Psychometric Properties of Scores from the Web-Based LibQUAL+ Study of Perceptions of Library Service Quality

COLLEEN COOK AND BRUCE THOMPSON

ABSTRACT

BASED ON DATA PROVIDED BY 4,407 PARTICIPANTS, the present study investigated the psychometric integrity of scores on thirty-four items of the LibQUAL+ evaluation of perceived library quality. The study investigated LibQUAL+ score structure, score reliability, score correlation and concurrent validity coefficients, scale means, and scale standardized norms. If both generic and specialized norms were eventually developed for a large sample of users at ARL institutions, LibQUAL+ norms could then facilitate the ultimate application of LibQUAL+—i.e., identifying areas of potential improvement at a given library, and identifying similar libraries with more favorable profiles whose behavior might then be modeled in pursuit of providing better service to library users.

INTRODUCTION

When most of us visit a surgeon prior to an operation, we probably are concerned about our physician's collection of surgical instruments, diplomas, and reference reprints on surgical procedures. But we probably are concerned about other things in addition to the physician's collections. We care at least as much that our surgeon is focused on our needs, empathic regarding our interests, and dedicated to providing quality service on a consistent basis.

Although users of research libraries may not have life-threatening interests at stake, many library users do feel that service quality is vital to

Colleen Cook, Texas A&M University, College Station, TX 77843-4225
Bruce Thompson, Texas A&M University, Department of Educational Psychology, College Station, TX 77843-4225
LIBRARY TRENDS, Vol. 49, No. 4, Spring 2001, pp. 585-604

their abilities to obtain academic degrees or external funding and is critical in creating and disseminating knowledge. Libraries ignore user perceptions of library service quality at their peril. In the modern research library, the singular use of resource-based metrics as the only index of library quality can no longer be regarded as reasonable.

So it is not surprising that libraries confront "pressure . . . to assess the degree to which their services demonstrate criteria of 'quality.' . . . The emphasis on these measures and services provided to library clientele requires librarians . . . not to equate 'quality' merely with collection size" (Hernon & McClure, 1990, p. 155). As Nitecki (1996b) noted: "A measure of library quality based solely on collections has become obsolete" (p. 181). As a matter of fact: "In recent years, LIS [Library and Information Science] researchers have drawn on marketing and other literatures to focus attention on *expectations* and an alternative view of *quality*, one representing the user's or customer's perspective on the services used" (Nitecki & Hernon, 2000, p. 259).

These dynamics led the Association of Research Libraries (ARL) to institute its "New Measures" initiatives. One of the "New Measures" initiatives is the LibQUAL+ study being conducted by ARL and the Texas A&M University Libraries (Cook & Heath, 2000a; Cook, Heath, & Thompson, 2000a). Continuing phases of the LibQUAL+ study are being supported in part by the Fund for the Improvement of Post-Secondary Education (FIPSE).

Briefly, the first iteration of the LibQUAL+ protocol was developed in Spring 2000. The initial phase of the study involved participation with Texas A&M University and twelve additional institutions:

—University of Arizona
—University of California, Santa Barbara
—University of Connecticut
—University of Houston
—University of Kansas
—Michigan State University
—University of Minnesota
—University of Pennsylvania
—University of Pittsburgh
—Virginia Tech
—University of Washington
—York University

In its first phase, the protocol built on the use of the twenty-two items in the well-established SERVQUAL instrument (Parasuraman, Zeithaml, & Berry, 1985, 1994).

The SERVQUAL protocol ostensibly measures perceptions of service *tangibles, reliability, responsiveness, assurance,* and *empathy* (Parasuraman, Berry,

& Zeithaml, 1991). Within this model, "only customers judge quality; all other judgments are essentially irrelevant" (Zeithaml, Parasuraman, & Berry, 1990, p. 16).

However, the twenty-two items of SERVQUAL have not yielded the expected five-factor structure when the instrument has been used within the library setting (Cook & Thompson, 2000, in press; Niteki, 1996a). Furthermore, it is critical to ground any evaluation of library service quality within the perceptual schemata evoked by users in their thinking about libraries. Thus, one of the initial steps in the LibQUAL+ inquiry involved conducting in-depth interviews with users at several of the institutions in our study.

The findings of this qualitative work have been described elsewhere (Cook & Heath, 2000b) and resulted in our adding nineteen items to the LibQUAL+ measure used in Spring 2000. The LibQUAL+ items will continue to evolve as the project moves forward. Revisions will continue to be informed by qualitative work plus quantitative analyses such as those reported here.

In short, LibQUAL+ is (1) not SERVQUAL, and (2) not (at least yet), a fixed core of unchanging items. LibQUAL+ is instead grounded in the epistemological view that, in the behavioral sciences, dynamic "theory building and construct measurement are joint bootstrap operations" (Hendrick & Hendrick, 1986, p. 393). The results described here apply to LibQUAL+ in its current form, but the reader is cautioned that this tool will continue to evolve as we collect new iterations of data from an increasing number of users and an even broader array of libraries.

The present inquiry was conducted to address five questions:

1. Can a meaningful and replicable structure underlying user perceptions of library services be identified?
2. Can psychometrically stable scores on LibQUAL+ dimensions be generated?
3. Are scores on different LibQUAL+ dimensions of user perceptions correlated with each other and user overall ratings of library service quality?
4. Do comparisons of LibQUAL+ subscale and total scores across user types suggest that LibQUAL+ scores are psychometrically valid?
5. Can standardized norms potentially be developed to assist librarians in understanding user perceptions of library service quality and targeting areas of needed or desired improvement?

METHOD

Participants

Under the guidance of a lead library contact at the twelve institutions, random samples of 600 faculty, 600 graduate students, and 900

undergraduate students were randomly selected at each institution. However, some institutions elected to oversample some respondent groups. Undergraduate students were uniformly oversampled because it was anticipated that their response rates would be disproportionately lower.

