© Health Research and Educational Trust
DOI: 10.1111/j.1475-6773.2005.00480.x

The Reliability of Survey Assessments of Characteristics of Medical Clinics

Peter V. Marsden, Bruce E. Landon, Ira B. Wilson, Keith McInnes, Lisa R. Hirschhorn, Lin Ding, and Paul D. Cleary

Objective. To assess the reliability of survey measures of organizational characteristics based on reports of single and multiple informants.
Data Source. Survey of 330 informants in 91 medical clinics providing care to HIV-infected persons under Title III of the Ryan White CARE Act.
Study Design. Cross-sectional survey.
Data Collection Methods. Surveys of clinicians and medical directors measured the implementation of quality improvement initiatives, priorities assigned to aspects of HIV care, barriers to providing high-quality HIV care, and quality improvement activities. Reliability of measures was assessed using generalizability coefficients. Components of variance and clinician–director differences were estimated using hierarchical regression models with survey items and informants nested within organizations.
Principal Findings. There is substantial item- and informant-related variability in clinic assessments that results in modest or low clinic-level reliability for many measures. Directors occasionally gave more optimistic assessments of clinics than did clinicians.
Conclusions. For most measures studied, obtaining adequate reliability requires multiple informants. Using multiple-item scales or multiple informants can improve the psychometric performance of measures of organizational characteristics. Studies of such characteristics should report the organizational level reliability of the measures used.

Key Words. Health care organizations, HIV care, reliability, survey research

Rising concern about the quality of medical care and preventable medical errors has increased interest in how systems of care operate. Health care organizations can shape the quality of care through the selection of clinical staff or educational programs for patients. Influencing clinician behavior, however, is arguably the most important way in which organizations affect care (Flood 1994; Landon, Wilson, and Cleary 1998). Organizations can influence clinicians using financial incentives, management strategies (e.g., utilization review, guidelines, profiling), structural arrangements (e.g., presence of particular facilities or domains of expertise, governance structures), and normative practice styles or organizational cultures.

Studies of organizational influences on the quality of care require measures of organizational characteristics that are rarely, if ever, recorded in a standardized way. Organizational data are commonly collected by surveying informants about their organizations. Surveys often ask for factual data such as the number of FTE medical staff or whether professionals with particular specialties are on site. They can also ask about subjective phenomena, such as an organization's culture or mission. Recent examples include Kralewski et al. (2000), who gathered data on revenue sources and methods of physician compensation from clinic medical directors or administrators, and Meterko, Mohr, and Young (2004), who measured hospital culture by surveying hospital employees.

Lazarsfeld and Menzel (1980) distinguish "global" and "analytical" organizational survey measures. Global measures refer to organization-level properties such as size or centralization of decision making. "Analytical" measures are organization-level averages of respondent-level data, such as the proportion of clinicians who are board certified in infectious diseases.

High reliability is necessary but not sufficient for the validity of measurement (Bohrnstedt 1983). Imprecise measurement (low reliability) will sometimes lead investigators to incorrect conclusions about relationships between an organizational factor and outcome measures of interest. Nonetheless, few organizational studies examine the reliability of informant reports.

If informant reliability is low, relying on a single informant per organization may be unwise. Just as using multiple-item scales can improve respondent-level survey measures, combining reports from multiple informants may raise reliability for organizational measurements. Assessing measure reliability can offer guidance about the number of informants needed to adequately measure different organizational properties.

Address correspondence to Peter V. Marsden, Ph.D., Professor, Department of Sociology, Harvard University, 630 William James Hall, 33 Kirkland Street, Cambridge, MA 02138. Bruce E. Landon, M.D., M.B.A., Associate Professor, Keith McInnes, M.S., EQHIV Project Director, Lin Ding, Ph.D., Senior Researcher and Paul D. Cleary, Ph.D., Professor, are with the Department of Health Care Policy, Harvard Medical School, Boston, MA. Bruce E. Landon, M.D., M.B.A., Associate Professor, is also a practicing internist in the Division of General Medicine, Beth Israel Deaconess Medical Center, Boston. Ira B. Wilson, M.D., M.Sc., Associate Professor, is with the Institute for Clinical Research and Health Policy Studies and the Department of Medicine, Tufts-New England Medical Center, Boston, MA. Lisa R. Hirschhorn, M.D., M.P.H., Assistant Clinical Professor of Medicine, is with Harvard Medical School and is Senior Clinical Advisor on HIV/AIDS at JSI Research and Training.

When organizations are the objects of measurement, studies usually can select among several possible informants, so researchers must decide which informants to approach. Standard advice is to seek out informants who are knowledgeable, motivated, and unbiased (Huber and Power 1985). Managerial or administrative informants are often chosen on the assumption that they have good access to information. Such informants, however, also may tend to present the organization positively (Seidler 1974). Studies rarely examine differences in descriptions of an organization between types of informants (e.g., medical directors and physicians).

This article addresses issues of measure reliability and differences across informant types using data from a national study of medical clinics, the Evaluation of Quality Improvement for HIV (EQHIV) study. That study gathered data about clinic characteristics from the clinic director and several clinicians in each practice studied. It asked about implementation and assessment of improvement initiatives, HIV care priorities, and barriers to improvement. We examine the reliability of single-informant organizational measures based on individual survey items as well as multiple-item scales, and how reliability can be improved by using multiple informants. We also calculate the number of informants required to obtain reliable organization-level measures, and assess clinician–director differences in descriptions of a clinic.

