Preprint – cite as:
Pomerantz, J., Mon, L., & McClure, C. R. (2008). Evaluating Remote Reference Service: A Practical Guide to Problems and Solutions. *portal: Libraries and the Academy*, 8(1), 15-30.

Evaluating Remote Reference Service:
A Practical Guide to Problems and Solutions

Jeffrey Pomerantz
School of Information and Library Science
University of North Carolina at Chapel Hill
CB 3360, 100 Manning Hall
Chapel Hill, NC 27599-3360
pomerantz@unc.edu
t: 919-962-8366
f: 919-962-8071

Lorri Mon
College of Information
Florida State University
268 Louis Shores Building
Tallahassee, FL 32306-2100
lmon@ci.fsu.edu
t: 850-644-5772
f: 850-644-9763

Charles R. McClure
College of Information, Information Institute
Florida State University
226 Louis Shores Building
Tallahassee, FL 32306-2100
cmcclure@lis.fsu.edu
t: 850-644-8109
f: 850-644-4522

Abstract

This paper identifies key methodological issues affecting quality of data in the evaluation of remote reference services. Despite a growing number of studies in this area, no comprehensive effort has been made to identify potential problems and suggest solutions. The strategies proposed in this paper offer practical ways in which libraries can improve the overall quality and usefulness of data gathered in remote reference evaluation studies.

Introduction

The current climate of cost-cutting and tight budgets requires libraries to present evidence justifying the value of the services they offer, as well as for planning and decision-making about the future. It is critical for libraries to have quality data on which to base decisions, and which may be reliably used to justify budgets. Indeed, the rise of the evidence-based librarianship movement demonstrates the importance of quality data: evidence-based librarianship is based on the premise that the practice of librarianship must be based upon the highest-quality data[1]. While the authors would argue that librarianship has always relied on evidence, this evidence has sometimes been based more on untested hypotheses and librarians' intuitions than on quality data.

For the purposes of this paper, "quality data" are those that are:

- *Reliable*: the measures produce the same results every time they are used to produce data,
- *Valid*: the measures actually measure that which they are intended to measure, and
- *Useful*: the data assists library decision makers make better decisions than if the data were not available.

In the evaluation of remote reference services, the evaluator should strive to obtain the highest quality data possible.

Any library service needs ongoing evaluation (formative) and evaluation conducted at specific key points in time, e.g., annually, semesterly, quarterly (summative). Such evaluation is essential if decision making is to be done that will improve the service. Remote reference service is no exception to this rule. Indeed, because of the complexity of delivering high quality remote reference service and the need for planning to provide high quality services, evaluation is perhaps even more critical than for other services.

Note that the term "remote reference" is used throughout this paper, instead of the perhaps more common terms *digital reference* or *virtual reference*. Anne Grodzins Lipow suggested that little agreement exists in the library literature as to the use of these terms, and defines digital reference broadly as all forms of "personalized reference service via the Internet," and defines virtual reference more narrowly as only reference services provided via synchronous technologies[2]. This paper addresses reference services provided via both synchronous and asynchronous technologies, so in order to encompass both, the authors use the more inclusive term remote reference.

This paper summarizes a number of issues and topics that can affect the success with which librarians can evaluate remote reference services. Although evaluation of remote reference services is no more complex than the evaluation of other types of library and information services, there are a number of issues and topics specific to remote reference that should be considered prior to conducting such evaluations. As a result of considering these issues and topics, evaluators will be able to strengthen their evaluation design and data collection methods.

Evaluations of remote reference services can be significantly improved if evaluators identify specific strategies to increase the quality of the data that they collect. Both methodological issues of data collection and specific strategies to address these issues to obtain high quality data are outlined in this paper. Evaluations of remote reference services can be much more useful for decision making and planning if evaluators take care in designing and implementing specific aspects of the evaluation as discussed in this paper.

Given the time and cost for performing evaluations of any kind, careful thought must go into the development of the evaluation design and data collection methods. Such “up front” thinking and design development will pay good dividends with more reliable, valid, and useful data. The practical suggestions identified here will result in better evaluations.

Data collected prior to the interaction

All forms of remote reference must offer a means for the user to submit a request to the service: users write to an email address or fill out a webform in email-based services, fill out a webform in chat-based services, or send an instant message to IM-based services. Features and constraints of the communication medium, in addition to the policies of the service, dictate what data are or are not collected.