For the analyses reported here, the 4,407 participants were divided into two subsamples (n_1 = 420; n_2 = 3,987) based on LibQUAL+ administration format. Descriptions of the samples are available elsewhere (Cook, Heath, & Thompson, 2000b; Cook, Heath, Thompson, & Thompson, in press–a; Cook, Heath, Thompson, & Thompson, in press–b; Thompson, Cook, & Heath, in press).

Procedure

Each randomly selected participant received an e-mail from the library administration at the home campus. This message requested participant assistance in improving library service quality by responding to a brief survey. The participants were informed that the survey was being administered on the Web. The invitation to participate included a hot hyperlink to the Web survey URL. However, participants were also told that they could access the Web site by typing the URL address into the destination box on the Web browser of their preference.

The URL initially sent the participants to the servers at ARL, which then connected the users to servers housing the survey at Texas A&M University. The first page of the survey included a colorized logo furnished by each of the participating universities. Thus, the survey appearance was somewhat individualized for each school.

Prior to responding to the forty-one LibQUAL+ items and some additional items, users were first asked to provide general demographic information. This was done to allow subsequent descriptions of the samples, and a direct explicit comparison of respondents with the institutional profiles of each campus. Cook, Heath, and Thompson (2000) provide a thorough meta-analysis of reasonable response rate expectations and influences in Web-based surveys. In the current political season, when national surveys of 600 voters are (reasonably) generalized to 150 million Americans, it is intriguing that some continue to focus more on sample size than on sample representativeness.

However, as Thompson (2000) emphasized, the representativeness of the respondents is what counts in research. Response rate counts only to the extent that it may (or may not) bear upon sample representativeness. As Krosnick (1999) emphasized in his recent survey of the paper-and-pencil response-rate literature: "But it is not necessarily true that representativeness increases monotonically with increasing response rate .... [R]ecent research has shown that surveys with very low response rates can be more accurate than surveys with much higher response rates" (p. 540).

As regards the present participants, Thompson (2000) reported the evidence regarding sample representativeness.

Instrumentation

For each of the forty-one LibQUAL+ items, users were asked to rate their minimum expectations, perceptions, and desires regarding library quality. There were two formats for responding, each associated with one of the two subsamples.

Cook, Heath, Thompson, and Thompson (in press–b) provide more information, including pictures of selected Web pages regarding both response formats. Arnau, Thompson, and Cook (2001) present taxonometric analyses suggesting that user perceptions of library quality are continuously scaled.

The subsample of 420 respondents, drawn from York University and Texas A&M University, answered the survey using graphical sliders. For each item, these portray a continuum, and the respondent clicks and drags the slider along the continuum to a given point to communicate ratings. This may have the advantage of providing more precise ratings data.

The subsample of 3,987 respondents provided their ratings data using a "radio button" (hereafter "nonslider") response format. In this response format, for each item on each rating (i.e., minimum, perceived, and desired), participants were presented nine equally spaced small circles, and they clicked on the appropriate circle for a given response to darken it and thus communicate their ratings. This Web response format is analogous to the use of a nine-point Likert scale. On the average, the participants using the nonslider response format took 71.2 seconds less to complete the survey (M_{SLIDERS} = 12.5 minutes [SD = 5.0]; $M_{\text{NONSLIDERS}}$ = 11.3 minutes [SD = 5.5]).

Results

Dimensions of Perception

The first analysis investigated the dimensions underlying users' perceptions of library service quality. This analysis invoked separate principal components analyses of the two subsamples (Hetzel, 1996). The analyses summarized here followed the guidelines presented by Thompson and Daniel (1996).

Based on reliability item analysis and factor analyses for both prior related data sets (Cook & Thompson, 2000) and the present data (Cook, Heath, & Thompson, 2000b), a subset of thirty-four of the original forty-one LibQUAL+ items was retained for further analyses. Retention of a smaller subset of items allows for addition of new items in the next phase of LibQUAL+ item evolution while still maximizing score psychometric integrity.

Gorsuch (1983) has noted that: "A prime use of factor analysis has been in the development of both the theoretical constructs for an area and the operational representatives for the theoretical constructs" (p. 350). In short, "factor analysis is intimately involved with questions of validity. . . . Factor analysis is at the heart of the measurement of psychological constructs" (Nunnally, 1978, pp. 112-13).

The KMO sampling adequacy coefficients for the two analyses were .95 for the slider subsample and .97 for the nonslider data. These values strongly suggest the ample adequacies of the sample sizes for both analyses.

Both the eigenvalue-greater-than-one rule (Λ_5 = .98 and Λ_5 = .94, respectively) and "scree" plots suggested that four factors should be extracted. Of course, as LibQUAL+ evolves with the addition and deletion of items, in an ongoing renewal process informed by both qualitative work and empirical analysis, the structure measured by the protocol may change as well. The pattern/structure coefficients rotated to the varimax criterion in both analyses are presented in Table 1.

Score Reliability

An important element of evaluating score integrity involves the evaluation of score reliability. Coefficient alpha (α) can be computed for this purpose (Reinhardt, 1996). Some researchers deem coefficients of .7 or higher acceptable (Nunnally, 1978, p. 245), though higher values are desired, particularly as scores are applied in making higher stakes judgments (Pedhazur & Schmelkin, 1991).

Item analyses can be conducted as part of such inquiries (Thompson & Levitov, 1985). First, items are expected to "discriminate" between higher and lower scorers on a scale. To evaluate this item behavior, item scores (e.g., here "1" to "9" for the nonslider data) are correlated with scale scores, and reasonably large positive values are desired. However, these "discrimination" or "item-total correlation" coefficients would be inflated if scores on a given item were correlated with scores on a scale to which the given item scores also made a contribution.