Assessing Reliability

Several health care studies have used surveys or interviews with informants to measure organizational characteristics. Studies relying on data from a single informant per organization have examined effects of group practice and payment methods on costs of care (Kralewski et al. 2000) and effects of care management processes on the quality of care (Casalino et al. 2003). Other studies used multiple informants in assessing organizational characteristics and performance in intensive care units (Shortell et al. 1991), long-term care teams (Temkin-Greener et al. 2004), and hospitals (Shortell et al. 1995; Aiken and Sloane 1997; Aiken and Patrician 2000; Meterko et al. 2004). Studies have used both single-item organizational measures (e.g. Aiken and Sloane 1997), and multiple-item scales (e.g. Shortell et al. 1991).

Multiple-informant studies often present one-way analyses of variance (ANOVA) of informant reports classified by organization to support combining informant assessments into organization-level measures. A statistically significant *F* ratio in such an analysis indicates a nonrandom resemblance in reports by informants within a given organization, but does not directly measure the extent of resemblance. The F ratio is sometimes supplemented by the correlation ratio η^2, equivalent to the coefficient of determination (R^2) for regressing an informant report on a set of indicator variables for organizational differences. Like R^2, η^2 can be misleadingly large when there are many indicator variables relative to the total number of reports.

Bohrnstedt (1983) generically defines the reliability of a measure as the ratio of true-score variance to total variance, or alternately as the complement of the ratio of error to total variance:

$$\rho_{\text{Measure}} = \frac{\sigma^2_{\text{True}}}{\sigma^2_{\text{Measure}}} = \frac{\sigma^2_{\text{True}}}{\sigma^2_{\text{True}} + \sigma^2_{\text{Error}}} = 1 - \frac{\sigma^2_{\text{Error}}}{\sigma^2_{\text{Measure}}}. \tag{1}$$

The last expression in (1) shows that reliability is low when error variance is large relative to total variance. The next-to-last expression shows that reliability is also low if variation in a phenomenon (σ^2_{True}) is limited within a given study population.

When several informant reports are available, it is common to use their average

$$\bar{r}_j = \frac{\sum_{h=1}^{n_j} r_{jh}}{n_j} = \frac{\sum_{h=1}^{n_j}\sum_{k=1}^{K} x_{kjh}}{n_j K} \tag{2}$$

as an organization-level measure. In (2) r_{jh} is the report of informant h about organization j and n_j is the number of informants for organization j; J is the number of organizations and $N = \sum_{j=1}^{J} n_j$ is the total number of informants. If $n_j = 1$, (2) is the report r_{jh} of a single informant. The measurement r_{jh} may be a scale averaging K items x_{kjh}; if $K = 1$, r_{jh} is a single item.

When r_{jh} is a scale score, two potential sources of error variation in (2) are distinguishable, measurement error in r_{jh} and error because of informant differences in r_{jh}. Since the object of measurement is the organization, (2) is reliable when organizational variability is high relative to these sources of error. Likewise, the informant-level measure r_{jh} is affected by organizational and informant differences as well as errors of measurement. Assuming that these sources are independent, the variance σ^2_r of r_{jh} is

$$\sigma^2_r = \sigma^2_{\text{o}} + \sigma^2_{\text{i}} + \sigma^2_{\text{e}} \tag{3}$$

where σ^2_{o}, σ^2_{i}, and σ^2_{e} refer, respectively, to organizational, informants-within-organizations, and error components of variance. The variance $\sigma^2_{\bar{r}}$ of the organizational measure $\bar{r}_j$ is then

$$\sigma^2_{\bar{r}} = \sigma^2_{\mathrm{o}} + \frac{\sigma^2_{\mathrm{i}}}{n_j} + \frac{\sigma^2_{\mathrm{e}}}{n_j K} \tag{4}$$

The latter two components of $\sigma^2_{\bar{r}}$ reflect error in $\bar{r}_j$, while σ^2_{o} is reliable organizational variance. Expressing σ^2_{o} as a fraction of $\sigma^2_{\bar{r}}$ yields a generalizability coefficient (O'Brien 1990; Shavelson and Webb 1991) measuring the reliability of $\bar{r}_j$:

$$\rho_{\bar{r}} = \frac{\sigma^2_{\mathrm{o}}}{\sigma^2_{\bar{r}}} = \frac{\sigma^2_{\mathrm{o}}}{\sigma^2_{\mathrm{o}} + \sigma^2_{\mathrm{i}}/n_j + \sigma^2_{\mathrm{e}}/n_j K}. \tag{5}$$

Measure (5) gives the fraction of variance in the organizational measure $\bar{r}_j$ attributable to systematic organizational differences rather than informant variations or measurement error.

If r_{jh} is a single item, informant and error variance are indistinguishable; σ^2_{i} and σ^2_{e} then combine into a single "error" variance component $\sigma^2_{\mathrm{i,e}}$, and the reliability of $\bar{r}_j = \bar{x}_j$ becomes

$$\rho_{\bar{x}} = \frac{\sigma^2_{\mathrm{o}}}{\sigma^2_{\mathrm{o}} + \sigma^2_{\mathrm{i,e}}/n_j}. \tag{6}$$

If, moreover, organizations are measured using a single informant ($n_j = 1$), (6) simplifies further to

$$\rho_x = \sigma^2_{\mathrm{o}}/(\sigma^2_{\mathrm{o}} + \sigma^2_{\mathrm{i,e}}) \tag{7}$$

a quantity known as the intraclass correlation (see, e.g., Scheffé 1959, p. 223).

METHODS

Setting

Title III of the Ryan White Comprehensive AIDS Resources Emergency (CARE) Act administered by the HIV/AIDS Bureau of the Health Resources and Services Administration (HRSA) supports comprehensive primary health care for HIV-infected persons. In 1999, HRSA required that clinical sites newly awarded funding under Title III participate in a quality improvement collaborative conducted by the Institute for Healthcare Improvement (IHI), and invited other Title III clinics to participate. The EQHIV study (Landon et al. 2004) conducted pre- and postintervention surveys of clinicians and medical directors in the participating clinics and a matched set of comparison clinics. Here we examine data from the preintervention surveys conducted between August 2000 and January 2001.

