Development and Testing of a Reference Transaction Assessment Instrument

Marjorie E. Murfin and Gary M. Gugelchuk

An instrument was designed to assess the reference outcome in terms of patron report of success. It was theorized that patrons were capable of separating reports of success in finding what was wanted from reports of general overall satisfaction, if given a suitably constructed instrument. Utilized in design were Cronbach's theories on response set and Cuadra's theories on the fail-safe reaction.

A survey was conducted. Fifteen academic libraries across the country participated. Their collections ranged from 56,000 to 4 million volumes. A total of 531 transactions with matching patron and librarian forms were analyzed in terms of 35 variables. The reliability of the instrument was found to be .81 overall—.89 for the patron form and .76 for the librarian form. It was also found that patron report of success on factual questions did not differ significantly from results obtained from unobtrusive observation.

The primary source of pride and satisfaction for the professional reference librarian is the quality of reference service he or she provides. Although quality reference service is a desired goal, its definition and assessment have been fraught with difficulty. Criteria for reference service quality abound in the literature, and numerous evaluation plans have been attempted to measure and assess quality. These valuable forays into the poorly mapped-out territory of reference service appear to some observers to exemplify the legend of the ten blind men trying to define an elephant by only touching its varied parts—the methodologies of evaluation seem restricted, and the evaluation criteria partial and incomplete.

Before 1960, evaluation of the actual outcome of the reference transaction was not a pressing issue because it was generally believed, on the basis of previous studies, that the outcome was highly satisfactory to reference patrons.[1] Summarizing these studies in 1964, Rothstein stated that "results could hardly be bettered by paying for testimonials."[2] However, in 1970, with the advent of unobtrusive observation as a research methodology, the outcome of the reference transaction began to be seriously questioned. Controlled studies using this method revealed that reference librarian success at answering factual

questions averaged around 50%.[3] Reference librarians were hard pressed to reconcile the favorable attitudes of their patrons with this low success rate.

In response to this dilemma, there has been an increased interest in research on the nature of reference service, and as a result of this, some important advances have been made. These include the definition of the reference question and its incorporation into a standard by Emerson and the LAMA Committee on Statistics for Reference Service,[4] the formulation of mutually exclusive categories of questions by Hieber,[5] the work of identifying reference process factors by Jahoda and Braunagel,[6] the considerable methodological contributions of Hallman to subject recording and data collection,[7] the seperation of outcomes of the reference transaction by Strong,[8] Howell and others,[9] Tessier and others,[10] Tagliacozzo,[11] and Gothberg,[12] and the work on scale construction for relevance judgments by Cuadra and others.[13]

Guided by these research breakthroughs, the authors have undertaken to design and develop a survey instrument and a data collection protocol that would assess the outcome of the reference transaction and identify factors related to successful and nonsuccessful performance. We also hoped to design this instrument in such a way that it would be practical for widespread use while demonstrating both validity and reliability. As far as the authors can determine, patron response to the traditional reference transaction has never been measured by use of a uniform instrument that (1) is designed to measure more than one or two types of outcome; (2) utilizes carefully constructed scales with clearly demonstrated power of discrimination; (3) has a sufficient number of items to demonstrate reliability; and (4) has been administered in a number of libraries under carefully controlled conditions. These are standards basic to measurement research in other disciplines, yet not been consistently utilized in reference research.

Before presenting the design and field-testing results of the reference transaction assessment instrument (RTAI), the research issues involved will be examined. These issues can best be understood in the historical context of evaluation research in reference.

THE LITERATURE AND METHODOLOGY OF REFERENCE EVALUATION

Between 1976 and 1982, growing pressures to evaluate reference service resulted in an outpouring of some 238 publications on assessment of reference service. Ninety-two of these items stressed the need for evaluation, suggested criteria, and proposed ideas on evaluation while 146 were studies of various aspects of reference service. In addition, some 60 items were written on communication and interaction.[14] The spate of publication has continued to the present day.

One trend apparent in the literature during this period sought the solution to reference service problems in the individual reference librarian.[15] Studies of reference personnel increased tenfold from 1976 through 1982, covering such aspects as personality characteristics, concepts and attitudes, cognitive styles, and a variety of other areas. Also explored was the role of the reference librarian in the communication process.[16]

Another trend since 1975 has been a greater emphasis on outcome studies, with some 29 appearing in the literature. Of these, 55% utilized unobtrusive observation or judgment of outcome by outside experts, 21% librarian judgment only, and 24% judgment by patron/librarian.[17] Studies of the outcome of online searching also began to appear frequently during this period, but in contrast to outcome studies of traditional reference service, some 81% utilized patron judgment, 5% patron/librarian judgment, and 14% focused on measures of precision and recall.[18]

This volume of publication, in the opinion of some experts, has not been associated with substantial progress. In recent years, Altman notes, "there is still no consensus on how to measure, let alone evaluate, reference services."[19] Mary George,

a reference head, regrets that there are no widely accepted evaluation methods on which to draw.[20]

"Altman notes, 'there is still no consensus on how to measure, let alone evaluate, reference services.' "

Various other reference experts discuss reasons why progress in development of reference evaluation methodology may have been slow. For instance, some claim that not all relevant aspects of the reference transaction have been properly studied. Ellsworth and Joan Mason comment:

. . . no consideration at all [has been paid] to the wide range of other factors that can make or break reference service. . . . Many things besides eye contact or body language speak volumes to library users about the reference department's intentions toward them.[21]

A second problem has been that indepth knowledge of the measurement process and of the reference process have seldom been found together in a single individual. Cross-disciplinary research is needed but has not taken place frequently enough. Zweizig comments on this, noting that reference managers do not feel that they understand the measurement process and that those who design measures do not understand the reference process.[22]

Unobtrusive Evaluation

In spite of these problems and concerns, some progress has been made and valuable insights gained by research done during this period. Perhaps the greatest single advance has been the development of the method of unobtrusive observation as a means of assessing the outcome of the reference transaction. In unobtrusive observation, proxies present factual test questions at the reference desk by phone or in person. Success is then judged by the percentage of correct answers obtained. Research studies using unobtrusive observation appear to have nearly doubled

since 1975,[23] and, in one instance, unobtrusive observation has been adapted for use as a performance measure in a single library.[24] The authors agree with McClure that, although unobtrusive testing is but one approach to the evaluation of reference service, it is a critically essential component.[25]

Unobtrusive observation, however, has certain limitations, as has been noted by various reference experts. Test questions present a problem because they must be of a certain level of difficulty, must be representative, can't easily be repeated, and take considerable time and effort to formulate. Results are usually not available for some time and are difficult to generalize to the department as a whole.[26]

Also, this approach is generally biased toward factual questions and does not assess the full range of questions asked by patrons. Nor does it usually take into account the difficulties of garbled and incoherent questions as often presented by actual patrons. Finally, this method ordinarily admits of only right or wrong answers and generally does not allow for the other less clear-cut outcomes, such as finding answers within an acceptable range or not finding what was sought but instead finding other useful information or materials.

For these reasons, practical methods of reference assessment in addition to unobtrusive observation are still needed. This need for a variety of methods is stressed by Crowley, among others, who notes the advantages of independent multiple approaches to measuring reference capacity.[27] Pierce also supports this when she notes that any library evaluating its reference services will find any method chosen lacking in some respects.[28]

One methodology available for use in assessing reference department quality is the questionniare survey. This approach is commonly used in sociological research, public opinion polling, and marketing, to name but a few contexts. Over the years there has been a vast improvement in the design of questionnaires to make results more reliable and valid. Thus, a suitably designed questionnaire might also work

well for evaluation of the reference transaction.

CRITERIA FOR A REFERENCE ASSESSMENT INSTRUMENT AND DESIGN OF THE RTAI

The literature of reference evaluation suggests a number of desired qualities for an effective reference assessment tool. An effective assessment instrument should

A. Demonstrate validity;
B. Demonstrate reliability;
C. Utilize the natural setting;
D. Assess multiple factors in the reference process; and
E. Include sufficient input, process, and outcome factors so that cause and effect relationships may become apparent.

In addition, the use of a standardized form to collect and report data should

F. Utilize an adequate and unbiased sample;
G. Provide for comparability of data;
H. Provide for timeliness of results; and
I. Provide for interpretability.

Each of these items will be considered below as it relates to the design of our reference transaction assessment instrument (RTAI) and data collection protocol.

Construct Validity and the Influence of Positive Bias

Perhaps the most important quality for any assessment instrument is construct validity, defined as the ability of an instrument to measure the characteristics it is intended to measure. Roloff notes in regard to reference research that ''Researchers rarely indicate reliability and validity measures . . . absence of descriptions of such fundamental concepts . . . is distressing.''[29]

As mentioned previously, a major question that arises in the area of validity is related to patron report of outcome. Why have so many studies found high levels of patron satisfaction when, on the other hand, unobtrusive observation suggests that only about 50% of reference questions are answered correctly?