For this reason, "corrected" discrimination coefficients are computed by correlating item scores with scores on a given scale computed without using the given item. For example, in the present study, the corrected discrimination coefficient for item 28, a Reliability scale item, was computed by correlating nonslider item 28 scores (ranging from "1" to "9") with scale scores computed using the remaining six of the seven items constituting this scale (ranging from 6 x 1 = "6" to 6 x 9 = "54").

Second, it is important that "if item deleted" statistics can be computed for each item. Good items hurt score reliability the most when they are not included. For example, for the nonslider data, the LibQUAL+ Reliability scale score alpha was .863, but if item #5 was omitted, it became .829. This suggests that item 5 was a very good item for the Reliability

Table 1. Varimax-Rotated Pattern/Structure Coefficients for Slider (n=420) and Nonslider (n=3987) Data.

	Slider Factors (n=420)				Nonslider Factors (n=3987)			
Item/Content Stem	I	II	III	IV	I	II	III	IV
19 Willingness to help users	.82	.10	.24	.18	.82	.11	.15	.25
18 Readiness to respond to user	.76	.19	.27	.16	.79	.11	.18	.28
24 Deal with users in caring fashion	.71	.19	.23	.23	.78	.24	.20	.14
20 Employees have knowledge	.73	.10	.26	.20	.76	.16	.22	.22
34 Employees who are courteous	.77	.18	.14	.18	.76	.19	.19	.14
9 Employees instill confidence	.62	.14	.41	.11	.71	.18	.17	.25
11 Employees understand needs	.60	.14	.54	.20	.70	.20	.24	.31
41 Giving users individual attention	.62	.28	.17	.24	.66	.22	.34	.13
15 Instruction in use, when needed	.49	.22	.34	.18	.61	.18	.20	.28
28 Performing services right	.61	.17	.24	.33	.58	.19	.32	.36
13 Users' best interests at heart	.46	.20	.58	.24	.58	.28	.24	.38
38 Employees have neat appearance	.47	.32	.04	.19	.48	.30	.28	.02
16 Maintain error free records	.45	.12	.29	.38	.40	.18	.30	.36
39 A meditative place	.13	.82	.01	.14	.16	.82	.18	.02
30 A haven for quiet and solitude	.17	.84	.04	.09	.16	.82	.16	.08
40 Space that facilitates quiet	.20	.82	.05	.14	.18	.80	.20	.05
12 A contemplative environment	.19	.81	.22	.10	.20	.79	.15	.19
4 A place for reflection	.10	.68	.34	.03	.12	.71	.08	.30
14 Comfortable and inviting location	.24	.72	.24	.15	.25	.69	.17	.24
29 Space group/individual study	.15	.64	.11	.33	.15	.66	.28	.13
22 Center intellectual interaction	.09	.72	.12	.22	.19	.63	.31	.05
21 A secure and safe place	.41	.45	.02	.12	.35	.36	.14	.26

(continued on page 592)

Table 1. (continued from page 591).

Item/Content Stem	Slider Factors (n=420) I	II	III	IV	Nonslider Factors (n=3987) I	II	III	IV
37 Complete runs of journal titles	.20	.21	.16	.71	.18	.21	.75	.11
27 Comprehensive print collection	.25	.15	.41	.54	.20	.26	.69	.17
36 Interdisciplinary needs addressed	.28	.16	.09	.68	.26	.21	.64	.15
10 Resources added to collection	.17	.10	.35	.50	.24	.13	.54	.31
25 Fulltext delivered electronically	.11	.32	.08	.50	.24	.23	.49	.12
35 Modern equipment	.40	.26	.11	.44	.28	.29	.48	.23
32 Library materials in the stacks	.32	.10	.33	.43	.28	.22	.47	.25
2 Providing services as promised	.36	.12	.66	.25	.35	.13	.21	.72
5 Service at promised time	.27	.04	.64	.15	.34	.19	.19	.72
3 Keep users informed	.19	.18	.67	.23	.32	.14	.13	.65
1 Convenient access collections	.17	.20	.66	.12	.22	.20	.32	.60
17 Timely document delivery	.35	.07	.38	.50	.31	.10	.38	.43

Note. Pattern/structure coefficients greater than .4 are underlined.

scale, because not using this item hurts the score integrity on this scale. The results of these various analyses are presented in Table 2.

Scale Relationships

Table 3 presents product-moment correlations of scores on the scales with each other and with total scores computed with all thirty-four LibQUAL+ items. Also presented in the table are correlations of subscale and total LibQUAL+ scores with scores on participants' rating of overall library quality.

This latter perception was collected at the end of the survey as a separate item. The correlations of LibQUAL+ scores with these global quality ratings are essentially concurrent validity coefficients.

Mean LibQUAL+ Differences

Also of interest were comparisons of LibQUAL+ means. These comparisons were made across both (a) LibQUAL+ scales, and (b) various demographic variables.

Comparisons Across Scales. The LibQUAL+ scales involve different numbers of items. To allow direct comparisons of scale means, for the purposes of

Table 2. Reliability Item Analysis Statistics for Slider (n=420) and Nonslider (n=3987) Data.