Note that the terms *request* and *response* are used throughout this paper instead of the more common terms *question* and *answer*, to indicate messages sent by the user to the librarian, and by the librarian to the user. These terms are used because not all requests sent by a user may be in the form of a question, and not all responses sent by a librarian may be an answer. During telephone and in-person interviews conducted with chat and email remote reference users, Lorri Mon found that there was some ambiguity for users about the “answers” received[3]. When asked questions such as whether an answer was fast enough to meet their needs, some users responded, “There wasn’t really an answer,” or “I didn’t really feel that I got an answer,” although the librarians had provided responses to the users’ requests. It is clear that users perceive a distinction between answers and responses, and that distinction is preserved here.

Self reported data

Jeffrey Pomerantz points out that most remote reference services have no mechanism to determine the truthfulness of a user's responses on question submission forms, and that it may be impossible to do so[4]. In some cases, self-reported data is critical: knowing if the user is affiliated with the institution that hosts the library determines whether the librarian is able to direct the user to subscription resources. In other cases, the requested data may be unimportant: the user's name, for example, is generally used simply to personalize the interaction[5], in which case a screen name might work as well as the user's real name.

Since determining the truthfulness of self-reported data is not always possible, reference services should take steps to encourage users to provide reliable data. One method for improving user self-reported data is for the library to explain in each case why specific data are being requested. For example, asking for a zipcode may be perceived as an unnecessary intrusion on privacy, but explaining that this data enables the librarian to suggest locally-available resources allows the user to perceive a benefit in tradeoff: supplying a zipcode means receiving more personalized information. Explanations for each piece of data requested can easily be included on a question submission webform: the Internet Public Library (IPL)'s webform provides an excellent example of this strategy (ipl.org/div/askus/).

What data are worth collecting

Services that use a webform for question submission will receive fairly consistent data from users, since webforms are designed to elicit specific information in specific fields. Email-, chat-, and IM-based services, on the other hand, often collect highly unstructured data from users, since users can type anything at all into messages in these media. It is therefore important for email-, chat-, and IM-based services to provide instructions to users about the type of data that is most important for users to provide in these messages.

Providing users with guidance on how to make sufficiently detailed requests to remote reference services may also help alleviate the problem observed by Lorri Mon, that users reported reference services had provided them with redundant information[3]. Mon found that users made comments such as "all the information I got was information I already had" and "I didn't value the information, and I already had the information." Designing the user interface to obtain more detailed information from users about their questions is one possible way to increase accuracy and reduce redundancy.

The Internet Public Library, for example, uses two questions designed to gain more detailed information about the query: *sources already consulted* and *how information will be used*. For the *sources already consulted* field, the IPL's webform explains, "Knowing where you've already looked will help us keep from sending you someplace you've already been," and for *information use*, the webform states that "Understanding the context and scope of your information needs helps us to deliver an answer that you will find useful." For a service that does not use webforms, similar guidance for users about how to more fully express their requests can be provided on the web page which contains the email or IM link. For example, in the past, the IPL provided a suggested "email form" on which users could model their emailed question submissions.

The webform or email form also offers an opportunity to collect data for broader service assessment. For example, requesting data such as a user's zipcode not only informs the immediate reference interaction by enabling referrals to local resources, but also makes possible subsequent analyses of the geographic scope of a remote reference service. Design of the remote reference intake mechanisms should thus take into consideration what sorts of data are needed from users to inform both the reference interaction and subsequent evaluation efforts. Encouragement and guidance for users to fully cooperate in providing the needed data should also be built in.

Data collected during the interaction

During the interaction, two types of data are collected: data *from* the interaction, and data *about* the interaction. Data from the interaction includes everything communicated between the user and the librarian, including messages that are primarily concerned with the task of the reference interaction as well as those concerned with the relationship between user and the librarian[6]. Data about the interaction includes web server logs and other data collected automatically by the reference management application in use by the service (e.g., Questionpoint, www.questionpoint.org, or AOL Instant Messenger, www.aim.com).

Consistency of automatically-collected data

All computer-based systems collect data automatically: web servers, for example, record data about the usage of the pages on websites; email clients record the date and time of the sending and receipt of emails. Many reference management applications collect additional data such as the length of time the user waited in a "queue" before connecting with a librarian, and the specific library through which a user connected to a consortial service. Web server logs may also be analyzed to collect data about which webpages are requested by users and the date and time of each request, the webpage from which users are referred to a reference service, a user's browser type and IP address, and a range of other data[7].