Selection of Sites

Of the 200 Title III sites in the continental United States in May 2000, we excluded 16 reporting HIV caseloads lower than 100 per year, 12 that initially enrolled in the collaborative but did not participate, and one that lost CARE Act funding shortly before the collaborative began. Of the remaining 171 sites, 62 participated in the collaborative, and 54 of those participated in the study and provided survey data. Control sites were matched with intervention sites on type (community health center, community-based organization, health department, hospital, or university medical center), location (rural, urban), number of locations delivering care, region, and number of active HIV cases. Of 40 control sites, 37 participated in the study and provided survey data. The Committee on Human Studies of Harvard Medical School approved the study protocol.

Selection of Informants/Respondents

EQHIV surveyed clinic directors and clinicians to assess clinic and clinician characteristics. Surveys were mailed to the medical director and random samples of up to five clinicians who had primary responsibility for HIV patients. If a site had five or fewer clinicians, all were selected. Completed surveys were returned by 79 medical directors (87 percent response rate) and 300 clinicians (89 percent response rate). At 49 sites, the medical director was also a sampled clinician, and completed both instruments, so there were 330 distinct informants.

Variables and Scales

Survey instruments asked about clinic characteristics such as leadership commitment to quality, quality improvement initiatives, teamwork, patient care priorities, clinic priorities and limitations, and use of computers, as well as individual characteristics including formal education and training, HIV care experience, HIV knowledge, and basic demographic information. We constructed eight scales including items with common substantive content, using guidance from factor analyses. The longest scale (seven items) assessed the organization's openness to quality improvement. Others measured HIV knowledge (six items), research emphasis (three items), clinician autonomy (three items), emphasis on helping patients (three items), stress on guidelines (two items), barriers to quality improvement (five items), and a clinician's patient load (three items).

Analyses

The director and clinician surveys had 15 identical items.[1] As we are concerned with the reliability of measures across multiple informants within organizations, we examined the items answered by clinicians, including responses by directors to identical items.

Assessing the reliability of organization-level measures via (5), (6), or (7) requires estimates of variance components. Estimates were obtained by maximum likelihood using *Stata* (StataCorp 2003) and *GLLAMM* (Rabe-Hesketh, Pickles, and Skrondal 2001).[2]

For K-item scales, we estimated three-level mixed-effects regressions for items nested in informants nested in organizations, including fixed effects for differences in item means:

$$x_{kjh} = \mu_K + \sum\nolimits_{k=1}^{K-1} \beta_k z_{kjh} + \nu_j + \eta_{jh} + \varepsilon_{kjh} \tag{8}$$

where μ_K is the mean for the last item in a scale, z_{kjh} is an indicator variable identifying observations on item k, β_k is the difference in means between item k and item K, υ_j is a random organization effect, η_{jh} is a random effect for informant h within organization j, and ε_{kjh} is a residual term for item-level error. Estimates for σ_{o}^2, σ_{i}^2, and σ_{e}^2 in (5) are variances of the random effects υ_j, η_{jh}, and ε_{kjh}, respectively.

For single items, we estimated random-effects regressions for informants nested in organizations:

$$x_{kjh} = \mu_k + \nu_j + \varepsilon_{kjh} \tag{9}$$

where μ_k is the mean for item k and ε_{kjh} is a residual combining item- and informant-level error. We calculated reliabilities in (6) and (7) using the estimated variances $\hat{\sigma}_{\mathrm{o}}^2$ and $\hat{\sigma}_{\mathrm{i,e}}^2$ of the random effects υ_j and ε_{kjh}, respectively.

With estimates of the variance components, we can calculate the implied number of informants n_j^* required to measure an organizational characteristic at any criterion level of reliability. We set reliability in (5) or (6) at the conventional threshold of 0.70 (Nunnally 1978; Shortell et al. 1991) and solve for n_j. For single items, this leads to[3]

$$n_j^* = \frac{0.7\hat{\sigma}_{\mathrm{i,e}}^2}{0.3\hat{\sigma}_{\mathrm{o}}^2}. \tag{10}$$

The necessary number of informants increases with informant/error variance and the criterion level of reliability, and declines with organization-level variance. For a K-item scale, similar manipulation of (5) yields

$$n_j^* = \frac{0.7(K\hat{\sigma}_{\mathrm{i}}^2 + \hat{\sigma}_{\mathrm{e}}^2)}{0.3K\hat{\sigma}_{\mathrm{o}}^2}. \tag{11}$$

We assessed clinician–director differences in reports about a clinic by adding an indicator variable identifying clinicians as a fixed effect in models (8) and (9).

Results

Sites and Informants

The EQHIV study sites were representative of Title III clinics nationally (Landon et al. 2004). Differences between intervention and control sites in terms of location (rural/urban, regional), site type, and clinic status (general medicine versus specialized HIV practice) were statistically insignificant. Just over three-quarters of the informants were clinicians rather than directors or clinician–directors, 51 percent were male, and 71 percent were physicians. Clinicians and clinician–directors had a mean age of 42. The mean number of informants per clinic was 3.4 for items on the clinician survey only, and 3.6 for those on both the director and clinician surveys.

Reliability of Global Organizational Measures

Table 1 presents estimated reliabilities for 26 single-item global measures that ask informants to report organization-level features. The first column presents the intraclass correlation ρ_x, interpretable here as the reliability of a single informant report. The second column presents the multiple-informant reliability $\rho_{\bar{x}}$ evaluated at the mean number of informants per organization. The implied number of informants required to reach $\rho_{\bar{x}} = 0.70$ appears in column 3; columns 4–6 give the numbers of informants and clinics for each item, the correlation ratio η^2, and the F ratio from one-way ANOVA.