Scale/ Item	Slider Data "Corrected" Item-Total Correlation	Slider Data α if Item Deleted	Nonslider Data "Corrected" Item-Total Correlation	Nonslider Data α if Item Deleted	"Corrected" Total Scale Discrimination Slider	"Corrected" Total Scale Discrimination Non-Slider
Service						
11	.78	.918	.79	.932	.72	.73
13	.69	.922	.72	.935	.70	.73
24	.75	.919	.81	.931	.69	.73
20	.74	.920	.79	.932	.65	.71
18	.80	.917	.80	.932	.71	.70
19	.80	.917	.82	.931	.69	.70
41	.68	.922	.72	.935	.66	.69
34	.74	.919	.76	.933	.66	.68
9	.72	.920	.75	.934	.64	.67
15	.61	.925	.67	.937	.60	.64
38	.48	.931	.53	.942	.52	.56
Scale α		.928		.940		
Library as Place						
14	.75	.907	.72	.902	.67	.66
12	.82	.903	.79	.897	.66	.65
40	.80	.904	.77	.899	.61	.61
30	.79	.904	.78	.898	.57	.60
29	.66	.914	.67	.906	.59	.59
39	.76	.907	.77	.899	.55	.59
22	.70	.911	.65	.907	.56	.58
4	.65	.914	.68	.905	.55	.57
21	.48	.923	.45	.918	.51	.54
Scale α		.919		.913		
Access to Collections						
27	.59	.759	.64	.790	.60	.60
35	.52	.774	.57	.803	.58	.60
36	.60	.762	.60	.800	.55	.58
32	.50	.777	.52	.810	.54	.57
37	.60	.757	.64	.791	.56	.56
10	.50	.777	.56	.804	.49	.55
25	.43	.793	.49	.818	.45	.50
Scale α		.797		.826		
Reliability						
28	.59	.820	.67	.840	.66	.71
5	.58	.822	.74	.829	.49	.65
2	.72	.800	.73	.830	.64	.63
16	.55	.824	.57	.852	.58	.59
1	.54	.828	.58	.851	.52	.59
3	.62	.814	.60	.850	.57	.56
17	.60	.817	.56	.853	.59	.55
Scale α		.840		.863		
Total α					.952	.958

Note. Subscale and total score alpha coefficients are underlined. Total score results are computed as regards a single score produced using all 34 items.

Table 3. Product-moment Correlation Coefficients for Nonslider Data.

		LibQUAL+ Subscale				
LibQUAL+ Scale	Overall Rating	Service	Library as Place	Access to Collections	Reliability	LibQUAL+ TOTAL
Service	.678 (3,769)	1.000 (3,987)	.567 (3,987)	.686 (3,987)	.773 (3,987)	.896 (3,987)
Library as Place	.532 (3,769)		1.000 (3,987)	.625 (3,987)	.546 (3,987)	.820 (3,987)
Access to Collections	.675 (3,769)			1.000 (3,987)	.704 (3,987)	.851 (3,987)
Reliability	.659 (3,769)				1.000 (3,987)	.858 (3,987)
TOTAL	.733 (3,769)					1.000 (3,987)

Note. Sample sizes are reported in parentheses. All correlation coefficients are statistically significant at $\alpha = .001$.

these comparisons, subscale scores were divided by the number of scale items (e.g., 7 for the Reliability subscale) so that all means would fall within the same "1" (low) to "9" (high) score interval.

Figure 1 presents box-and-whisker plots for the four LibQUAL+ subscales for the 3,987 nonslider participants. Box-and-whisker plots present the score median as a bolder horizontal line within a box. The upper boundary of the box represents the third quartile (i.e., 75th percentile) while the lower boundary of the box represents the first quartile (i.e., 25 percentile). The location of the "whiskers" indicates the extreme score boundaries.

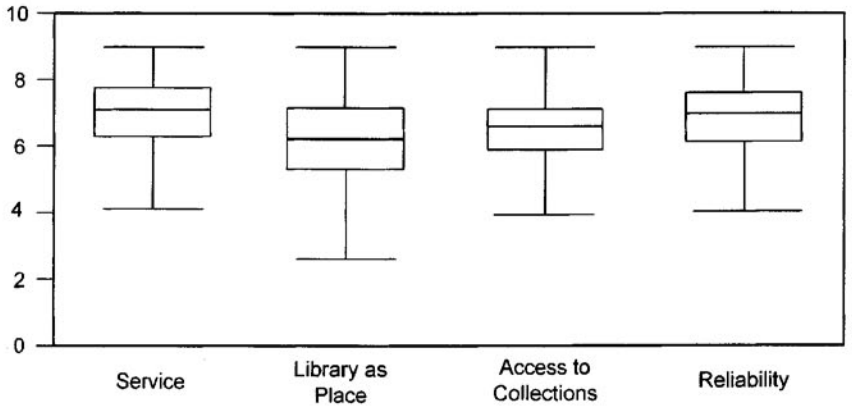

Figure 1. Box-and-Whisker Plots for LibQUAL+ Subscales Each Scaled "1" to "9."

Comparisons Across Demographic Variables. To facilitate comparisons across demographic variables, the LibQUAL+ scores were converted to so-called *T*-scores (i.e., scores with means of 50 and standard deviations of 10). Some of these comparisons were expected to be trivial. For example, there seems to be no theoretical basis on which to expect female and male users to perceive libraries differently. On the LibQUAL+ nonslider scores, the mean total scores across gender were trivially different (i.e., $M_F = 49.9$; $SD_F = 9.8$, versus $M_M = 50.1$; $SD_M = 10.2$, $p = .461$). Table 4 presents comparisons of LibQUAL+ subscale and total score means across frequencies of library use, across participant role groups, and across participant disciplines.

LibQUAL+ Norms

Norms are used quite frequently in education and psychology. Norms tables allow the conversion of observed scores for a person into derived scores. These tables are developed by administering a given measure to a large representative sample of a target group. For example, an educational achievement test might be administered to a normative sample of 1,000 high school seniors whose demographic profile (e.g., gender, ethnicity, geographic location) closely matches that in the most recent U.S. Census.

Once a *generic* norm table is in hand, observed scores can be converted into normative scores or standard scores. For example, if high school senior Patrick got 87 items correct out of 93, the norms table could be consulted to determine that a score of 87 in the normative sample equaled a *T*-score (i.e., scores with means of 50 and standard deviations of 10) of 73. Or the norms table might indicate that Patrick's score of 87 correct answers was higher than 93 percent of the 1,000 high school seniors in the normative sample (i.e., Patrick's percentile rank was 93).