Data collected automatically are data collected consistently. When a reference management application captures the time that a user connected to the service, for example, that data point has the same meaning across all user sessions. Just because data is collected consistently, however, does not mean that it is important. If a reference service wishes to identify traffic patterns by time of day, users' queue times may be an important piece of data. If a reference service wishes to identify users' satisfaction with the service, however, queue time may be irrelevant. A reference management application may capture a great deal of data, but the reference service is not obligated to use all of it. The use of automatically-captured data should be dictated by the evaluation questions that the reference service wants answered.

Two potential concerns with automatically-collected data focus on the issue of missing data. One problem occurs when data may have been captured but then deleted at some point in the past. Email-based services are particularly prone to this, as it is quite simple to delete a single email message or even entire directories of saved emails. The possibility thus arises that analyses may be working with incomplete data sets. If this situation is detected, it may be possible to get data restored from an archived backup, as network administrators in most institutions archive the contents of servers periodically and store this data for some period of time.

Data may also be missing not because it was deleted, but because it was never collected in the first place. In email-based services, individual librarians may decide to respond to users outside of the reference management application, perhaps to compose their response at a more convenient time, or perhaps because they are more comfortable using their own preferred email software. In instant messaging services, if the instant messenger (IM) options to archive transcripts are not set explicitly, transcripts will not be archived and are lost for future data analysis.

The result in either case is that no record remains as to whether the reference interaction was ever completed. While this may be one way to deal with the issue of patron privacy, it also leaves the library unable to conduct any sort of subsequent analysis or evaluation of services provided. For services that wish to be able to assess remote reference interactions, it is therefore important to set policies about how and where interactions are conducted, to ensure reliability of data collection.

Comparability of automatically-collected data

While data collected automatically may be consistent with itself, it may not always be consistent with other data. Denise Troll Covey reports that libraries "want digital library usage statistics to be comparable with traditional usage statistics."[8] Such usage statistics include data such as "virtual visits" to the library's website and user satisfaction. John Carlo Bertot, Charles R. McClure, and Joe Ryan argue that remote and in-person reference transactions are comparable, and can be combined to arrive at a number for a library's total number of reference transactions[9]. On the other hand, Fornell et al. suggest that customer satisfaction has several components, and different measures will elicit data on different aspects of satisfaction[10]. Evaluators may therefore need to make a judgment call as to the degree to which data collected by different methods or at different times are, in fact, comparable.

There are several efforts currently underway to remedy issues of data consistency in different areas of library data collection: the Counting Online Usage of Networked Electronic Resources project (COUNTER, www.projectcounter.org), the Association of Research Libraries' line of StatsQUAL™ data collection instruments (including LibQUAL+™, DigiQUAL™, and MINES for Libraries™, www.arl.org/stats/initiatives/), ScholarlyStats (www.scholarlystats.com), Donald S. Elliott and colleagues' methodology for cost-benefit analysis[11], and others. While not all of these instruments may be appropriate for all libraries, use of one or more of them may enable a library to collect consistent, quality data which will enable meaningful analyses and be comparable across data collection efforts.

Interpreting automatically collected data

Various types of post-processing may need to take place in order to make automatically-collected data useful for analysis. User IP addresses, for example, may need to be "geocoded" by obtaining a latitude and longitude in order to be useful for understanding patterns of user locations and the geographic scope of the remote reference service.

Interpretation of automatically-collected transcripts can pose challenges when attempting to understand "aboutness" of questions. Indeed, users themselves may have difficulty when asked to classify their questions. David S. Carter and Joseph Janes found that when Internet Public Library users were asked to indicate the subject of their questions using a webform drop-down list with categories such as history, science, government, and literature, the most commonly chosen subject was "Other/Misc.," which was not the default selection[12]. While there is obviously a need to understand what types of questions library users are asking, there is inherent difficulty in attempting to categorize question types.