Most estimated one-informant reliabilities ρ_x are small; the median intraclass correlation is 0.18 for the 26 measures. An exception is the priority placed on research, with estimated reliability over 0.60. The remaining 25 estimates of ρ_x range between 0.04 (funding limitations as a barrier to improvement) and 0.36 (whether a computer is available for patient care).

Table 1: Reliability Measures for Single Items—Global Organizational Properties

Item	ρ_x	$\rho_{\bar{x}}$ *at* $n_j = N/J$	n_j^* *Needed for* $\rho_{\bar{x}} = 0.7$	*N, J*	η^2	*F Ratio (p-Value)*
Clinic priorities						
High-quality clinical care	0.105	0.298	19.8	325, 90	0.334	1.32 (.05)
Research to improve HIV care	0.604	0.846	1.5	324, 90	0.719	6.73 (<.001)
Helping patients and families access resources	0.133	0.357	15.2	325, 90	0.369	1.54 (.005)
Community outreach/ prevention	0.300	0.607	5.4	324, 90	0.490	2.53 (<.001)
Clinic barriers						
Limited staff	0.126	0.342	16.1	324, 90	0.372	1.56 (.004)
Limited funding	0.036	0.120	62.1	325, 90	0.301	1.14 (.221)
Limited expertise	0.172	0.427	11.2	323, 90	0.418	1.88 (<.001)
Limited travel resources	0.101	0.288	20.8	325, 90	0.359	1.48 (.011)
Limited pt visit time	0.281	0.586	6.0	325, 90	0.483	2.47 (<.001)
Clinical leadership and QI						
Clarity of vision	0.191	0.437	9.9	292, 89	0.433	1.76 (<.001)
Responsiveness to suggestions	0.247	0.519	7.1	293, 89	0.474	2.09 (<.001)
Ability to implement QI	0.208	0.462	8.9	292, 89	0.445	1.85 (<.001)
Supportiveness of collaborative*	0.180	0.410	10.6	159, 52	0.431	1.59 (.023)
HIV clinical staff						
Initiative	0.210	0.486	8.8	320, 90	0.434	1.98 (<.001)
Collaboration	0.311	0.599	5.2	295, 89	0.518	2.52 (<.001)
Education/training	0.204	0.474	9.1	318, 90	0.423	1.88 (<.001)
Receptiveness	0.248	0.522	7.1	295, 89	0.485	2.20 (<.001)
Clinic practices						
Decentralization	0.159	0.382	12.3	286, 88	0.416	1.62 (.003)
Specific quantifiable goals	0.196	0.486	9.6	324, 90	0.415	1.86 (<.001)
Routine progress measurement	0.060	0.167	36.5	276, 88	0.343	1.13 (.251)
Consultation of pts re QI	0.184	0.426	10.4	293, 89	0.424	1.71 (.011)
Link pts/families to resources	0.175	0.414	11.0	296, 89	0.440	1.85 (<.001)
Guidelines	0.110	0.291	18.8	295, 89	0.365	1.35 (.044)
Computer available for pt care	0.364	0.657	4.1	298, 89	0.552	2.93 (<.001)
QI experience						
Was there a recent QI initiative?	0.106	0.300	19.7	326, 90	0.368	1.54 (.005)
Was the initiative worthwhile?†	0.083	0.201	25.8	228, 82	0.379	1.10 (.306)

*Item was asked only at intervention clinics.

†Item was asked only when an initiative was reported.

The estimated reliabilities $\rho_{\bar{x}}$ for clinic means are higher than the intraclass correlations for individual items because averaging across multiple informants lowers error variance. Nonetheless, with the number of informants per organization in the EQHIV study (around 3.3 for most items after deletion of informants with missing values), only the organization-level mean for research emphasis has an estimated reliability greater than 0.70. Other estimates range from 0.12 (funding limitations) to 0.66 (computer availability). The median $\rho_{\bar{x}}$ in Table 1 is 0.43. Other clinic-level measures that approach 0.70 reliability include the priority assigned to outreach/prevention activities ($\rho_{\bar{x}} = 0.61$), limited visit time as a barrier to improvement (0.59), and collaboration among clinical staff (0.60).

Given informant variations and item-level measurement errors, a substantial number of informants would be required to obtain reliable measures of many global organizational features. Values of n_j^* range from 1.5 (research emphasis) to over 60 (limited funding), with a median of 10.5. While appreciably higher than the number of informants per organization in EQHIV, n_j^* for these single-item measures is usually lower than the number of informants per organization in other multiple-informant studies in health care settings. Both the Shortell et al. (1991) and Temkin-Greener et al. (2004) studies, for instance, had over 40 informants per organization.

All but three F ratios from ANOVAs for the global items are significant at the 0.05 level. Thus, finding significant organizational differences does not imply high reliability. Values of the correlation ratio η^2 range from 0.30 (limited funding) to 0.72 (research emphasis). Because of the relatively small number of informants per organization, values of η^2 are high by comparison with the intraclass correlations ρ_x.[4]

Reliability of Analytical Organizational Measures

Analytical organizational characteristics such as a clinic's specialty composition can be measured using means of individual characteristics reported by sampled respondents within an organization. For such measures, respondent-level variance reflects heterogeneity rather than disagreement. Such heterogeneity nonetheless reduces the reliability of an analytical measure.