Many in the profession have hypothesized that patron evaluation is not valid because patrons are biased in a positive di-

rection toward the reference librarian and allow this bias to influence their rating of success in finding information. These opinions are summarized by Young, who says:

The very fact of receiving help or the pleasant demeanor of the reference librarian, rather than accuracy of information, will inevitably influence patron perception of service.[30]

Others express similar concerns, including Rothstein, Zweizig, Pierce, Altman, Kantor, Vathis, Weech, and McClure.[31] These concerns are based on results of past studies that are described by Rothstein and Weech in their reviews of research in reference.[32]

''Satisfaction with service did not influence patrons' rating of the information found.''

On the other hand, it is important to note that Gothberg, using an instrument where outcomes of (1) satisfaction with the interview, and (2) satisfaction with the transfer of information were separated, found that satisfaction with service did not influence patrons' rating of the information found.[33] This supports the contention that unrealistically high ratings found in previous studies were due in great part to the use of inadequate instruments and methods to study a complex phenomenon. Other researchers also note imprecision in previous measurement instruments. Certain common types of imprecision may actually encourage positive bias and lead to inflated ratings.

The Too-Broad Question. One example of imprecise measurement is the use of a broad, general satisfaction question not tied to any particular aspect of the reference transaction. That is, the satisfaction question was not linked specifically to the information actually provided. Thus, responses represented only broad, general satisfaction with the entire transaction or experience. Tagliacozzo also points out the ineffectiveness of the broad satisfaction question when she notes, ''Especially

those questions which elicit general overall judgments of the service provided by the system should be looked at suspiciously."[34]

The Single Question. In addition to the too-broad question, another reason for lack of precision and inflated ratings in previous attempts to measure satisfaction may have been the use of a *single* question that did not take into account the multiple outcomes of the reference transaction. Tagliacozzo emphasizes this when she comments "one should be wary of taking the data of a single rating scale to assess the value that an information system has for its users."[35] The importance of this is stressed by Cronbach, who recommends measuring all outcomes that might be worth striving for.[36] With regard to outcomes of the information interview in particular, Auster says that satisfaction is multidimensional. She notes that a set of questions is required, rather than a single question, and that we must identify the distinct aspects and tap the dimensions of satisfaction identified in the literature.[37] Tessier, Crouch, and Atherton (Cochrane) discuss a variety of possible outcomes of the information transaction.[39]

In designing the patron form, particular care was taken to prevent the blurring together of different outcomes in order to increase the likelihood of valid ratings. Separate scales were made for (1) success in finding what was wanted; (2) satisfaction with what was found or not found; (3) satisfaction with helpfulness, courtesy, etc.; and (4) amount learned. Other outcomes, such as communication difficulty, could also be measured by combining separate items. A copy of these forms may be seen in appendix A.

However, even avoiding the above common types of imprecision is probably not sufficient in and of itself to ensure validity of response. Positive bias toward the reference librarian does exist and must be actively controlled through sophisticated design. More specifically, positive bias toward socially desirable types of answers has been known to measurement experts for some time as one type of response set. Mehrens and Lehman note that steps can be taken to try to control for response set.[39]

Cronbach notes that response set is particularly prevalent on tests that (1) contain ambiguous items; (2) require the individual to respond on a disagree-agree continuum; and (3) lend themselves to responses in a favorable or unfavorable direction.

Best Answer Anchoring. One technique suggested by Cronbach is the best answer choice.[40] This technique has been utilized in the patron scale for finding what was wanted, where six choices of answer (descriptive phrases not necessarily on a continuum) are presented (see appendix B). Using versions of this technique, Cuadra and others found, in studies of anchored versus unanchored scales, that anchoring the midpoint on the scale with a phrase tied to a particular reference point produced lower ratings and was especially effective in producing differentiated responses. Cuadra notes "it might prove especially effective to add additional anchors at each scale extreme, to obtain the desired greater difference effect."[41] Loftus and others also remark on the effectiveness of anchored scales when they note, "if the task required is one of estimation, we can use our knowledge of the effects of anchoring and other heuristics to guide the respondent's judgment."[42]

Specificity of Scale Items. Cuadra and Katter related positive bias on the part of raters to the specificity of scale items. They found that if a rater was uncertain about how to represent the rating on a particular scale, the tendency was to assign a higher rating. A study was done in which some 24 subjects judged the relevance of 12 articles to a certain statement of requirements. These judgments were made under two different conditions. One condition involved a very specific statement of requirements, the other a more general statement of requirements. Using the general statement, the subjects expressed greater uncertainty and tended to give higher relevance ratings. Using the more specific statement, they expressed less uncertainty and tended to give lower relevance ratings. Cuadra terms this the fail-safe reaction and discusses it further.[43]

Number of Categories in Scales. In one study, Cuadra showed that certainty of

judgment at the ends of the scale is much greater when there are at least six to eight categories in the scale. Six to eight category scales were shown to have superior information-preserving qualities over two and four category sales without apparent disadvantages.[44]

"In the natural setting, patron input is an essential dimension of evaluation of reference service."

In line with these research findings and to encourage valid responses the following goals were set for the design of the RTAI:

1. Control possible bias in favor of the reference librarian (as a reward for helpfulness) by effectively separating ratings of service from those of success in finding what was wanted.

2. Prevent blurring of the rating of success with other outcomes by effectively providing separate ratings for (a) success in finding what was wanted, (b) satisfaction with the information/materials found or not found, (c) satisfaction with service (helpfulness, etc.), (d) amount learned, and (e) communication difficulty.

3. Prevent rater uncertainty in judging success by using precise and specific language in items.

4. Prevent rater uncertainty in judging success by use of a scale with six anchored points.

Additional controls for positive bias were applied by establishing a rigorous scoring system. The criterion for complete success required three conditions: (1) that the patron mark "found just what was wanted" on the success scale, (2) that the patron mark yes on the satisfaction scale, and (3) that there be no marks for any of the nine listed reasons for dissatisfaction. Thus, each transaction had to meet a three-point criterion in order to be judged completely successful.

Similarly, in regard to the six service items (understood what was wanted, enough help, time, courtesy, clear expla-

nations, knowledge), a partly negative mark by the patron was grouped with fully negative marks as "less than completely satisfactory" on that item. So it can be seen that the standard of judgment was set at a high level.

Reliability

Mehrens and Lehman note, "In general longer tests are more reliable. Test length influences reliability."[45] The RTAI includes multiple factors on both patron and librarian forms that, it is hoped, will enhance reliability and permit cross checking of results. Results of reliability tests on the RTAI are treated in detail in the following section on pilot testing.

Utilization of the Natural Setting

Shavit says:

Qualitative research has the natural setting as the direct source of data. . . . Reference service can best be understood when it is observed in the setting in which it occurs.[46]

In the natural setting, patron input is an essential dimension of evaluation of reference service. Cronin and others support this conclusion:

It would only be possible to carry out a proper evaluation exercise by enlisting the cooperation of the user. . . . Few evaluation studies can be conducted without . . . the views and requirements of the end user being taken into account.[47]

However, discussions such as those between Cooper[48] and Soergel[49] argue the issue of whether patron judgment of success is sufficient, in and of itself, as the criterion for success. One argument given against use of the patron's report of success is that it often doesn't represent a judgment of the best possible information or material that *could have* been provided, as judged against an external criterion of quality. We believe this to be true, and this dimension of reference evaluation must be left to expert peer review.

On the other hand, patrons judge the success and satisfactoriness of the information or material on how well these meet their own complex pattern of internal and external needs, purposes, and states of knowledge. The work of Cuadra and Kat-

ter illustrates this in demonstrating that the patron's purpose or intended use for information strongly influences his or her judgment of relevance, as, for example, self-generated needs for certain materials, following someone else's instructions, prejudices against certain sources or types of materials or against certain styles of presentations, and certain knowledge or lack of knowledge that distorts the interpretation of certain kinds of material.[50]

This suggests that unless we fully understand the patron's state of knowledge, purpose, and intended use for information, we cannot adequately judge how successfully his or her needs have been met. The patron is the only one in a position to know his or her own complex pattern of needs, motivations, abilities, and situational constraints, many of which are not expressed and can never be grasped in their entirety by the reference librarian or intermediary. Tagliacozzo emphasizes the need for patron judgment of outcomes when she notes, "Perhaps we will never be able to determine what the real information needs of the users are and should therefore content ourselves with what the users believe—and tell us—that they are."[51]

Unless we are unfailingly expert at interviewing and have limitless time at our disposal, we may miss some key aspect or fact that renders our interpretation of need wrong. If the patron appears to be dissatisfied with good material or is satisfied with scanty material it is just as likely that we have not uncovered some key constraint or requirement as that this response is due to the patron's inability to judge information critically.[52]

Others have, in fact, advocated or used a measure of user satisfaction. These include Pritchard, Auckland, and Castens,[53] and Rzasa and Baker,[54] who considered user satisfaction one component of effectiveness. Lancaster defines effectiveness as "to what extent a service satisfies its user's demands."[55] The Consortium for Public Library Innovation also includes user satisfaction as a measure of effectiveness[56] as did Strong[57] in his study. Rzasa and Baker note that "the conclusion can be drawn that a measure of effectiveness

must adequately reflect satisfaction of user needs."[58]