The choice of a classification scheme for reference requests may differ depending on the purpose and the audience for the research. Many schemes have been used to categorize users' questions, including the general subject of a question (for example, as defined in the Dewey Decimal Classification System or Library of Congress Subject Headings), the overarching information need in terms of what the user wishes to do with the information, and the type of source in which an answer may be found[13]. Classification schemes that are meaningful to one audience, however, may not be appropriate for communicating with another. For example, categories such as "ready reference" or "directional"[14] are meaningful to librarians, but may not be especially meaningful to non-librarians such as users, funders, legislators, or voters. For the purposes of describing the outcomes of library question-answering in terms of impacts on the community, categories such as "homework help" or "e-government" may be more effective in demonstrating the achievements of the service to a wider audience.

Data collected after the interaction

The conclusion of the reference interaction is an obvious and natural point in time to attempt to collect data from the user. The user is still "present" in the service's virtual space, and may still have the service's response in front of them. By capitalizing on the user's virtual presence, the service may collect data from the user at nearly the same point in time as the user receives the final reference interaction response.

Data collection occurs at this point in a variety of ways. An email may be sent automatically to a user who has provided a valid email address; a popup survey window may appear on the user's screen; links to an online survey can be automatically appended to emails and IMs. Surveys are the most immediate method for collecting data following the interaction. Interviews with users provide richer data than surveys can elicit, but require some time to schedule and can be labor-intensive for the evaluator.

Popup windows are commonly used by commercial chat applications, but can be problematic for data collection since many users set their web browsers to block popups. Even if a users' computer settings do not interfere with a popup survey, other inadvertent problems such as a network connection failure or the user exiting the browser application instead of closing the chat session from within the chat software may also cause a failure of the automated chat popup survey. Additionally, even if the popup survey does appear, a user may not notice it among many open windows on the desktop. Popup windows should therefore be avoided for data collection whenever possible. Instead, a link to a web-based survey may be automatically added to the closing exchange of a chat interaction or in a follow-up email, so that a survey link can be more consistently made available to all users.

How long after the interaction?

Immediately following the interaction, users are best able to comment on their initial impressions of the service. The sorts of data that can be collected from users at this point in time include, for example, the speed with which the librarian responded to the user's question, the librarian's helpfulness and politeness, and the ease of use of the software. This data is useful in evaluating the usability of the service, and the quality of the interactions between users and librarians.

Some questions commonly asked of users immediately following the interaction, however, are those for which users cannot provide valid data based upon immediate impressions, such as fully evaluating the accuracy, completeness, and usefulness of the information provided. In order to be able to accurately answer these questions, users may need more time to read, synthesize, and use the information provided[15]. Reference services that collect data from users at different points in time after the interaction therefore should only ask users for data that they can reliably provide at a given point in time.

Jeffrey Pomerantz and Lili Luo made use of a method to collect data about users' evaluations of the completeness and usefulness of the information provided, by following up with users two weeks after the conclusion of the reference interaction[16]. The researchers selected two weeks as an appropriate period of time because they judged this to be long enough for the user to have had time to use the information provided, but still short enough for the user to clearly remember the interaction. What the researchers found, however, is that some users were unable to recall some details of the interaction, and a small number had no recollection of the interaction at all.

Any reference service that wishes to collect follow-up data from users needs to carefully consider what an appropriate period of time is for this follow-up, as there is a clear tradeoff: the more time elapses, the less able users may be to recall the data of interest. Furthermore, the appropriate amount of time may be affected by a number of factors: for example, the demographics of the user community (e.g., less time may be appropriate for younger users), the immediacy of the information need (e.g., less time may be appropriate if the user indicates that the information is needed right away), even the date on which the user posed their question (e.g., follow-up should perhaps be faster at the end of a semester).

Response rate

After interaction data is collected, a common problem is that response rates to subsequent survey attempts tend to be low. Jeffrey Pomerantz and Lili Luo report an 8.6% response rate on an exit survey immediately following the interaction, and a 25.7% response rate for follow-up interviews[16]. Other authors report similarly low response rates on exit surveys, from 14.2% on the low end[17] to 32% on the high end[18].

While low response rates are a problem for any study, these findings are at least consistent with a two decades-long trend of increasing survey nonresponse in all fields. Motivated by this trend, there has been a great deal of work done to identify the causes and potential remedies for survey nonresponse[19], and some of these may be appropriate for remote reference services to use.