Table 2 evaluates 30 one-item analytical measures. The estimated reliabilities vary widely, although F ratios indicate organizational differences on most measures ($p < .05$ for 24 of 30). No organizational commonalities are evident for some, including mean hours devoted to administrative work and mean frequency of discussing guidelines. Other clinic means are relatively

Table 2: Reliability Measures for Single Items—Analytical Organizational Properties

Item	ρ_x	$\rho_{\bar{x}}$ *at* $n_j = N/J$	n_j^* *Needed for* $\rho_{\bar{x}} = 0.7$	*N, J*	η^2	*F Ratio (p-Value)*
Knowledge and expertise						
Response to rising HIV viral load	0.063	0.184	34.7	299, 89	0.344	1.25 (.100)
Contraindication for AZT	0.151	0.375	13.1	299, 89	0.407	1.64 (.002)
When to add fourth drug to regimen	0.071	0.205	30.4	299, 89	0.360	1.34 (.046)
# determinations for baseline VL	0.079	0.224	27.1	299, 89	0.357	1.33 (.052)
Resistance, reverse transcriptase	0.169	0.406	11.5	299, 89	0.419	1.72 (<.001)
Resistance, protease inhibitors	0.120	0.314	17.1	299, 89	0.379	1.45 (.016)
Self-assessed HIV expertise	0.279	0.563	6.0	293, 88	0.498	2.34 (<.001)
Infectious disease certification*	0.557	0.763	1.9	205, 80	0.727	4.22 (<.001)
Time allocation						
Hrs/week on patient care	0.278	0.559	6.1	293, 88	0.491	2.23 (<.001)
. . . on administration	0	0	—	291, 89	0.253	0.78 (.908)
. . . on teaching/precepting	0.253	0.525	6.9	286, 88	0.472	2.04 (<.001)
. . . on research	0.291	0.574	5.7	292, 89	0.500	2.30 (<.001)
Behaviors with patients						
Frequency discuss guidelines	0	0	—	297, 89	0.286	0.61 (.606)
Give patients resource info	0.109	0.588	19.0	297, 89	0.383	1.47 (0.014)
Give patients written materials	0.115	0.302	18.0	297, 89	0.359	1.32 (0.054)
Educate family/friends of patients	0.179	0.423	10.7	298, 89	0.431	1.80 (<.001)
Patient load						
# outpatients seen per week	0.648	0.860	1.3	296, 89	0.743	6.81 (<.001)
% of patients seen with HIV	0.573	0.816	1.7	294, 89	0.695	5.32 (<.001)
# HIV patients in clinician panel	0.531	0.781	2.1	292, 89	0.662	4.30 (<.001)
Other clinic activities						
Participation in clinic decisions	0.077	0.214	28.0	292, 89	0.353	1.26 (.093)
Use computer for patient care†	0.461	0.706	2.7	210, 75	0.646	3.32 (<.001)
Use e-mail with patients	0.363	0.648	4.1	287, 89	0.561	2.87 (<.001)
% HIV patients in clinical trials	0.348	0.630	4.4	281, 88	0.557	2.79 (<.001)
Clinician practice satisfaction	0.154	0.383	12.6	296, 88	0.406	1.64 (.002)
On-site access to HIV expert	0.115	0.305	17.9	300, 89	0.408	1.65 (.002)
Sociodemographic composition						
Gender	0.044	0.141	50.9	322, 90	0.303	1.13 (.230)
White/nonwhite	0.320	0.603	5.0	287, 89	0.539	2.63 (<.001)
Age	0	0	—	298, 89	0.282	0.93 (.645)
Years since MD*	0.068	0.158	32.1	206, 80	0.453	1.32 (.081)
Physician/nonphysician	0.196	0.452	9.5	299, 89	0.434	1.83 (<.001)

*Asked only of physicians.

†Asked only when computer reported available in clinic.

reliable, however, even with the limited number of informants in this study. The proportion of physicians who are board certified in infectious diseases, for example, has an estimated organization-level reliability $\rho_{\bar{x}}$ of 0.76. Clinic means on measures of patient load—outpatients per week, percentage of outpatients with HIV, number of HIV patients in a clinician's panel—have estimated reliabilities of 0.86, 0.82, and 0.78, respectively. Across the 30 measures in Table 2, the median value of n_j^* needed to obtain clinic-level reliability of 0.70 is just over 12.

Multiple-Item Scales

Multiple-item scales can yield more reliable organizational measures than single items, as item-level errors tend to cancel out when items are combined. Table 3 assesses the reliability of organization-level scale means in the EQHIV study. Scales include measures of both global and analytical properties.

The first column of Table 3 presents ρ_r, i.e., (5) evaluated assuming one informant per organization. The second column presents estimated reliabilities $\rho_{\bar{r}}$ for scale means, i.e., (5) evaluated at the mean number of informants per organization in EQHIV. Column 3 gives the implied number of informants per organization needed to obtain a mean with 0.70 reliability, and column 4 gives the *p* level for testing the hypothesis of no organizational variance using a likelihood-ratio statistic.

To highlight differences between reliability assessments taking organizational and informant standpoints, column 5 presents an informant-level reliability measure

$$\rho_{r(i)} = \frac{\sigma_{\mathrm{o}}^2 + \sigma_{\mathrm{i}}^2}{\sigma_{\mathrm{o}}^2 + \sigma_{\mathrm{i}}^2 + \sigma_{\mathrm{e}}^2/K}. \tag{12}$$

Table 3: Reliability Measures for Multiple-Item Scales

Scale (# items)	ρ_r	$\rho_{\bar{r}}$ *at* $n_j = N/J$	n_j^* *Needed for* $\rho_{\bar{r}} = 0.7$	*p-Value, LLR Test of* $H_0 : \sigma_{\mathrm{o}}^2 = 0$	$\rho_{r(i)}$	*Cronbach's* α	*N, J*
Openness to QI (7)	0.365	0.674	4.05	<.001	0.803	0.792	328,91
HIV knowledge (6)	0.255	0.522	7.32	<.001	0.566	0.567	299, 89
Research emphasis (3)	0.693	0.891	1.02	<.001	0.693	0.697	330, 91
Autonomy (3)	0.271	0.570	6.30	<.001	0.566	0.572	326, 91
Patient help (3)	0.172	0.410	11.33	<.001	0.602	0.604	298, 89
Guidelines emphasis (2)	0.101	0.272	20.51	.035	0.530	0.513	297, 89
Barriers to QI (5)	0.115	0.323	17.59	.019	0.651	0.652	325, 90
Patient load (3)	0.221	0.486	8.31	<.001	0.502	0.498	298, 89

