal of the American Academy of Child and Adolescent Psychiatry*. 2004;43(11):1345–1354.
16. Canino G, Shrout PE, Alegria M, et al. Methodological challenges in assessing children's mental health services utilization. *Mental Health Services Research*. 2002;4(2):97–107.
17. Bean DL, Rotheram-Borus MJ, Leibowitz A, et al. Spanish-language services assessment for children and adolescents (SACA): reliability of parent and adolescent reports. *Journal of the American Academy of Child & Adolescent Psychiatry*. 2003;42(2):241–248.
18. Farmer E, Angold A, Burns B, et al. Reliability of self-reported service use: test–retest consistency of children's responses to the Child and Adolescent Services Assessment (CASA). *Journal of Child and Family Studies*. 1994;3(3):307–325.
19. Goldberg RW, Seybolt DC, Lehman A. Reliable self-report of health service use by individuals with serious mental illness. *Psychiatric Services*. 2002;53:879–881.
20. Santelli AJ, Klein J, Graff C, et al. Reliability in adolescent reporting of clinician counseling, health care use, and health behaviors. *Medical Care*. 2002;40(1):26–37.
21. Nelson DE, Holtzman D, Bolen J, et al. Reliability and validity of measures from the Behavioral Risk Factor Surveillance System (BRFSS). *Sozial- und Präventivmedizin*. 2001;46(S1):S3–S42.
22. Roberts AR. The organizational structure and function of shelters for battered women and their children: a national survey. In: Roberts AR, ed. 2nd edn. *Battered Women and their Families*. New York, NY: Springer Publishing; 1998:58–75.
23. McHugo GJ, Kammerer N, Jackson EW, et al. Women, Co-occurring Disorders, and Violence Study: evaluation design and study population. *Journal of Substance Abuse Treatment*. 2005;28(2):91–107.
24. Derogatis LR. *Brief Symptom Inventory (BSI): Administration, Scoring and Procedures Manual*. 4th edn. Minneapolis, MN: National Computer Systems; 1993.
25. Cohen J. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*. 1960;20:37–46.
26. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin*. 1979;86:420–428.
27. Shrout PE. Measurement reliability and agreement in psychiatry. *Statistical Methods in Medical Research*. 1998;7:301–317.
28. Spitznagel EL, Helzer JE. A proposed solution to the base rate in the kappa statistic. *Archives of General Psychiatry*. 1985;42:725–728.
29. Kraemer HC. Ramifications of a population model for k as a coefficient of reliability. *Psychometrika*. 1979;44:461–472.
30. Carsjo K, Thorslund M, Warneryd B. The validity of survey data on utilization of health and social services among the very old. *Journals of Gerontology. Series B, Social Sciences*. 1994;49:S156–S164.
31. Domino ME, Morrissey JP, Chung S, et al. Twelve-month service use and costs for women with co-occurring mental and substance use disorders and a history of violence. *Psychiatric Services*. 2005;56:1223–1232.
32. Schwarz N. Self-reports: How the questions shape the answers. *The American Psychologist*. 1999;54(2):93–105.
33. Clark RE, Ricketts SK, McHugo GJ. Measuring hospital use without claims: a comparison of patient and provider reports. *Health Services Research*. 1996;31:153–169.
34. Kashner TM. Agreement between administrative files and written medical records—A case of the Department of Veterans Affairs. *Medical Care*. 1998;36(9):1324–1336.
35. Calsyn RJ, Allen G, Morse GA, et al. Can you trust self-report data provided by homeless mentally ill individuals. *Evaluation Review*. 1993;17(3):353–366.
36. Tsemberis S, McHugo G, Williams V, et al. Measuring homelessness and residential stability: the residential time-line follow-back inventory. *Journal of Community Psychology*. 2007;35(1):29–42.
37. Klinkenberg WD, Calsyn RJ, Morse GA, et al. Consistency of recall of sexual and drug-using behaviors for homeless persons with dual diagnosis. *AIDS and Behavior*. 2002;6(4):295–307.
38. Nieves K, Draine J, Solomon P. The validity of self-reported criminal arrest history among clients of a psychiatric probation and parole service. *Journal of Offender Rehabilitation*. 2000;30(3–4):133–151.
39. Sohler N, Colson PW, Meyer-Bahlburg HF, et al. Reliability of self-reports about sexual risk behavior for HIV among homeless men with severe mental illness. *Psychiatric Services*. 2000;51(6):814–816.