Multidimensional Nature of the Reference Process

The reference process is complex, consisting of a considerable number of input, process, and outcome factors. Reference experts stress the need to take these multiple factors into account. For example, the interview process alone is recognized in other disciplines as extremely complex. Pittenger, Hockett, and Danehy devote 81 pages of text to analysis of only the first five minutes of the therapeutic interview.[59] Taylor underscores this by noting that the reference interview is one of the most complex acts of human communication for "in this act, one person tries to describe for another person, not something he knows, but rather something he does not know."[60]

As for the complexity of the process overall, Lancaster identifies at least 55 factors[61] and S. D. Neill identifies some 66 factors that may play a part in the reference process at one time or another.[62] In order for a reference transaction assessment instrument to have content validity, it would need to represent this complexity. Dick and Hagerty define content validity as how well items selected for inclusion in an instrument represent the actual situation being considered.[63] Mehrens and Lehmann note that poor content validity is always deplorable but particularly bad when results are to be used for purposes of accountability.[64]

Analysis of Cause and Effect

Shavit summarizes well when he comments that the question is "not only how accurate is the reference librarian's response . . . but also why is the answer inaccurate?" He adds that we need to know what causes searches to fail.[65] Lancaster, McClure, and Pierce also support the need to identify factors associated with failure.[66]

In order to link cause and effect, it is necessary to include as many input factors as possible in a reference evaluation instrument because it cannot be known ahead of time with certainty which inputs will later prove to be causal factors. All outcomes

should also be included so that complex patterns of cause and effect can be discerned. If cause-and-effect relationships cannot be discerned, it is often difficult to interpret results, and the likelihood of misinterpretation is increased. Mehrens and Lehmann note that to establish causal relationships, certain variables must be taken into account, including input, surrounding conditions, process, and outcome.[67]

"Norm-referenced measurement can often assist in decision making and is necessary both for program evaluation and for making predictions."

In designing the RTAI, this model was followed, and an attempt was made to represent multiple input, condition, process, and outcome factors. A librarian form was designed based on input and process variables suggested by previous studies and by other reference practitioners and experts[68] and judged by the authors as most likely to affect the outcome of the reference transaction. This method of constructing an instrument, termed *common sense* or *deductive*, is supported by Burisch, who says "[there is] no consistent superiority of any strategy in terms of validity or predictive effectiveness . . . deductive scales normally communicate more directly . . . whenever there is a genuine choice, the simple deductive approach is recommended."[69]

Input factors in this model included type of staff handling questions, patron status and discipline, type and subject of question, and importance of question to patron. Process factors included response (directing and suggesting only or helping with the search), time taken, instruction given, and number and types of sources used. Factors related to surroundings were included in a 17-item checklist of problems. As further support for the selection of these factors, mobility or helping with the search has been shown by Gers and Seward to affect the outcome of the reference transaction.[70] A copy of the

librarian form may be seen in appendix A. The authors were aware that the librarian's recording task would be increased to some extent by a librarian form with multiple factors. But it was felt that the greater good would come of being able to link cause and effect, thus tending to prevent harmful misinterpretation of results. Also, inclusion of multiple factors should contribute to reliability and permit cross-checking of results.

Utilization of an Adequate and Unbiased Sample

The sample utilized should be, as Roloff notes, both adequate to justify generalization and selected so as to avoid potential bias.[71] To counteract the possibility of bias, we required that all transactions during the sample period, both reference and directional, be recorded. Directional transactions were clearly defined, and all other questions were to be recorded as reference, in line with the definition as outlined in ANSI Standard Z239.7—1983.[72]

Comparability

Brown notes that "without a common language and common understandings . . . we cannot effectively and efficiently address our common problems. . . . It is becoming less and less possible for the individual library to exist in isolation from others."[73] The RTAI addresses this need for comparability by providing norms.

Mehrens and Lehmann argue the merits of norm-referenced instruments when they note that without norms for purposes of comparison, we have no clear idea of the meaning of a person's score in and of itself. Norm-referenced measurement can often assist in decision making and is necessary both for program evaluation and for making predictions. They note that norms or "group means are always more reliable than individual scores because the random errors of the individual scores tend to cancel themselves out, thus making the means reasonably accurate."[74] The RTAI is a norm-referenced survey instrument designed to provide a profile for an individual library, with data for each item in the profile for (1) all libraries of the same size (and type), (2) the top-scoring library

in that size group (three size groups), and (3) all libraries in the sample, based upon use of a standardized form and sampling strategy.

Timeliness of Results

Zweizig notes that situations should be avoided in which data arrive too late to affect decisions.[75] By utilizing the technology of computerized scanning, the RTAI should have a turnaround time of several weeks at the most.

Interpretability

Shavit, following John van Maanen, notes that high technical sophistication often makes research results incomprehensible to all but the highly trained few.[76] Brown notes that "data should be organized into a framework intelligible to both lay and library persons."[77] The RTAI reports all scores in terms of simple percentages. Thus, if a percentage as a criterion for reference department performance were to be established in the future, the RTAI could be used as a criterion-referenced instrument.[78] Mehrens and Lehmann note that the most meaningful test should be related to both norms and criteria.[79] They also note that presenting a number of scores in profile form, as in the RTAI, is a useful aid in interpretation.[80]

In summary, the RTAI was designed over a period of several years, utilizing the first author's own reference work as a basis and in consultation with the second author with regard to measurement and statistical questions. It was hoped that the RTAI would be able to be used by a library for the following purposes:

1. To provide feedback from the patron as to whether his or her particular needs were met.

2. To identify factors associated with success or failure on each individual transaction.

3. To assess quality of service on all types of questions, rather than on factual questions alone.

4. To assess quality of service in the actual reference setting, using the questions asked by a library's own patrons.

5. To provide feedback on a variety of outcomes of the reference transaction, such as success; satisfaction; quality of helpfulness, courtesy, etc.; amount learned; quality of communication and the interview.

6. To compare other factors, as well as success, with other libraries of the same size and type.

7. To provide feedback to those in charge from librarians on factors that they report as having hindered their efforts on a particular transaction.

PILOT-TESTING THE REFERENCE TRANSACTION ASSESSMENT INSTRUMENT (RTAI)

In order to evaluate the reference transaction assessment instrument and its proposed method of utilization, heads of reference departments across the country were asked to volunteer in a pilot-test of the procedure. Of the libraries volunteering to participate in this pilot test, one public and 17 academic libraries were selected in order to vary as much as possible the range of institutional contexts in which the assessment scheme would be used. These 18 libraries ranged in collection sizes from 56,000 to 4 million volumes and included both publicly and privately funded institutions located in regions of the country from California to New England. Also included in these 17 academic libraries were one medical school and one undergraduate library.

Because of the particular natures of the public, undergraduate, and medical libraries, it was decided to eliminate them from the results presented here. Thus, the data presented in the remainder of this paper are restricted to the 15 academic library reference departments contained in our sample and described below in table 1.

Each library received a packet of 50 forms for reference questions and 50 forms for directional questions, plus detailed instructions. Reference forms consisted of a checklist to be marked by the reference staff member for each question and a tear-off sheet to be marked by the patron for the same question. The two forms were machine coded so that librarian and patron responses could be matched after the forms were separated. A copy of the assessment instrument appears as appendix A.

Each library was instructed to adminis-

TABLE 1

CHARACTERISTICS OF LIBRARIES
PARTICIPATING IN PILOT TEST COLLECTION SIZE

	Small 500,000 vols.	Medium 500,000–1,000,000 vols.	Large 1,000,000 vols.
Publicly funded	1	4	4
Privately funded	3	1	2
Total number of transactions	138	179	214

ter the forms on a typical day in the Autumn quarter and to record each transaction until all 50 forms were administered. After a request for assistance had been made, the patron was informed that the reference department was conducting a survey and was asked to fill out a brief checklist at a later time and deposit the completed form in a labeled container some distance away. The patron portion of the form was then torn off and given to the patron. The staff member noted the question in the upper right-hand corner of the librarian's form and completed the rest of the items on the form as soon as possible after the transaction had been completed.[81]

Upon completion of the 50 forms, librarian and patron forms were collected and mailed to the Ohio State University. The forms were computer scanned, and data were read onto magnetic tapes. The computer also matched serial numbers from the librarian and patron forms so that both halves of the transaction could be reunited.

Return Rate

In the 15 participating libraries, 88.46% of all forms given out were returned. The number of transactions per library for which both the patron and librarian halves of the form were returned ranged from a low of 80% to a high of 97.67%.

Reliability

Reliability of measures, the degree to which replication of measurement yields the same result, is a key concern in the formation of any measuring instrument. By definition, the best way to obtain an estimate of the reliability of an instrument is via the test-retest method. Because of the time and costs involved in conducting such a study with the reference transaction assessment instrument, an alternate approach to estimating the instrument's reliability was employed: an analysis of the intercorrelation of multiple, parallel measures.