Measure (12) treats both organizational and informant variance as reliable; only item-level variation is regarded as erroneous. Values of $\rho_{r(i)}$ are comparable with those of Cronbach's α presented in column 6. Contrasting $\rho_{r(i)}$ and α with the organization-level reliabilities in columns 1 and 2 illustrates differences in scale reliability at organizational and informant levels.

Significant ($p<.05$) organization-level variance is present for all eight scales. For two, reliable organizational differences can be detected using the number of informants in EQHIV. The one-informant reliability ρ_r is almost 0.70 for the research emphasis scale, and clinic means on this scale have a reliability of nearly 0.90 at 3.6 informants per organization. Likewise, the multiple-informant reliability of the seven-item scale measuring openness to quality improvement is 0.67. Other scales perform less well. The results imply that reliable organizational measures could be obtained with fewer than 10 informants for most scales.

Even though these scales are relatively short, their estimated within-informant reliabilities often approach or exceed 0.7. Estimates of $\rho_{r(i)}$ and α range from 0.50 (patient load scale) to 0.80 (openness to QI scale). A scale can be reliable at the informant level and yet be a weak organization-level measure if informant-level variance σ_{i}^2 is large. For example, informants answer the five items on barriers to improvement consistently ($\rho_{r(i)} = \alpha = 0.65$), but appreciable informant differences produce ρ_r of only 0.12, and an estimated reliability for organization means (at 3.6 informants) of 0.32. Informant variations are much smaller for openness and research emphasis, so their internal consistency and organizational reliability are both high.

Multiple-item scales clearly can improve organizational measurement, but informant differences limit the improvements possible through adding scale items. Assuming one informant and arbitrarily many items, organizational reliability in (5) cannot exceed $\sigma_{\mathrm{o}}^2/(\sigma_{\mathrm{o}}^2 + \sigma_{\mathrm{i}}^2)$. For the EQHIV data, this upper bound on the organizational reliability of a scale ranges from 0.18 for the barriers to improvement scale, where the informant-level variance is over four times the organizational variance, to 1.0 for research emphasis, which had estimated informant variance of 0. Further improvements in reliability would require multiple informants.

Clinician/Director Differences

We compared the responses of clinicians with those of directors (including clinician-directors) on all items and scales in Tables 1–3. Differences significant at or below the 0.10 level are displayed in Table 4. The first column gives

Table 4: Clinician/Director Differences on Items and Scales

	Clinician/Director Difference	*Clinician/Director Difference (SD Units)*	*p-Level*
Global organizational items (Table 1)			
Priority: high-quality clinical care	– 0.140	– 0.259	.034
Priority: research	0.272	0.213	.01
HIV clinical staff: education and training	0.306	0.336	.004
Clinic practice: decision decentralization	– 0.187	– 0.262	.089
Was there a recent QI initiative?	– 0.105	– 0.238	.054
Individual characteristic items (Table 2)			
Knowledge: contraindication for AZT	– 0.175	– 0.419	.005
Knowledge: when to add fourth drug to regimen	– 0.108	– 0.346	.024
Knowledge: # determinations for baseline viral load	– 0.149	– 0.380	.013
Self-assessed HIV expertise	– 0.231	0.504	<.001
Participation in clinic decisions	– 0.742	– 0.792	<.001
Frequency discuss guidelines with patients	0.228	0.408	.008
# outpatients seen per week	– 9.75	– 0.230	.026
% of patients seen with HIV	– 9.05	– 0.240	.033
# HIV patients in panel	– 77.67	– 0.444	<.001
On-site access to HIV expert	– 0.243	– 0.737	<.001
Years since MD	– 2.55	– 0.320	.054
Physician	– 0.305	– 0.667	<.001
Gender (female)	0.197	0.394	.002
Scale scores (Table 3)			
HIV knowledge	– 0.105	– 0.263	.001
Autonomy	– 0.250	– 0.266	.004
Guidelines emphasis	0.203	0.346	.005

the clinician/director difference using the units of measure in the EQHIV surveys; the second column uses standard deviation units.

Directors and clinicians assessed a few organizational characteristics differently. Significant differences, ranging between a quarter and a third of a standard deviation, were found for five of the 26 global organizational indicators from Table 1. Clinicians characterized their clinics as placing a lower priority on clinical care and a higher priority on research than did directors. Clinicians rated the education of HIV clinical staff somewhat higher than directors did, reported less decentralization, and were less likely to report a recent QI initiative.

Clinicians and directors differed on three of eight scales. Clinicians reported more emphasis on guidelines, somewhat less autonomy, and scored lower than clinician-directors on the HIV knowledge scale. There were several clinician–director differences on the individual characteristics from Table 2, most of which reflect factual rather than perceptual differences.

Discussion

This study found that survey measures of organizational properties for Title III HIV clinics had low to modest reliability. Reports reflect common organizational phenomena, but vary substantially among informants within organizations. This can reflect perceptual differences, different interpretations of questions, and other measurement errors. Multiple-item scales can improve organizational measures, but scale scores also vary substantially within organizations. Our analyses suggest that obtaining reliable organizational measurements usually requires aggregation of reports across multiple informants.