Furthermore, *specialized* norms can also be developed. Separate educational norms are frequently provided for both urban and nonurban school districts. For example, if Patrick resided in a rural school district, the rural norms might be relevant for some interpretations. These rural norms might indicate that his 87 correct answers corresponded to a *T*-score in this normative group of 71 while his percentile rank was 90.

Table 5 presents illustrative generic norms for LibQUAL+ total scores. Similar norms could easily be derived for LibQUAL+ subscale scores. The table indicates, for example, that a LibQUAL+ total score (computed by adding together the 34 items and then dividing the sum by 34) of 6.05 equaled a *T*-score of 45.14 in the sample of 3,987 participants, which was higher than 27 percent of the 3,987 total scores.

Consider, for example, that the Table 5 norms were deemed representative of users at all ARL libraries. If, in a future sample, an ARL library received a LibQUAL+ total score of 6.65, then librarians at that institution

Table 4. LibQUAL+ Subscale and Total Score Comparisons Across Library Use, Role Groups, and Disciplines for the Nonslider Data.

Variable/ Category		LibQUAL+ Subscales									
		Service		Library as Place		Access to Collections		Reliability		LibQUAL+ TOTAL	
	n	M	(SD)	M	(SD)	M	(SD)	M	(SD)	M	(SD)
Use											
Daily	660	50.37	(10.03)	9.55	(10.20)	51.46	(9.65)	49.77	(10.41)	50.25	(10.13)
Weekly	1574	50.10	(10.08)	49.95	(10.25)	49.77	(10.37)	50.40	(9.98)	50.06	(10.17)
Monthly	1008	50.25	(9.54)	50.53	(9.70)	50.06	(9.65)	50.21	(9.39)	50.34	(9.55)
Quarterly	455	49.41	(10.75)	50.38	(9.91)	49.18	(10.32)	49.29	(10.54)	49.57	(10.67)
Never	51	46.58	(10.21)	48.15	(9.97)	45.96	(10.62)	45.33	(9.83)	46.14	(9.25)
p	0.060		0.172		<.0001		0.002		0.044		
eta^2	0.2%		0.2%		0.7%		0.4%		0.3%		
Role											
Other	126	52.64	(9.25)	51.01	(9.95)	52.46	(9.09)	51.73	(9.54)	52.26	(9.81)
Under-graduate	998	48.64	(10.49)	53.09	(9.23)	51.08	(8.93)	49.01	(9.66)	50.59	(9.90)
Research Scientist	23	49.90	(8.58)	51.75	(9.29)	50.15	(9.34)	48.92	(10.01)	50.37	(9.35)
Graduate	1281	50.25	(9.88)	49.88	(9.87)	49.91	(10.28)	50.78	(10.03)	50.21	(10.01)
Librarian	537	50.50	(8.48)	48.53	(9.76)	51.47	(8.64)	48.71	(9.05)	49.70	(8.98)
Faculty	1022	50.43	(10.38)	47.73	(10.26)	47.98	(11.03)	50.48	(10.65)	49.04	(10.55)

p	<.0001		<.0001		<.0001		<.0001		0.001	
eta^2	0.8%		4.1%		1.8%		0.8%		0.5%	
Discipline										
Business	322	51.30	(9.97)	53.14	(9.66)	51.68	(9.62)	51.01	(10.34)	52.18	(10.12)
Architecture	61	52.35	(9.36)	51.89	(10.26)	50.94	(10.03)	51.86	(9.66)	52.16	(9.80)
Education	260	51.27	(11.14)	51.90	(10.35)	51.06	(10.22)	51.01	(11.01)	51.60	(11.04)
Health Science	374	50.93	(9.09)	49.59	(9.28)	50.35	(9.51)	51.18	(9.04)	50.55	(8.94)
Engineering	434	49.77	(9.49)	51.58	(9.46)	49.79	(9.30)	50.49	(9.58)	50.53	(9.51)
Fine Arts	153	49.12	(10.82)	51.64	(9.53)	50.44	(9.69)	49.62	(9.57)	50.25	(10.10)
Vet Medicine	41	51.13	(7.48)	49.37	(8.94)	49.92	(7.90)	49.88	(8.31)	50.16	(7.56)
Science	682	49.73	(10.00)	50.71	(9.40)	49.32	(10.59)	50.09	(10.17)	50.02	(9.99)
Libraries	479	50.72	(8.30)	48.63	(9.86)	51.55	(8.76)	48.99	(8.90)	49.90	(8.92)
Law	69	49.20	(9.95)	49.84	(9.94)	49.90	(9.24)	49.53	(8.41)	49.52	(9.04)
Social Science	662	49.06	(10.58)	48.80	(10.01)	48.99	(10.46)	49.49	(9.91)	48.89	(10.09)
Humanities	450	48.81	(10.99)	46.89	(11.03)	48.72	(10.80)	48.89	(11.51)	47.94	(11.30)
p	<.0001		<.0001		<.0001		0.005		<.0001	
eta^2	0.9%		3.1%		1.1%		0.7%		1.4%	

Note. "Eta^2" is a variance-accounted-for effect size. It indicates what percentage of the variance in the LibQUAL+ scores could be explained by knowledge of the user groups to which different participants belonged.

Table 5. Illustrative Table of Norms for LibQUAL+ Total Scores Based on Nonslider Data (n = 3987).