To this end, the Reference Transaction Assessment Instrument was partitioned into a set of 35 variables that appeared to be distinct measures of the quality of library reference service. The definitions of these 35 variables are found in appendix B, along with the method of their quantification. Variables whose manner of quantification was expected to correlate negatively with quality of library reference service were inverted to correlate positively. In addition, examination of the distributions of the variables derived from the 531 responses in our sample revealed that several of them were skewed with respect to a normal distribution (that is, the bell-shaped curve). These variables were transformed to make the shape of their distributions more closely conform to a normal distribution. After performing some necessary modifications, Pearson zero-order product-moment correlations were computed for each pair of these 35 variables. The resulting correlation matrix appears as figure 1.

As an initial measure of the instrument's reliability. Cronbach's alpha, a commonly used estimator of reliability, was computed.[82] Cronbach's $\alpha = N\varrho/1+(n-1)\varrho$, where ϱ is the average interitem correlation and N is the total number of variables, 35 in our case. The average interitem correlation for the matrix in figure 1 is 0.1128, which yields a value for Cronbach's α of 0.81652.[83] Although the average interitem correlation

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1.	TOTCKPPB	---															
2.	QT	.05	---														
3.	ZRT	.13	-.04	---													
4.	ZLRS	-.12	-.02	-.01	---												
5.	ZBD	.06	.07	.04	.37	---											
6.	ZI	.34	.10	.24	-.25	.05	---										
7.	TOTCKLR3	.62	.08	.77	-.09	.36	.46	---									
8.	TOTCKSPB	.21	.08	.17	-.10	.02	.34	.50	---								
9.	TCTCKCFB	.12	.08	.30	.06	.07	.19	.42	.08	---							
10.	TOTCKAPB	.16	.06	.03	-.03	.77	.21	.54	.10	.09	---						
11.	TOTCKQPB	-.02	-.05	.07	.01	.00	.11	.46	.08	.01	.03	---					
12.	TOTCALT	-.24	-.07	-.03	.13	-.00	-.22	-.19	-.08	-.03	-.07	-.04	---				
13.	LLPERVM	.31	.14	.26	-.17	.05	.55	.39	.26	.16	.13	.15	-.14	---			
14.	ZFRD	-.00	.05	.40	.03	.11	.10	.10	.09	.18	.02	.00	-.06	.12	---		
15.	ZFRS	-.05	-.03	.03	.22	.21	-.15	-.07	-.05	.02	-.07	-.01	-.04	-.10	.24	---	
16.	PR3	-.02	.07	.28	.00	.12	.08	.03	.05	.07	.05	-.05	-.08	.08	.80	.19	---
17.	TOTCKNEI	-.03	.06	.25	.04	.11	.17	.07	.08	.08	.04	.01	.03	.13	.19	.14	.72
18.	TOTCKLEV	.03	.09	.03	-.05	.64	.09	.00	-.02	-.01	-.01	-.01	.01	-.05	.24	.16	.24
19.	ZI	-.05	.02	.12	-.08	.08	-.00	.04	.03	.00	.12	-.01	.05	-.02	.18	-.01	.11
20.	TOTCKSER	.00	.12	.01	.03	.05	-.05	-.01	.03	-.05	-.04	-.00	.01	-.06	.28	.25	.29
21.	TOTCKTH	-.03	.08	.02	.06	.05	-.06	-.02	.02	-.03	-.03	.02	.01	-.01	.32	.27	.31
22.	TOTCKUK	.04	.08	-.01	-.03	.01	.04	-.01	.06	-.07	-.04	-.03	-.01	.01	.12	.12	.19
23.	ZPR7	.03	.03	.02	.02	.00	-.01	-.00	.00	.01	-.05	-.02	.00	.00	.11	.14	.12
24.	ZPR8	-.08	.07	.07	.07	.04	-.04	-.06	-.00	.01	-.05	-.02	.01	-.03	.30	.25	.28
25.	ZPR9	.02	.06	.01	-.04	.08	-.09	-.01	-.02	-.02	.02	-.03	.04	-.05	.20	.16	.19
26.	ZFR10	.03	.07	-.01	-.02	.03	.04	-.01	.07	-.09	-.02	-.03	.02	.00	.10	.06	.18
27.	ZPR11	-.03	.01	-.04	.10	-.02	-.11	-.02	-.06	.07	-.06	.02	.00	-.10	.06	.06	.06
28.	ZPR12	-.00	.05	-.04	.00	.06	-.09	-.03	-.01	-.05	-.01	.00	.00	-.06	.25	.25	.24
29.	TOTCKLRN	-.10	-.03	-.01	.09	-.00	-.20	-.10	-.03	-.07	-.05	-.02	.13	-.20	.09	.03	.07
30.	TOTCKREL	.05	.08	.11	-.06	.04	.02	.01	.00	.10	-.00	-.10	-.06	.03	.47	.15	.44
31.	TOTCKPCD	.04	.07	.12	-.04	-.01	.04	.03	.02	.12	-.03	-.06	-.01	-.00	.39	.12	.35
32.	TOTCKCD	.28	.09	.13	-.07	.02	.11	.19	.13	.13	.01	-.06	-.07	.04	.32	.10	.28
33.	ZLRFS	-.13	-.04	.06	.81	.44	-.26	-.12	-.11	.06	-.07	-.02	.06	-.18	.20	.66	.13
34.	AS	-.12	-.00	.12	.19	.18	.00	-.06	-.00	-.01	.03	-.02	.02	.02	.40	.20	.32
35.	TOTCKRNS	.02	.07	.24	.01	.10	.09	.06	.07	.09	.02	-.02	-.07	.10	.76	.19	.78

FIGURE 1
Correlation Matrix

is low, the sheer number of variables under consideration here accounts for the large value for Cronbach's α.

Computation of Cronbach's α assumes that the 35 variables are parallel measures—that is, different measures of the same underlying phenomenon. It ignores the fact that two or more distinct dimensions of a complex phenomenon may be measured. If such is the case, Cronbach's α is a poor estimator of the true reliability of the measuring instrument. Close inspection of the correlation matrix (figure 1) suggests that this may be the case for the reference transaction assessment instrument. In some sections of the matrix, correlations are quite small, while in other sections, groups of highly intercorrelated variables can be seen. To explore this possibility in a more formal manner, a cluster analysis of the 35 variables was conducted.

Cluster analysis is one of a number of statistical methods used in the examination of multivariate data sets and is akin to factor analysis, principal components analysis, and canonical correlation analysis. Whereas these other techniques aim to create a new set of variables that represent a mathematically simpler combination of the original set of variables, the results of cluster analysis are groupings of the original variables based upon their overall degree of similarity. Variables in the same cluster are more closely related to each other than to any variable found in a different cluster. In addition, these groups, termed *clusters*, are nonoverlapping. A variable will be assigned to only one cluster. In the other previously mentioned multivariate techniques, a single variable may be divided up into two or more new variables, making interpretation of the results rather difficult.

There are a number of distinct techniques for performing cluster analyses.[84] The procedure used in this study is a hierarchical, divisive method using the SAS program VARCLUS.[85] The procedure is divisive in that it starts out with the full 35-

17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34
.14	..:																
.05	.06	---															
.24	.60	.04	---														
.28	.46	.04	.79	---													
.11	.23	.01	.70	.24	---												
.05	.30	.02	.51	.22	.65	---											
.25	.42	-.01	.68	.84	.23	.22	---										
.17	.86	.06	.65	.49	.23	.25	.46	---									
.10	.14	.00	.60	.18	.90	.30	.19	.19	---								
.04	.08	-.04	.37	.21	.17	.21	.14	.07	.10	---							
.19	.34	.02	.60	.50	.15	.18	.49	.35	.09	.17	---						
.06	.10	.05	.19	.23	.06	-.03	.22	.14	.14	.06	.16	---					
.26	.16	.05	.17	.10	.14	.13	.13	.15	.09	.06	.06	-.00	---				
.23	.24	.04	.35	.16	.42	.59	.19	.21	.22	.08	.12	-.04	.72	---			
.17	.20	.04	.28	.11	.35	.52	.12	.18	.17	.06	.11	-.06	.58	.82	---		
.12	.08	-.07	.16	.18	.05	.10	.19	.09	.03	.08	.13	.12	.06	.04	.01	---	
.29	.13	.07	.16	.24	.01	.05	.18	.14	.00	.03	.22	.06	.22	.15	.15	.27	---
.88	.31	.05	.29	.29	.17	.14	.27	.22	.12	.07	.20	.04	.59	.43	.34	.13	.32

FIGURE 1 *Continued*

variable set and begins to divide them into clusters in a step-by-step manner. It first divides the total variable set into a best two-cluster solution. It then examines the two clusters, finds the cluster containing the greater amount of dissimilarity among the variables contained in that cluster, and divides it into two clusters. The result is a three-cluster solution of the variables. The three clusters are then examined, a four-cluster solution is obtained, and so on. Ultimately, if this procedure were allowed to keep clustering in this manner, it would eventually produce 35 clusters, with one variable in each cluster. However, the subdividing process is stopped when there is statistically little improvement to be gained by further subdivision. In the context of reliability studies, variables contained in the same cluster when the program terminates can be considered sufficiently similar so as to represent parallel measures of an underlying dimension. In addition, the VARCLUS program is a form of oblique principal components

analysis so that the reliability estimator theta (Θ), an extension of Cronbach's α to the results of principal components analysis,[86] can be computed for each resulting cluster.