The relatively low reliabilities for organizational means reflect a limited number of informants per organization, rather than especially low informant-level agreement. Informants can be familiar with the full organization in the relatively small EQHIV sites. One would expect lower concordance in studies of larger health care organizations such as hospitals.

The EQHIV intraclass correlations are high relative to those we calculated from other multiple-informant health care organization studies. Approximate intraclass correlations for constructs in a study of PACE teams (Temkin-Greener et al. 2004) range between 0.06 (conflict management) and 0.07 (perceived team effectiveness).[5] Teams there were assessed, on average, by over 40 informants, so organization-level means have relatively high reliability; we calculate a range from 0.72 (conflict management) to 0.76 (effectiveness).

Reducing the informant and error components of variance in (4) can increase measure reliability. Pretesting, clarifications in item wording, and specific probes (Casalino et al. 2003) can reduce item-level error. Ensuring that the object of measurement (e.g., a clinic rather than a floor or team) is salient to informants also can reduce informant variations. Adding both scale items and informants can improve reliability. Additional items raise reliability by reducing item/error variance, while additional informants lower both informant and item/error variance. Improvements in reliability from adding informants are potentially greater than those from adding items. Recruiting new informants is, however, more expensive than lengthening a scale.

Directors occasionally gave more optimistic assessments than did clinicians. Such differences occurred only slightly more often than expected by chance, though, and were relatively small. Other informant differences also may influence assessments, however. Temkin-Greener et al. (2004) found that professionals assessed teams more positively than did paraprofessionals.

Our reliability estimates reflect variation in phenomena within the EQHIV study population as well as agreement among informants. If true variation is limited, a measure will have low reliability if there is even modest informant disagreement. Agreement coefficients (James, Demaree and Wolf 1984; LeBreton, James, and Lindell 2005) assess agreement per se by comparing observed disagreement with a conceivable level calculated using a null (e.g., uniform) distribution, rather than, with observed variation within a study population. As variation in several EQHIV measures is highly restricted, agreement coefficients are much higher than reliabilities for these. For example, the priority assigned to high-quality HIV care is high and varies little across organizations; the mean priority on a 1–5 scale is 4.75, with a standard deviation of 0.54. The pooled agreement coefficient r^*_{WGp} (LeBreton, James, and Lindell 2005) is 0.87 for high-quality care, while the single-informant reliability in Table 1 is only 0.11. Leadership responsiveness is another example of high agreement but low reliability ($r^*_{WGp} = 0.72, \rho_x = 0.25$). These comparisons suggest that our measures might be more reliable if assessed using more heterogeneous organizations. While agreement coefficients are generally higher than the corresponding reliabilities, agreement levels are low for many EQHIV measures; examples include limited staff as a barrier to improvement ($r^*_{WGp} = 0.38$), decentralization ($r^*_{WGp} = 0.36$) and presence of a recent QI initiative ($r^*_{WGp} = 0.31$).

Another limitation of this study is that its findings for Title III clinics may not generalize to other health care organizations. As well, the clinician survey included many indicators prone to subjective interpretation. It is likely that informant reliability is higher for objective features such as the size of the medical staff or total clinic caseload. EQHIV assembled such information in a single-informant site survey, so we were unable to assess the reliability of such data.

Multiple-informant organizational measures are usually constructed by taking a mean across several reports. Informant variation reduces the reliability of such measures, but it also can be of substantive interest. Temkin-Greener et al. (2004), for example, use an ethnic diversity index to predict team performance. Our study did not attempt to assess the reliability of measures of organizational diversity or variation.

Surprisingly few studies of clinic or hospital characteristics report the organization-level reliability of their measures. Many that do rely on statistics such as the *F*-statistic or the correlation ratio do not adequately describe unit level reliability. Some studies report the informant-level internal consistency of scales, but a scale can be internally consistent within informants yet be unreliable as an organizational-level measure. This study found substantial item and respondent variability in clinic assessments, and modest or low clinic-level reliability for many measures. We suggest that studies of organizational characteristics should report the organizational-level reliability of the measures used, if possible.

ACKNOWLEDGMENT

Supported by a grant (R-01 HS10227) from the Agency for Healthcare Research and Quality (ARHQ). Carol Cosenza, MSW, and Patricia Gallagher, PhD, of the Center for Survey Research assisted with instrument development and survey administration. We also thank colleagues at the Health Resources and Services Administration and at the Institute for Healthcare Improvement who participated in and facilitated the EQHIV study, and two anonymous reviewers for helpful comments on a previous version of the this article.

NOTES

1. Informants who responded to both the director and clinician questionnaires answered the 15 overlapping items twice. Paired *t*-tests detected significant differences between the "director" and "clinician" responses on two items: informants gave significantly higher assessments of the priority placed on community outreach activities ($p = .044$) and the barriers to improvement posed by limited funding ($p = .033$) when responding as directors rather than clinicians. We used the "director" responses of these informants on the 15 overlapping items.
2. Most indicators in the EQHIV surveys are ordered and dichotomous measures. We follow typical practice by assigning equally spaced scores to these and treating them as quantitative variables. We reached similar conclusions about reliability using logit and ordinal logit models that treat the indicators as discrete variables (Snijders and Bosker 1999).
3. It is possible for n_j^* to exceed the number of eligible respondents in some organizations, since n_j^* rises with both error and informant variance. Large values of n_j^* reflect low reliability.
4. The expected value of the between-group sum of squares in ANOVA (the numerator of η^2) depends on both the within-group variance and the between-group variance (Searle, Casella, and McCulloch 1992), so η^2 is positive even with no

between-group variance. When $\sigma_o^2 = 0$, η^2 is $(J-1)/(N-1)$; this ratio is substantial, 0.274, for illustrative values of $J=91$ and $N=330$ from EQHIV.