Raw Score	%tile	T Score	Raw Score	%tile	T Score	Raw Score	%tile	T Score
3.38	1	20.12	3.94	2	25.38	4.26	3	28.39
4.50	4	30.60	4.65	5	31.98	4.75	6	32.97
4.90	7	34.34	5.00	8	35.28	5.10	9	36.24
5.18	10	36.93	5.26	11	37.71	5.33	12	38.36
5.40	13	39.00	5.47	14	39.68	5.53	15	40.23
5.62	16	41.06	5.65	17	41.40	5.71	18	41.88
5.77	19	42.25	5.79	20	42.71	5.82	21	42.99
5.88	22	43.53	5.93	23	43.94	5.97	24	44.33
6.00	25	44.63	6.03	26	44.91	6.05	27	45.14
6.08	28	45.35	6.10	29	45.61	6.14	30	45.97
6.18	31	46.28	6.20	32	46.50	6.24	33	46.83
6.26	34	47.11	6.29	35	47.38	6.32	36	47.66
6.37	37	48.06	6.39	38	48.26	6.42	39	48.57
6.45	40	48.81	6.47	41	49.04	6.50	42	49.31
6.53	43	49.58	6.56	44	49.86	6.58	45	50.08
6.61	46	50.31	6.62	47	50.46	6.65	48	50.68
6.58	49	50.96	6.70	50	51.14	6.73	51	51.43
6.74	52	51.55	6.76	53	51.79	6.79	54	52.06
6.82	55	52.28	6.84	56	52.47	6.86	57	52.68
6.88	58	52.89	6.91	59	53.16	6.93	60	53.34
6.95	61	53.49	6.97	62	53.71	7.00	63	53.99
7.02	64	54.21	7.05	65	54.46	7.07	66	54.63
7.09	67	54.83	7.12	68	55.09	7.15	69	55.36
7.17	70	55.53	7.19	71	55.74	7.21	72	55.95
7.24	73	56.19	7.26	74	56.46	7.29	75	56.74
7.32	76	57.01	7.35	77	57.29	7.38	78	57.56
7.41	79	57.84	7.44	80	58.12	7.48	81	58.44
7.50	82	58.70	7.54	83	59.05	7.57	84	59.32
7.61	85	59.70	7.65	86	60.04	7.68	87	60.31
7.72	88	60.71	7.76	89	61.14	7.81	90	61.58
7.86	91	62.02	7.93	92	62.68	7.98	93	63.12
8.01	94	63.47	8.09	95	64.16	8.17	96	64.95
8.29	97	66.09	8.42	98	67.25	8.68	99	69.67

could re-express the rating as a normative score of $T = 50.68$. Furthermore, the staff could then say, "if perceptions of use were compared to those of all ARL libraries, we would score higher than approximately 48 percent of all the ratings provided in the normative sample."

To make the use of norms even more concrete, Figure 2 presents T-scores for three respondent groups for one of the schools (pseudonym "Higher University") in the LibQUAL+ phase one study. For the present heuristic purposes, imagine that the Table 5 norms and the related generic norms for the four subscales were created from an independent normative sample measured at some prior time and not created using data involving the current respondents from Higher University.

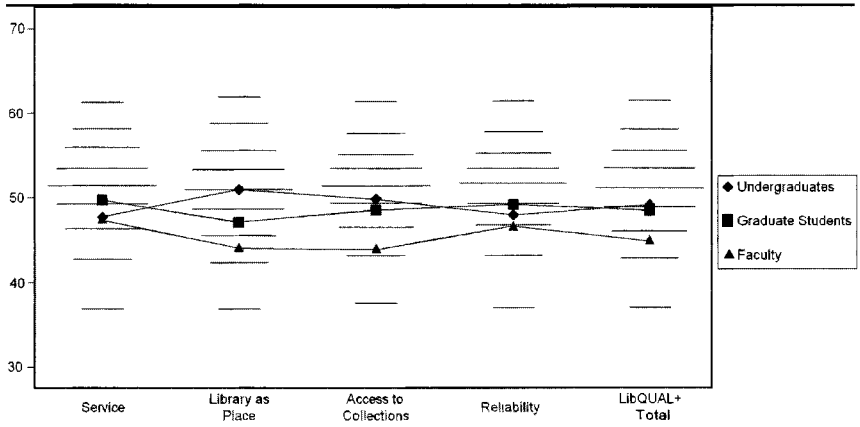

Figure 2. Hypothetical Comparisons of *T*-scores of Three User Groups at "Higher University" on LibQUAL+ Subscale and Total Scores
Note. The 10th through 90th percentiles (i.e., deciles) are indicated on each score by horizontal lines widest for the 50th percentile (i.e., the median) and narrowest for the 10th and the 90th percentiles.

The normative comparisons presented in Figure 2 suggest a number of conclusions. First, relative to the normative sample, the current respondents at Higher University rated the HU Library at or below the 50th percentile (or median) on all LibQUAL+ dimensions, including the total score. Second, respondents were most homogeneous in their ratings as regards the Service and Reliability subscales. Third, faculty were uniformly most critical of the HU Library. For example, the mean rating by faculty of Access to Collections (M = 43.9, as indicated by the location of the triangle in Figure 2) was only higher than roughly 20 percent of the ratings in the prior normative sample on this LibQUAL+ dimension.

Discussion

The present study was conducted to address five research questions:

1. Can a meaningful and replicable structure underlying user perceptions of library services be identified?
2. Can psychometrically-stable scores on LibQUAL+ dimensions be generated?
3. Are scores on different LibQUAL+ dimensions of user perceptions correlated with each other and with user overall ratings of library service quality?
4. Do comparisons of LibQUAL+ subscale and total scores across user types suggest that LibQUAL+ scores are psychometrically valid?
5. Can standardized norms potentially be developed to assist librarians in understanding user perceptions of library service quality and targeting areas of needed or desired improvement?

The answers to all five questions appear to be "yes." However, these answers warrant some further elaboration.

LibQUAL+ Dimensions

It is striking that the factor structure reported in Table 1 was generally replicated so well across the two independent subsamples. The factors appear to be meaningful. The items are generally "univocal" (i.e., "speak" primarily through a single factor). And the results are consistent with related analyses using different methods and the wider set of all forty-one items (cf. Cook, Heath, & Thompson, 2000b).