Table 2 lists the 12 clusters produced from the original 35 variables by the procedure described above. The steps in the formation of this 12-cluster solution are represented in figure 2, where gaps in the column of Xs correspond to separate clusters as produced at each step in the analysis.

The results observed in figure 2 are quite revealing. At the two-cluster stage, the analysis almost exactly divides the variables into those derived from the librarian half of the form and those from the patron half. The correlation between the two clusters is 0.04. The only exception is that TOTCKLRN, a patron variable measuring how much the patron learned about the library or reference sources, is contained in the librarian cluster. It should be noted that at the final stage, this variable is clus-

TABLE 2

CLUSTERS PRODUCED FROM ORIGINAL 35 VARIABLES

Cluster	Variable Name	Definition
1	TOTCKSER	Number of times patron checks negative responses on 6 service items (understanding, knowledgeability, courtesy, clear explanations, enough time, enough help)
	TOTCKTH	Number of times patron reports receiving only partly or not enough time and/or help
	ZPR12	Patron reports whether enough time was given (yes, partly, no)
	ZPR8	Patron reports whether enough help was given (yes, partly, no)
2	TOTCKLR3	Number of times librarian checks 17 listed problems (see appendix B)
	TOTCKSPB	Number of times librarian checks 3 searching problems (difficult to think of source, difficult to find subject headings, source is difficult to consult)
	TOTCKQPB	Number of times librarian checks 4 question problems (lacking and misinformation, very recent information wanted, concerned with foreign country/language, difficult citation)
3	ZFND	Patron reports degree to which he/she found what was wanted and was satisfied (found exactly and satisfied, found approximately and satisfied, found approximately and not satisfied, not found and not satisfied)
	PR3	Patron reports degree of satisfaction with what was found or not found (satisfied, partly satisfied, not satisfied)
	TOTCKNEI	Number of times patron checks negative items in regard to *quantity* of information (found nothing, couldn't find information in source, not enough, not enough in-depth)
	AG	Number of times patron and librarian agree on outcome (agree found, agree partly found, agree not found)
	TOTCKRNS	Number of times patron checks 9 listed reasons for being partly or not satisfied (found nothing, not enough, need more in-depth, not relevant enough, want different viewpoint, couldn't find information in source, not sure if correct)
4	TOTCKUK	Number of times that patron reports that librarian only partly or did not understand, and/or was only partly or not knowledgeable
	ZPR7	Patron reports whether librarian understood what was wanted (yes, partly, no)
	ZPR10	Patron reports whether librarian appeared knowledgeable (yes, partly, no)
	ZPR11	Patron reports whether librarian was courteous (yes, partly, no)
5	ZBD	Librarian reports whether busy and whether directed or helped with search (busy and direct, busy and search, not busy and direct, not busy and search)
	TOTCKAPB	Number of times librarian reports 3 administrative problems (busy, books off shelf, cataloging or technical problem)
6	TOTCKREL	Number of times patron checks negative items in regard to *quality* of information (not relevant enough, want different viewpoint, not sure information correct, what I asked for but not really what I wanted)
	TOTCKPCD	Number of times checks items indicating communication difficulty (what I asked for but not really what I wanted, librarian only partly or did not understand what was wanted, want different viewpoint)
	TOTCKCD	Number of times patron checks items indicating communication difficulty *plus* number of times librarian reports that patron was difficult to communicate with
7	ZPRS	Patron reports that librarian helped with the search or directed and suggested
	ZLRPS	Librarian *and* patron report that librarian helped with search Librarian *or* patron report that librarian directed and suggested only

TABLE 2 Continued

Cluster	Variable Name	Definition
8	TOTCKPPB	Number of times librarian reports 5 patron problems (needs extra help, in hurry, wants number of things, returns frequently, difficult to communicate with)
	QT	Librarian reports type of question (4 types, most specific to least specific)
	ZT	Librarian reports time taken in minutes (0–2, 3–5, 5–15, more than 15)
	LLR5NUM	Librarian reports number of sources used (1–5+)
	TOTCKLI	Number of times librarian reports giving 2 types of library instruction (sources, citations, search strategy *and/or* catalog, computer, holdings, locations)
	TOTCKLRN	Number of times patron reports learning (degree of learning about sources *and/or* about library in general)
9	ZRT	Librarian reports degree of success in finding what was wanted by patron (found, partly found, not found)
	TOTCKCPB	Number of times librarian reports collection problems (collection weak or out-of-date *and/or* needs books in another area)
10	TOTCKLEV	Number of times patron checks items indicating difficulty in comprehending and utilizing information (need more simple, too much, explanations not clear)
	ZPR9	Patron reports whether librarian's explanations were clear (yes, partly, no)
11	LRS	Librarian reports whether helped with the search or directed and suggested only
12	ZI	Patron reports degree of importance of transaction to him/her (very important, important, moderately important, somewhat important, not important)

tered with TOTCKLI, a measure of the amount of instruction the librarian gave the patron. The views of both the librarian and patron appear to agree on the issue of instruction.

At the third step in the derivation, the patron variables are divided into a cluster of variables evaluating the services provided by the librarian and a cluster of variables measuring the quality of the information the patron received. The correlation between these two clusters is 0.39.

The other intermediate steps in the derivation of the 12-cluster solution are not so easily interpretable. In general, however, the variables within each of the 12 clusters are semantically related, suggesting that the results are sufficiently valid for reliability considerations.

Given that the 12 clusters produced from our 35 variables (see table 2) represent distinct submeasurements of the quality of library reference departments, reliability estimates for each cluster can be computed. These are listed in table 3. It

should be noted that theta for a single-variable cluster cannot be computed.

Clearly, cluster nine has an extremely low reliability estimate vis-à-vis the other clusters. It combines the presence of collection problems in the area of the patron's question with the librarian's report of whether the information needed was found. Although this is a sensible relationship of the two variables in that a weak collection in a given area will adversely affect the ability of the librarian to find the information the patron requests, the relationship read the other way around (a strong collection in a given area) is not so directly related to finding the information. A host of factors intervene in going from the patron's question to determining the correct source. It is significant that cluster nine was produced from a larger cluster that contained cluster two, a set of variables consisting of overall problems with the transaction, the presence of searching problems, and the presence of question problems. If clusters two and nine are recombined, the resulting cluster has a reli-

```
                    ITERATION                      VARIABLES

     XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX     TOTCKSER
     XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX     TOTCKTH
     XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX     ZPR8
     XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX     ZPR12
     XXXXXXXXXXXXXXXXXXXXXXXXXXXX
     XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX     TOTCKLEV
     XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX     ZPR9
     XXXXXXXXXXXXXXXXXXXXXX
     XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX     ZPRS
     XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX     ZLRPS
     XXXXXXXXXX
     XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX     TOTCKUK
     XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX     ZPR7
     XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX     ZPR10
     XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX     ZPR11
     XXXXX
     XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX     ZFND
     XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX     PR3
     XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX     TOTCKNEI
     XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX     AG
     XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX     TOTCKRNS
     XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
     XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX     ZI
     XXXXXXXXXXXXXX
     XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX     TOTCKREL
     XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX     TOTCKPCD
     XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX     TOTCKCD
     XX
     XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX     TOTCKPPB
     XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX     QT
     XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX     ZT
     XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX     TOTCKLI
     XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX     LLR5NUM
     XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX     TOTCKLRN
     XXXXXXXXXXXXXXXXXXXXXX
     XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX     TOTCKLR3
     XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX     TOTCKSPB
     XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX     TOTCKQPB
     XXXXXXXXXXXXXXXXXXXXXXXXXXXX
     XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX     ZRT
     XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX     TOTCKCPB
     XXXXXXXXXXXXXX
     XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX     ZLRS
     XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
     XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX     ZBD
     XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX     TOTCKAPB
```

FIGURE 2
Cluster Solution

TABLE 3
RELIABILITY ESTIMATES FOR CLUSTERS

Cluster	THETA	Cluster	THETA
1	0.90	7	0.79
2	0.62	8	0.62
3	0.89	9	0.46
4	0.76	10	0.93
5	0.87	11	—
6	0.88	12	—

ability estimate (theta) of 0.63.

Finally, the reliability estimate theta can be computed for intermediary clusters. In particular, at the two-cluster stage, the reliability of the patron-derived variable cluster is 0.89 and for the librarian-derived cluster, 0.76.

External Validity

In addition to being reliable, measures must be valid. In contrast to reliability, validity of measures concerns the degree to which an instrument measures the concept it is designed to measure. Evidence for the validity of an instrument can come from a number of sources, but external validity, the degree to which results from a measurement instrument correspond to the results from other studies of the same phenomenon, is most commonly utilized in addition to the construct validity issues previously discussed.