5. Our calculations assume that the number of informants is the same in all organizations. F statistics then imply intraclass correlations $\rho_x = (F-1)/(F-1+N/J)$, and organization-level reliabilities $\rho_{\bar{x}} = (F-1)/F$. If the number of informants differs across organizations, reliabilities are higher than calculated, but only slightly so unless the variation in informants is very large.

REFERENCES

Aiken, L. H., and P. A. Patrician. 2000. "Measuring Organizational Traits of Hospitals: The Revised Nursing Work Index." *Nursing Research* 49 (3): 146–53.

Aiken, L. H., and D. M. Sloane. 1997. "Effects of Specialization and Client Differentiation on the Status of Nurses: The Case of AIDS." *Journal of Health and Social Behavior* 38 (3): 203–22.

Bohrnstedt, G. 1983. "Measurement." In *Handbook of Survey Research*, edited by P. H. Rossi, J. D. Wright, and A. B. Anderson. New York: Academic Press.

Casalino, L., R. R. Gillies, S. M. Shortell, J. A. Schmittdiel, T. Bodenheimer, J. C. Robinson, T. Rundall, N. Oswald, H. Schauffler, and M. C. Wang. 2003. "External Incentives, Information Technology, and Organized Processes to Improve Health Care Quality for Patients with Chronic Diseases." *Journal of the American Medical Association* 289 (4): 434–41.

Flood, A. B. 1994. "The Impact of Organizational and Managerial Factors on the Quality of Care in Health Care Organizations." *Medical Care Review* 51 (4): 381–428.

Huber, G. P., and D. J. Power. 1985. "Retrospective Reports of Strategic-Level Managers: Guidelines for Increasing Their Accuracy." *Strategic Management Journal* 6 (2): 171–80.

James, L. R., R. G. Demaree, and G. Wolf. 1984. "Estimating Within-Group Interrater Reliability with and without Response Bias." *Journal of Applied Psychology* 69 (1): 86–98.

Kralewski, J. E., E. C. Rich, R. Feldman, B. E. Dowd, T. Bernhardt, C. Johnson, and W. Gold. 2000. "The Effects of Medical Group Practice and Physician Payment Methods on Costs of Care." *Health Services Research* 35 (3): 591–613.

Landon, B. E., I. B. Wilson, and P. D. Cleary. 1998. "A Conceptual Model of the Effects of Health Care Organizations on the Quality of Medical Care." *Journal of the American Medical Association* 279 (17): 1377–82.

Landon, B. E., I. B. Wilson, K. McInnes, M. B. Landrum, L. Hirschhorn, P. V. Marsden, D. Gustafson, and P. D. Cleary. 2004. "Effects of a Quality Improvement Collaborative on the Outcome of Care of Patients with HIV Infection: The EQHIV Study." *Annals of Internal Medicine* 140 (11): 887–96.

Lazarsfeld, P. F., and H. Menzel. 1980. "On the Relation between Individual and Collective Properties." In *A Sociological Reader on Complex Organizations*, 3rd Edition, edited by A. Etzioni and E. W. Lehman. New York: Holt, Rinehart and Winston.

LeBreton, J. M., L. R. James, and M. K. Lindell. 2005. "Recent Issues Regarding r_{WG}, r^{*}_{WG}, $r_{WG(J)}$, and $r^{*}_{WG(J)}$." *Organizational Research Methods* 8 (1): 128–38.

Meterko, M., D. C. Mohr, and G. J. Young. 2004. "Teamwork Culture and Patient Satisfaction in Hospitals." *Medical Care* 42 (5): 492–8.

Nunnally, J. C. 1978. *Psychometric Theory*, 2nd Edition. New York: McGraw-Hill.

O'Brien, R. M. 1990. "Estimating the Reliability of Aggregate-Level Variables Based on Individual-Level Characteristics." *Sociological Methods and Research* 18 (4): 473–504.

Rabe-Hesketh, S., A. Pickles, and A. Skrondal. 2001. *GLLAMM Manual.* Technical Report 2001/01, Department of Biostatistics and Computing, Institute of Psychiatry, King's College, University of London. Downloadable from http://www.gllamm.org/.

Scheffé, H. 1959. *The Analysis of Variance.* New York: Wiley.

Searle, S. R., G. Casella, and C. E. McCulloch. 1992. *Variance Components.* New York: Wiley.

Seidler, J. 1974. "On Using Informants: A Technique for Collecting Quantitative Data and Controlling Measurement Error in Organization Analysis." *American Sociological Review* 39 (6): 816–31.

Shavelson, R. J., and N. Webb. 1991. *Generalizability Theory: A Primer.* Newbury Park, CA: Sage.

Shortell, S. M., J. L. O'Brien, J. M. Carman, R. W. Foster, E. F. X. Hughes, H. Boerstler, and E. J. O'Connor. 1995. "Assessing the Impact of Continuous Quality Improvement/Total Quality Management: Concept versus Implementation." *Health Services Research* 30 (2): 377–401.

Shortell, S. M., D. M. Rousseau, R. R. Gillies, K. J. Devers, and T. L. Simons. 1991. "Organizational Assessment in Intensive Care Units (ICUs): Construct Development, Reliability, and Validity of the ICU Nurse–Physician Questionnaire." *Medical Care* 29 (8): 709–26.

Snijders, T. A. B., and R. Bosker. 1999. *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling.* London: Sage.

StataCorp. 2003. *Stata Statistical Software: Release 8.0.* College Station, TX: Stata Corporation.

Temkin-Greener, H., D. Gross, S. J. Kunitz, and D. Mukamel. 2004. "Measuring Interdisciplinary Team Performance in a Long-Term Care Setting." *Medical Care* 42 (5): 472–81.