Score Reliability

As reported in Table 2, the LibQUAL+ subscale and total scores had impressive reliability coefficients. Especially noteworthy were the reliabilities for the LibQUAL+ total scores which were .952 and .958 for the slider and nonslider data, respectively.

Of course, it is important to bear in mind that tests are *not* reliable (Thompson & Vacha-Haase, 2000). As the APA Task Force on Statistical Inference recently emphasized:

> It is important to remember that a test is not reliable or unreliable. Reliability is a property of the scores on a test for a particular population of examinees . . . . Thus, authors should provide reliability coefficients of the scores for the data being analyzed even when the focus of their research is not psychometric. (Wilkinson & APA Task Force on Statistical Inference, 1999, p. 596)

The important implication is that each time LibQUAL+ is administered, it will be necessary to conduct analyses to assure that each given data set is psychometrically sound. This will be particularly important as items are added and deleted during continuing refinement of the protocol.

Score Correlations

It is certainly important that LibQUAL+ scores correlated highly with independent global ratings of library quality as reported in Table 3. And it is important that LibQUAL+ subscale scores were all highly correlated with total scores.

However, the large correlations among the LibQUAL+ subscale scores, ranging from .546 to .773, suggest that a single dimension may be used to characterize user perceptions. The "corrected" item discrimination (item-score-to-total-score correlations) presented for LibQUAL+ total scores in the last two columns of Table 2 are also consistent with this view. For the slider data, these corrected item discrimination coefficients ranged from .45 to .72, and for the nonslider data ranged from .50 to .73. The service items tended to be most highly correlated with the total scores, suggesting that perceptions of service saturate the ratings.

Tables 1, 2, and 3 suggest that users *simultaneously* think about library quality both using first-order subscale dimensions and at a second-order aggregate level. This interpretation is supported by "higher-order" factor analyses we have reported elsewhere for both these and other data (Cook, Heath, & Thompson, 2000b; Cook & Thompson, in press).

Figure 3 graphically presents a hierarchical LibQUAL+ factor model. The model posits that selected items measure one of the four first-order factors (e.g., Affect of Service, Library as Place). However, the first-order factors are themselves correlated and aggregate at the second-order level into a single overarching Service Quality perceptions factor. We believe users think simultaneously at both levels. If our view is correct, for most applications, both LibQUAL+ subscale and total scores will be necessary to summarize user perceptions.

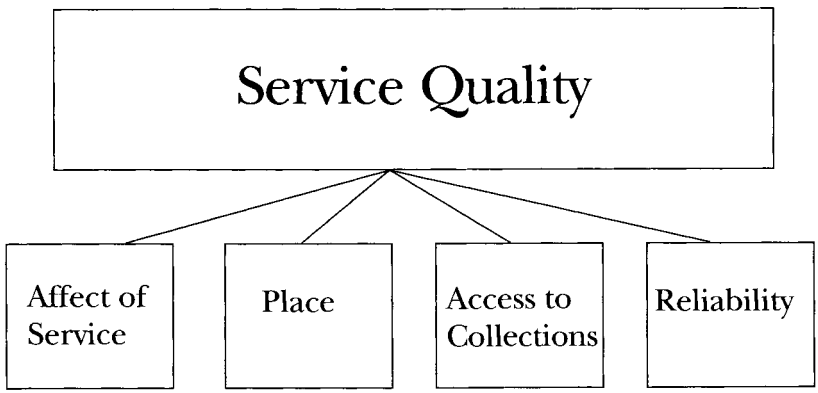

Figure 3. A Hierarchical LibQUAL+ Factor Model.

Mean Comparisons

It is heartening that, as expected, LibQUAL+ scores did not differ across gender. It is also heartening that user perceptions did not differ much across user frequency of library use, as reported in Table 4. Only users who reported using the library "never" differed appreciably in their ratings of the libraries.

Users also tended to be fairly homogeneous in their views across role groups. There was the most variation ($eta^2 = 4.1\%$) on the Library as Place subscale. The undergraduate students tended to be most favorable (*T*-score mean = 53.1) and the faculty the least favorable (*T*-score mean = 47.7) as regards this dimension.

Regarding user disciplines, observed differences were relatively small. The largest differences ($eta^2 = 3.1\%$) occurred on the Library as Place subscale. Business respondents were most positive (*T*-score mean = 53.1)

and Humanities respondents were most negative (*T*-score mean = 46.9) on this dimension.

Regarding comparisons across subscales, the 3,879 respondents rated all four dimensions fairly highly, as reported in Figure 1. However, respondents were somewhat more homogeneous and rated somewhat more highly perceived Service and perceived Reliability. It is noteworthy that Service and Reliability items tended to be most highly correlated with LibQUAL+ total scores, as reported in Table 2.

LibQUAL+ Norms

Table 5 and Figure 2 illustrate the development and use of norms for LibQUAL+. Although only *generic* norms for total scores were presented in Table 5, *generic* norms were also computed for the four LibQUAL+ subscales. Furthermore, *specialized* norms have been developed and may also be useful. For example, norms can be developed by (a) user group (e.g., faculty, graduate students), (b) discipline, or (c) campus type or setting (e.g., urban, private).

The potential to develop norms for specialized comparisons across ARL members hints at the potential of the LibQUAL+ protocol. If both generic and specialized norms were eventually developed for a large sample of users at ARL institutions, LibQUAL+ could then be used to make a series of intelligent comparisons with various reference groups. Such comparisons could then facilitate the ultimate application of LibQUAL+: identifying areas of potential improvement at a given library and identifying similar libraries with more favorable profiles whose behavior might then be modeled in pursuit of providing better service to library users.

REFERENCES

Arnau, R. C.; Thompson, R. L.; & Cook, C. (2001). Do different response formats change the latent structure of responses? An empirical investigation using taxonometric analysis. *Educational and Psychological Measurement, 61*, 23-44.