Many of the variables assessed by the reference transaction assessment instrument have not been independently examined on a sample of library transactions and so cannot be assessed in terms of external validity. One item, however—the percentage of times the patron found exactly what was wanted—should be comparable to the results of unobtrusive observation by Myers and by Jirjees. In these studies, proxies presented test questions by telephone to reference librarians in academic libraries in the Southeast and Northeast. Myers, in her study of academic reference service in the U. S. Southeast, found that 49.12% of the responses to her list of 14 factual questions were answered correctly.[87] Jirjees, using 35 factual questions, found that 56.6% were answered correctly.[88] In our sample of 15 libraries, patrons report that they received "just what was wanted" on 46.32% of the "facts and statistics" questions. A chi-square test comparing these percentages in the three studies is not statistically significant (table 4). Thus, results of success on factual-type questions as reported by patrons in this study correspond to that obtained by testing reference librarians by unobtrusive observation on a preselected set of reference questions.

A second comparison with external studies involves previously reported data on patron satisfaction with reference department service, as opposed to quality of answer, where approximately 81% and 84% of reference department patrons reported they were satisfied with the "interpersonal" service they received.[89] In this study, where patrons were asked to evaluate the quality of the "interpersonal" service they received, independently of their evaluation of their satisfaction with the information they obtained, 79.48% of the

TABLE 4
COMPARISON OF RESULTS OF UNOBTRUSIVE OBSERVATION WITH RESULTS OF THIS STUDY

	Myers	JIRJEES	RTAI	Total
Found Exactly	236	99	44	379
Not Found Exactly	244	76	51	371
Total	480	175	95	750
	$x^2 = 3.64$		$df = 2$	$P > 0.05$

patrons in the sample reported they were completely satisfied with the service. Although no statistical test of these percentages can be provided, the results of these three studies of patron satisfaction with service are quite similar.

DISCUSSION

For the two variables on the reference transaction assessment instrument for which independent measures exist (i.e., ZFND and TOTCKSER), the results from our pilot study agree with those of other studies. Assessment of external validity for the other variables derived from the RTAI must await independent study. For the complete set of 35 variables, estimated reliability as measures of reference department quality is 0.81, which is quite respectable for a questionnaire. Given the high return rate, the RIAI and its employment as an evaluation tool appear to be a viable means of conducting reference department evaluation.

Results from the cluster analysis indicate that the 35 variables derived from the RTAI group into 12 clusters, with reliability estimates ranging from 0.46 to 0.93. Further interpretations and inferences from the clustering results must be cautiously drawn as the clustering technique used in this analysis is designed to be exploratory and, as such, does not serve to confirm or refute a priori hypotheses concerning the relationships among variables. Alternate techniques would be needed to test different models for the relationships among clusters, and a more rigorous sampling strategy would be required for testing hypotheses and generalizing results to the total population of reference transactions.

Two aspects of the clustering solution presented in the reliability section, which were only mentioned in passing, require further explanation and exploration. As was mentioned, at the two-cluster stage the procedure employed in this study divided the 35 variable set into a patron form-derived cluster and a librarian form-derived cluster, with a correlation of 0.04 between these two clusters. This low correlation may surprise the reader, suggesting that there is no statistically significant

relationship between patron and librarian judgments. This may not be a valid inference, however, because the librarian form consists of variables that, by and large, measure factors in the situation that are expected to influence librarian performance. The patron form variables largely measure outcomes of the interaction. Although factors that influence outcomes should relate to those same outcomes more strongly than this result indicates, the clustering procedure used here was selected to create clusters that are maximally distinct from one another while at the same time, placing variables that are as similar as possible into a cluster. The most likely interpretation of our results is that outcome variables, as a group, are more closely related to each other than they are to influencing variables as a group. Alternate analyses would need to be performed to determine the strength of relationship between particular situational factors and particular outcome measures.

At the 3-cluster stage of the 12-cluster final solution, the procedure separates the patron cluster into a service evaluation subcluster and a quality of information received subcluster, with a correlation of 0.39 between these two subclusters. This result may have some bearing on the controversy surrounding the ability of patrons to evaluate reference department quality. Our results indicate that, as judged by the patron, reference librarians satisfactorily respond to the patron's information needs on factual questions approximately 50% of the time. Patrons, however, also appear to distinguish between their satisfaction with the services provided (courtesy, amount of help and time given, understanding and knowledge, clear explanations) and their satisfaction with the information they obtained. With a correlation of 0.39 between the service measures and the information measures, only 16% of the variance in service satisfaction could be explained by their evaluation of the quality of information received. Although quality of information received is a factor influencing the patron's judgment of the quality of service, this is not the sole determinant. Inferences concerning the patron's ability to

judge information critically, based only upon measures of their satisfaction with the services provided, are unwarranted.

These interpretations, nevertheless, require further research. Our results are merely suggestive. This, however, indicates an additional use of the reference transaction assessment instrument as a research tool. By presenting the results of our pilot test and its use as an evaluation tool, it is our hope that other researchers in the area of reference department behavior will be encouraged to further study the issue of what makes a reference department good.

REFERENCES AND NOTES

1. Mary Jo Lynch, "Research in Library Reference/Information Service," *Library Trends* 12:462 (Winter 1983).
2. Samuel Rothstein, "The Measurement and Evaluation of Reference Service," *Library Trends* 12:464 (Winter 1983).
3. Alvin M. Schrader, "Performance Standards for Accuracy in Reference and Information Services: The Impact of Unobtrusive Measurement Methodology," *The Reference Librarian* 11:197–214 (Fall/Winter 1984).
4. Katherine Emerson, "National Reporting on Reference Transactions, 1976–78," *RQ* 16:199–207 (Spring 1977); *American National Standard for Library and Information Sciences and Related Publishing Practices: Library Statistics* (New York: American National Standards Institute, 1983) (ANSI Z239.7-1983).
5. Carolyn Hieber, *An Analysis of Questions and Answers in Libraries,* Studies in the Man-System Interface in Libraries, Report no.1 (Bethlehem, Pa.: Lehigh University Center for Information Sciences, 1966) (ED 014 995).
6. Gerald Jahoda, *The Process of Answering Reference Questions: A Test of a Descriptive Model* (Tallahassee, Fla.: Florida State University, School of Library Science, 1977) (ED 136 769); Gerald Jahoda and Judith Braunagel, *The Librarian and Reference Queries: A Systematic Approach* (New York: Academic Pr., 1980).
7. Clark Hallman, "Designing Optical Mark Forms for Reference Statistics," *RQ* 20:257–64 (Spring 1981).
8. Gary Strong, "Evaluating the Reference Product," *RQ* 19:367–72 (Summer 1980).
9. Benita J. Howell, Edward B. Reeves, and John Van Willigen, "Fleeting Encounters—A Role Analysis for Reference Librarian-Patron Interaction," *RQ* 16:124–29 (Winter 1976).
10. Judith A. Tessier, Wayne W. Crouch, and Pauline Atherton (Cochrane), "New Measures of User Satisfaction with Computer-Based Literature Searches," *Special Libraries* 68:383–89 (Nov. 1977).
11. Renata Tagliacozzo, "Estimating the Satisfaction of Information Users," *Bulletin of the Medical Library Association* 65:243–49 (Apr. 1977).
12. Helen Gothberg, "Immediacy: A Study of Communication Effect on the Reference Process," *Journal of Academic Librarianship* 2:128 (July 1986).
13. Carlos Cuadra, Robert Katter, Emory Holmes, and Everett Wallace, *Experimental Studies of Relevance Judgments: Final Report. Volume II. Description of Individual Studies* (Santa Monica, Calif.: System Development Corporation, 1967).
14. Marjorie Murfin and Lubomyr Wynar, *Reference Service: An Annotated Bibliographic Guide; Supplement 1976-1982* (Littleton, Colo.: Libraries Unlimited, 1984), p.179–215, 9–10, 104–14, 245–47.
15. William F. Young, "Evaluating the Reference Librarian," *The Reference Librarian* 11:125–26 (Winter 1984).
16. Murfin and Wynar, *Supplement 1976–1982,* p.193–97, 201–2.
17. Murfin and Wynar, *Supplement 1976–1982,* p.208–15.
18. Murfin and Wynar, *Supplement 1976–1982,* p.285–97.
19. Ellen Altman, "Assessment of Reference Services," in *The Service Imperative for Libraries: Essays in Honor of Margaret E. Monroe,* ed. Gail Schlachter (Littleton, Colo.: Libraries Unlimited, 1982), p.170.
20. Mary W. George, "Questions Sheehy Can't Answer: Reflections on Becoming Head of Referece," *The Reference Librarian* 3:13 (Spring 1982).
21. Ellsworth Mason and Joan Mason, "The Whole Shebang—Comprehensive Evaluation of Reference Operations," *The Reference Librarian* 11:26 (Fall/Winter 1984).
22. Douglas Zweizig, "Tailoring Measures to Fit Your Service: A Guide for the Managers of Reference

Services," *The Reference Librarian* 11:53 (Fall/Winter 1984).
23. Murfin and Wynar, *Supplement 1976–1982*, p.212–15.
24. Linda M. Olson, "Reference Service Evaluation in Medium-Sized Academic Libraries: A Model," *Journal of Academic Librarianship* 9:322–29 (Jan. 1984).
25. Charles McClure, "Output Measures, Unotrusive Testing, and Assessing the Quality of Reference Services," *The Reference Librarian* 11:232 (Fall/Winter 1984).
26. McClure, "Output Measures," p.229; Sydney Pierce, "In Pursuit of the Possible: Evaluating Reference Services," *The Reference Librarian* 11:16 (Fall/Winter 1984).
27. Terence Crowley, "Referred Reference Questions: How Well Are They Answered?" *The Reference Librarian* 11:90–91 (Fall/Winter 1984).
28. Pierce, "In Pursuit of the Possible," p.18.
29. Michael Roloff, "Communication at the User-System Interface: A Review of Research," *Library Research* 1:16 (Spring 1979).
30. Young, "Evaluating the Reference Librarian," p.124.
31. Rothstein, "The Measurement and Evaluation of Reference Service," p.464; Zweizig, "Tailoring Measures," p.56; Pierce, "In Pursuit of the Possible," p.15–16; Altman, "Assessment of Reference Services," p.175–76; Paul Kantor, "Evaluation of and Feedback in Information Storage and Retrieval Systems," in *Annual Review of Information Science and Technology*, v.17 (White Plains, N.Y.: Knowledge Industry, 1982), p.114; Alma C. Vathis, "Reference Transaction and End Product as Viewed by the Patron," *RQ* 23:61 (Fall 1983); Terry Weech, "Who's Giving All Those Wrong Answers? Direct Service and Reference Personnel Evaluation," *The Reference Librarian* 11:119–20 (Fall/Winter 1984); McClure, "Output Measures," p.231.
32. Rothstein, "The Measurement and Evaluation of Reference Service," p.456–72; Terry Weech, "Evaluation of Adult Reference Service," *Library Trends* 22:315–35 (Jan. 1974).
33. Gothberg, "Immediacy: A Study of Communication Effect on the Reference Process," p.128.
34. Tagliacozzo, "Estimating the Satisfaction of Information Users," p.248.
35. Tagliacozzo, "Estimating the Satisfaction of Information Users," p.243.
36. L. J. Cronbach, "Course Improvement through Evaluation," *Teacher's College Record* 64:680 (May 1983).
37. Ethel Auster, "User Satisfaction with the Online Negotiation Interview: Contemporary Concern in Traditional Perspective," *RQ* 23:55 (Fall 1983).
38. Tessier, Crouch, and Atherton (Cochrane), "New Measures of User Satisfaction," p.383–89.
39. William Mehrens and Irvin Lehmann, *Standardized Tests in Education*, 3d ed. (New York: Holt, 1980), p.215.
40. L. J. Cronbach, "Further Evidence on Response Sets and Test Design," *Educational and Psychological Measurement* 10, no.1:4–5, 22–23 (1950).
41. Cuadra and others, *Experimental Studies*, p.75.
42. Elizabeth Loftus, Stephen Fienberg, and Judith Tanur, "Cognitive Psychology Meets the National Survey," *American Psychologist* 40:178 (Feb. 1985).
43. Carlos Cuadra and Robert V. Katter, "Opening the Black Box of Relevance," *Journal of Documentation* 23:291, 293 (Dec. 1967).
44. Cuadra and others, *Experimental Studies*, p.66.
45. Mehrens and Lehman, *Standardized Tests*, p.61, 53–54.
46. David Shavit, "Qualitative Evaluation of Reference Service," *The Reference Librarian* 11:239–40 (Fall/Winter 1984).
47. Blaise Cronin, "Performance Measurement and Information Management," *Aslib Proceedings* 34:231 (May 1982). See also Shavit, "Qualitative Evaluation," p.241; Douglas Zweizig and Brenda Dervin, "Public Library Use, Users, Uses—Advances in Knowledge of the Characteristics and Needs of the Adult Clientele of American Public Libraries," in *Advances in Librarianship* v.7, ed. Melvin Vogt (New York: Academic, 1977), p.249; P. V. Rzasa and N. R. Baker, "Measures of Effectiveness for a University Library," *Journal of the American Society for Information Science* 23:248 (July 1972); Vathis, "Reference Transactions and End Product," p.60; F. W. Lancaster, *The Measurement and Evaluation of Library Services* (Washington, D.C.: Information Resources, 1977), p.1; Helen Focke, "Library Users of Information and Their Needs," in *Present Status and Future Prospects of Reference/Information Service* (Chicago: American Library Assn., 1967), p.32.
48. William S. Cooper, "On Selecting a Measure of Retrieval Effectiveness," *Journal of the American Society for Information Science* 24:87–100 (Mar./Apr. 1973).
49. Dagobert Soergel, "Is User Satisfaction a Hobgoblin?" *Journal of the American Society for Information Science* 27:256–59 (July/Aug. 1976).
50. Cuadra and Katter, "Opening the Black Box of Relevance," p.294.

51. Tagliacozzo, "Estimating the Satisfaction of Information Users," p.243.
52. For arguments in support of this viewpoint, see Cooper, "On Selecting a Measure," p.87–99. Also see Tessier, Crouch, and Atherton (Cochrane), "New Measures of User Satisfaction," p.386.
53. A. Pritchard, M. Auckland, and M. Castens, *Library Effectiveness Study* (London: City of London Polytechnic, 1973) (ED 089 761).
54. Rzasa and Baker, "Measures of Effectiveness," p.248.
55. Lancaster, *The Measurement and Evaluation of Library Service*, p.1.
56. Gary E. Strong, "The Consortium for Public Library Innovation," *Public Library Quarterly* 1:253 (1979).
57. Strong, "Evaluating the Reference Product," p.367–72.
58. Rzasa and Baker, "Measures of Effectiveness," p.248.
59. Robert Pittenger, Charles Hockett, and John Danehy, *The First Five Minutes: A Sample of Microscopic Interview Analysis* (Ithaca, N.Y.: Martineau, 1960).
60. Robert S. Taylor, "Question-Negotiation and Information Seeking in Libraries," *College & Research Libraries* 29:180 (May 1968).
61. F. W. Lancaster, "Factors Influencing the Effectiveness of Question-Answering Services in Libraries," *The Reference Librarian* 11:95–108 (Fall/Winter 1984).
62. S. D. Neill, "The Reference Process and the Philosophy of Karl Popper," *RQ* 24:314–18 (Spring 1985).
63. Walter Dick and Nancy Hagerty, *Topics in Measurement: Reliability and Validity* (New York: McGraw-Hill, 1971), p.85.
64. Mehrens and Lehmann, *Standardized Tests*, p.267.
65. Shavit, "Qualitative Evaluation," p.237, 241–42.
66. Lancaster, "Factors Influencing the Effectiveness of Question-Answering," p.106-7; McClure, "Output Measures," p.218; Pierce, "In Pursuit of the Possible," p.19–20.
67. Mehrens and Lehmann, *Standardized Tests*, p.266–67.
68. Jahoda, *Process of Answering Reference Questions*.
69. Matthias Burisch, "Approaches to Personality Inventory Construction: A Comparison of Merits," *American Psychologist* 39:214 (Mar. 1984).
70. Ralph Gers and Lillie Seward, "Improving Reference Performance," *Library Journal* 110:34 (Nov. 1, 1985).
71. Roloff, "Communication at the User-System Interface: A Review of Research," p.16.
72. *American National Standard for Library and Information Sciences and Related Publishing Practices: Library Statistics* (New York: American National Standards Institute, 1983) (ANSI Z39.7-1983).
73. Maryann Kevin Brown, "Library Data, Statistics and Information: Progress toward Comparability," *Special Libraries* 71:477, 484, 478 (Nov. 1980). See also Marcella Ciucki, "Recording of Reference/Information Service Activities: A Study of Forms Currently Used," *RQ* 16:273–83 (Summer 1977); Auster, "User Satisfaction," p.56; Katherine Emerson, "Definitions for Planning and Evaluating Reference Service," *The Reference Librarian* 11:74 (Fall/Winter 1984); Pierce, "In Pursuit of the Possible," p.14.
74. Mehrens and Lehmann, *Standardized Tests*, p.27, 13, 86, 57.
75. Zweizig, "Tailoring Measures," p.58–59.
76. Shavit, "Qualitative Evaluation," p.235.
77. Brown, "Library Data," p.478.
78. Mehrens and Lehmann, *Standardized Tests*, p.12–15, 27.
79. Ibid., p.27.
80. Ibid., p.106.
81. The 50-form limit was decided upon for two major reasons. First, the questionnaire form of research is somewhat intrusive. Therefore, restricting the intrusion to as small a degree as possible seemed a desirable goal. Second, a sample of fewer than 30 transactions per library would severely threaten the security of making statistical inferences for a given library. Thus, 30 forms would be the minimally acceptable number of transactions. Not knowing the degree of compliance of both librarians and patrons to filling out and depositing the forms, some number above 30 would be necessary to insure that a sufficient number of complete transactions would be obtained. A 60% return rate was assumed, thereby yielding the number 50.
82. For a discussion of Cronbach's alpha, see E. G. Carmines and R. A. Zeller, *Reliability and Validity Assessment* (Beverly Hills, Calif.: Sage Publ., 1979), p.43–46.
83. The value of Cronbach's alpha reported here differs slightly from that published by C. Bunge. Bunge provides preliminary results based on a simplified correlation matrix where all statistically nonsignificant correlations were set equal to zero. The results here are based on the empirically

derived correlation matrix given in figure 1. See C. Bunge, "Factors Related to Reference Question Answering Success," *RQ* 24:483 (Summer 1985).