Cook, C., & Heath, F. (2000a). The Association of Research Libraries LibQUAL+ Project: An update. *ARL Newsletter: A Bimonthly Report on Research Library Issues and Actions from ARL, CNI, and SPARC, 211*(July), 12-14.

Cook, C., & Heath, F. (2000b). *Users' perceptions of library service quality: A "LibQUAL+" qualitative interview study.* Unpublished paper presented at the Association of Research Libraries (ARL) Measuring Service Quality Symposium on the New Culture of Assessment: Measuring Service Quality, October, Washington, DC.

Cook, C.; Heath, F.; & Thompson, B. (2000a). LibQUAL+: One instrument in the New Measures toolbox. *ARL Newsletter: A Bimonthly Report on Research Library Issues and Actions from ARL, CNI, and SPARC, 212.* Retrieved March 26, 2001 from the World Wide Web: http://www.arl.org/newsltr/212/libqual.html.

Cook, C.; Heath, F.; & Thompson, B. (2000b). *Users' hierarchical perspectives on library service quality: A "LibQUAL+" study.* Unpublished paper presented at the Association of Research Libraries (ARL) Measuring Service Quality Symposium on the New Culture of Assessment: Measuring Service Quality, October, Washington, DC.

Cook, C.; Heath, F.; & Thompson, R. L. (2001). A meta-analysis of response rates in Web- or Internet-based surveys. *Educational and Psychological Measurement, 60*, 821-836.

Cook, C.; Heath, F.; Thompson, B.; & Thompson, R. L. (in press–a). The search for new measures: The ARL "LibQUAL+" study—a preliminary report. *Portal: Libraries and the Academy.*

Cook, C.; Heath, F.; Thompson, R. L.; & Thompson, B. (in press–b). Score reliabilities in Web- or Internet-based surveys: Unnumbered graphic rating scales versus Likert scales. *Educational and Psychological Measurement.*

Cook, C., & Thompson, B. (2000). Reliability and validity of SERVQUAL scores used to evaluate perceptions of library service quality. *Journal of Academic Librarianship, 26*(4), 248-258.

Cook, C., & Thompson, B. (in press). Higher-order factor analytic perspectives on users' perceptions of library service quality. *Library & Information Science Research.*

Gorsuch, R.L. (1983). *Factor analysis* (2d ed.). Hillsdale, NJ: Erlbaum.

Hendrick, C., & Hendrick, S. (1986). A theory and method of love. *Journal of Personality and Social Psychology, 50*(2), 392-402.

Hernon, P., & McClure, C. R. (1990). *Evaluation and library decision making.* Norwood, NJ: Ablex.

Hetzel, R. D. (1996). A primer on factor analysis with comments on analysis and interpretation patterns. In B. Thompson (Ed.), *Advances in social science methodology* (vol. 4, pp. 175-206). Greenwich, CT: JAI Press.

Krosnick, J. (1999). Survey research. *Annual Review of Psychology, 50,* 537-567.

Nitecki, D. A. (1995). An assessment of the applicability of SERVQUAL dimensions: A customer-based criteria for evaluating quality of services in an academic library. Doctoral dissertation, University of Maryland, 1995. *Dissertation Abstracts International, 56*(8), 2918A(University Microfilms No. 95-39, 711).

Nitecki, D. A. (1996). Changing the concept and measure of service quality in academic libraries. *Journal of Academic Librarianship, 22*(3), 181-190.

Nitecki, D. A., & Hernon, P. (2000). Measuring service quality at Yale University's libraries. *Journal of Academic Librarianship, 26*(4), 259-273.

Nunnally, J. C. (1978). *Psychometric theory* (2d ed.). New York: McGraw-Hill.

Parasuraman, A.; Berry, L. L.; & Zeithaml, V. A. (1991). Refinement and reassessment of the SERVQUAL scale. *Journal of Retailing, 67*(4), 420-450.

Parasuraman, A.; Zeithaml, V. A.; & Berry, L. L. (1985). A conceptual model of service quality and its implications for future research. *Journal of Marketing, 49,* 41-50.

Parasuraman, A.; Zeithaml, V. A.; & Berry, L. L. (1994). Alternative scales for measuring service quality: A comparative assessment based on psychometric and diagnostic criteria. *Journal of Retailing, 70*(3), 201-230.

Pedhazur, E. J., & Schmelkin, L. P. (1991). *Measurement, design, and analysis: An integrated approach.* Hillsdale, NJ: Erlbaum.

Reinhardt, B. (1996). Factors affecting coefficient alpha: A mini Monte Carlo study. In B. Thompson (Ed.), *Advances in social science methodology* (vol. 4, pp. 3-20). Greenwich, CT: JAI Press.

Thompson, B. (2000). *Representativeness versus response rate: It ain't the response rate!* Unpublished paper presented at the Association of Research Libraries (ARL) Measuring Service Quality Symposium on the New Culture of Assessment: Measuring Service Quality, October, Washington, DC.

Thompson, B.; Cook, C.; & Heath, F. (in press). How many dimensions does it take to measure users' perceptions of libraries? A "LibQUAL+" study. *Portal: Libraries and the Academy.*

Thompson, B., & Daniel, L. G. (1996). Factor analytic evidence for the construct validity of scores: An historical overview and some guidelines. *Educational and Psychological Measurement, 56*(2), 197-209.

Thompson, B., & Levitov, J. E. (1985). Using microcomputers to score and evaluate test items. *Collegiate Microcomputer, 3,* 163-168.

Thompson, B., & Vacha-Haase, T. (2000). Psychometrics *is* datametrics: The test is not reliable. *Educational and Psychological Measurement, 60*(2), 174-195.

Wilkinson, L., & APA Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54*(8), 594-

This Page Intentionally Left Blank