84. For an overview of various clustering techniques, see M. S. Aldenderfer and R. K. Blashfield, *Cluster Analysis* (Beverly Hills, Calif.: Sage Publ., 1984).

85. SAS Institute, *SAS User's Guide: Statistics*, 1982 ed. (Cary, N.C.: SAS Institute, 1982), p.461–62. An *N* of 448 forms representing all forms that were completely filled out for all variables was used in this analysis.

86. Carmines and Zeller, *Reliability and Validity*, p.60–61.

87. Marcia J. Myers and Jassim M. Jirgees, *The Accuracy of Telephone Reference/Information Services in Academic Libraries* (Metuchen, N.J.: Scarecrow, 1983), p.60.

88. Myers and Jirgess, p.240.

89. Strong, "Evaluating the Reference Product," p.370; Howell and others, "Fleeting Encounters," p.127.

APPENDIX A: REFERENCE TRANSACTION ASSESSMENT INSTRUMENT

APPENDIX A. *Continued*

APPENDIX B: VARIABLES USED IN RELIABILITY ASSESSMENT

1. TOTCKPPB—Total Number of Patron Problems Checked by Librarian
   Range = 0 to 5
   a. Patron wants a number of things
   b. Patron in hurry
   c. Confused question or other communication difficulty
   d. Needs extra help
   e. Returns frequently
2. QT—Question Type by Level of Specificity as reported by librarian and scored as follows:
   1. Wants specific smaller item in larger publication or works by known author (known item books or serial titles are excluded from this category)
   2. Wants specific fact
   3. Wants explanation of source, etc.
   4. Wants something on a subject
3. ZRT—Results as Reported by Librarian scored as follows:
   1. Found

APPENDIX B. *Continued*

  2. Partly found
  3. Not found
4. ZLRS—Librarian Reports Response Type scored as follows:
  1. Direct and suggest only
  2. Help or make search
  3. Defer or refer patron
5. ZBD—Librarian reports whether busy and whether directed or searched
  1. Busy and direct
  2. Busy and search
  3. Not busy and direct
  4. Not busy and search
6. ZT—Librarian reports time spent
  1. 0–2 minutes
  2. 3–5 minutes
  3. 5–15 minutes
  4. More than 15 minutes
7. TOTCKLR3—Total number of times librarian checks a problem
  Range = 0–17
  a. Lacking and misinformation
  b. Concerned with foreign country or language
  c. Concerned with government document
  d. Very recent information/citation needed
  e. Patron wants number of things
  f. Difficult citation
  g. Patron in hurry
  h. Communication difficulty or confused question
  i. Needs extra help or returns frequently
  j. Difficult to think of source
  k. Difficult to find subject heading
  l. Books off shelf
  m. Source difficult to consult
  n. Busy or very busy
  o. Cataloging or technical problem
  p. Collection weak in that area or out-of-date
  q. Need books in another area or location
8. TOTCKSPB—Total number of times librarian checks searching problems
  Range = 0–3
  a. Difficult to think of source
  b. Difficult to find subject heading
  c. Source difficult to consult
9. TOTCKCPB—Total number of times librarian checks collection problems
  Range = 0–2
  a. Collection weak in that area or out-of-date
  b. Need books in another area or location
10. TOTCKAPB—Total number of times librarian checks administrative problems
  Range = 0–3
  a. Books off shelf
  b. Cataloging or technical problems
  c. Busy or very busy
11. TOTCKQPB—Total number of times librarian checks question problems
  Range = 0–4
  a. Concerned with foreign country/language
  b. Very recent information/citation needed
  c. Difficult citation
  d. Lacking and misinformation
12. TOTCKLI—Total number of times librarian checks giving library instruction
  Range = 0–2
  a. Librarian explains sources, citations, search strategy
  b. Librarian explains catalog, computer, holdings, locations, etc.

APPENDIX B. *Continued*

13. LLR5NUM—Librarian reports number of sources used or suggested
    1. One source
    2. Two sources
    3. Three sources
    4. Four sources
    5. Five or more sources
14. ZFND—Patron reports whether the desired information was found and whether he or she was satisfied with information
    1. Found exactly and satisfied
    2. Found approximately and satisfied
    3. Found approximately and not satisfied
    4. Not found and not satisfied
    (Not found but satisfied—small number only—excluded from this variable)
15. ZPRS—Patron reports whether librarian directed and suggested or helped search
    1. Librarian helped search for information
    2. Patron followed librarian's suggestion and found information
16. PR3—Patron reports degree of satisfaction
    1. Satisfied
    2. Partly satisfied
    3. Not satisfied
17. PR3—Patron reports reasons for being partly or not satisfied with *quantity* of information
    Range = 0–4
    a. Not enough
    b. Need more in-depth
    c. Couldn't find information in source
    d. Found nothing
18. TOTCKLEV—Total number of times patron reports reasons for being partly or not satisfied with appropriateness or clarity of information
    Range = 0–3
    a. Need more simple
    b. Too much
    c. Explanations not clear
19. ZI—Patron reports degree of importance of the question to him or her
    1. Very important
    2. Important
    3. Moderately important
    4. Somewhat important
    5. Not important
20. TOTCKSER—Total number of times patron checks negative service marks
    Range = 0–6
    a. Librarian only partly or did not understand what was wanted
    b. Librarian only partly or did not give enough help and explanation
    c. Explanations were only partly or not clear
    d. Librarian appeared only partly or not knowledgeable
    e. Librarian only partly or not courteous and considerate
    f. Librarian gave patron only partly or not enough time
21. TOTCKTH—Total number of times patron checks not enough assistance
    Range = 0–2
    a. Librarian gave only partly or not enough help and explanation
    b. Librarian gave patron only partly or not enough time
22. TOTCKUK—Total number of patron checks librarian only partly or did not understand or appeared only partly or not knowledgeable.
    Range = 0–2
    a. Librarian only partly or did not understand what was wanted
    b. Librarian appeared only partly or not knowledgeable
23. ZPR7—Librarian understood what was wanted
    0. No
    1. Partly
    2. Yes
24. ZPR8—Librarian gave enough help and explanation
    0. No
    1. Partly
    2. Yes
25. ZPR9—Librarian's explanation were clear
    0. No

APPENDIX B. *Continued*

1. Partly
2. Yes
26. ZPR10—Librarian appeared knowledgeable
    0. No
    1. Partly
    2. Yes
27. ZPR11—Librarian was courteous and considerate
    0. No
    1. Partly
    2. Yes
28. ZPR12—Librarian gave enough time
    0. No
    1. Partly
    2. Yes
29. TOTCKCD—Range = 0–4
    Total number of times patron *or* librarian marks responses indicating communication difficulty.
    Patron reports:
    a. Librarian only partly or did not understand what was wanted
    b. Information was what I asked for, but not really what I wanted
    c. Want different viewpoint
    Librarian reports:
    a. Communication difficulty or confused question
30. TOTCKLRN—Total number of times patron checks that something was learned about library and/or about reference sources
    Range = 0–4

Patron reports learning something about the library	*Patron reports learning new reference sources*
0. No	0. No new sources
1. Partly	1. One new source
2. Yes	2. Two or more new sources

31. TOTCKREL—Total number of times patron checks that information/materials were not relevant or doubtful in some way
    Range = 0–4
    a. Not relevant enough
    b. Want different viewpoint
    c. Not sure if information is correct
    d. Information was what I asked for, but not really what I wanted
32. TOTCKPCD—Total number of times patron checks responses indicating communication difficulties
    Range = 0–3
    a. Librarian only partly or did not understand what was wanted
    b. Information was what I asked for, but not really what I wanted
    c. Want different viewpoint
33. ZLRS—Librarian and patron report whether the librarian suggested and directed only or helped search
    1. Librarian *or* patron report directed and suggested only
    2. Librarian *and* patron report "helped search"
34. AG—Number of times patron and librarian agree on outcome
    a. Librarian and patron agree found
    b. Librarian and patron agree partly found
    c. Librarian and patron agree not found
35. TOTCKRNS—Total number of times patron checks reasons not satisfied
    Range = 0–9

a. Found nothing	f. Not relevant enough
b. Not enough	g. Want different viewpoint
c. Need more simple	h. Couldn't find information in source
d. Too much	i. Not sure if information given me is correct
e. Need more in-depth	