The motivation to respond to any data collection effort varies from person to person. This motivation is essentially a cost-benefit equation: the user must perceive that the benefit of responding (e.g., to herself, to the library, to society at large) outweighs the cost (e.g., the inconvenience, the time required). There are a variety of ways to change this cost-benefit equation so that users may be more inclined to respond to a library's data collection efforts. These include:

- Convincing respondents of the importance of their response: Offer online "suggestion" and "complaints" boxes to obtain feedback, and post responses publicly to demonstrate that efforts taken to give feedback are taken seriously and acted upon[20].
- Reducing the perceived burden to respondents: Make it easy for library users to give feedback quickly. For example, a library could employ some of the interactive web-based tools that are becoming commonplace, such as incorporating a "rate this answer" clickable feature on every response.
- Providing incentives: While a library may have budgetary or policy constraints that make this difficult or impossible, incentives can be a powerful motivator. Some examples include: a lottery giving away gift certificates to local or online vendors, or waiving library fines for participation.

Ambiguous data from users

For any data collection effort, it is important to pilot test the data collection instruments to ensure that the instruments in fact collect the desired data. Pre-testing can reveal whether questions are stated ambiguously, or respondents are misunderstanding questions or instructions. If these problems are not detected in advance of the survey or interviews, the data collected may be difficult to interpret, or simply useless. Pilot tests should be conducted with members of the library's user community from whom data will be collected, so that the perspectives of these users can be taken into account when creating data collection instruments.

One of the most important issues to address in a pilot test of a data collection instrument is to clarify ambiguous words. For example, the word "use" is especially problematic in the context of reference services, as there can be disagreement about what constitutes the use of information provided by the service. Lorri Mon found that some remote reference users felt that simply reading a librarian's response was "using" the information, even if they did nothing further with it afterward[3]. Jeffrey Pomerantz and Lili Luo noted that some librarians who were themselves users of another library's reference service did not perceive providing information to patrons to be a use of the information[16]. To avoid this problem, it may be more effective to ask whether information was "useful" rather than whether it was "used." Additionally, definitions and examples of other important but potentially ambiguous terms should be provided on data collection instruments to reduce misunderstandings.

While these researchers were able to discover during interviews the unexpected ways in which their study participants were interpreting the questions asked, participants in an online survey situation are typically confined to "yes/no" or multiple-choice selections with no opportunity for further clarifying their responses. Although multiple-choice, Likert scale, yes/no, and other closed-ended questions are useful for collecting certain types of data, they should be used sparingly on data collection instruments as they do not allow users explain their views in their own words. Data collection instruments should include open-ended essay questions that allow users to respond more freely. Some example questions include: "What was helpful to you?" "What was not helpful to you?" "What suggestions or recommendations do you have?" In some cases, essay questions may also be implemented on web-based surveys with too-short space limits, causing users' feedback to be cut off abruptly. To prevent truncation of users' responses[21], the placing of limits on the length or number of characters to be typed by users should be avoided.

Users are often asked in surveys and interviews to rate their satisfaction with reference services on scales typically ranging from 'very satisfied' to 'very dissatisfied.' However, Likert scales can be problematic for reference assessment, as they seem to lead to data on users' assessments of the service being skewed towards the positive. It is common for services to report high satisfaction rates of over 60%[22]. But there have long been questions raised about exactly what this expressed satisfaction means.

Herbert Goldhor, for example, speculated that users may "appreciate the effort without scrutinizing the results too closely,"[23] and some researchers have also suggested that users may have low expectations of librarians in general[24]. It is not uncommon, for instance, for users to express surprise upon hearing that librarians hold a Master's degree. If a user is unaware of the research efforts a skilled librarian is capable of, it may be difficult for the user to accurately assess the quality of librarian efforts in answering their request. Rachel Applegate discusses such situations as a potential problem of “false positives” wherein users may report satisfaction with an inferior product[25].

While many services would be thrilled by a greater than 60% satisfaction rate, artificially high satisfaction rates may have the unfortunate effect of blinding a service to user dissatisfactions that have not been captured. Service evaluations should therefore “triangulate” satisfaction data, supplementing it with other assessment approaches. For example, users should be asked not only about their satisfaction but also about their past experiences with the service, their expectations, whether their expectations were met, and whether the information received was useful and helpful. Pairing user assessments of an interaction with librarian assessments of the same interaction is another method of obtaining a more detailed evaluation.

Data not collected

Conducting more detailed evaluations requires collecting more detailed data. This section addresses measures that many services do not collect, but which would be useful to more fully inform evaluations.

Repeat users

The first of these measures is the percentage of repeat versus first-time users of the service. Many services track trends in use over time, but most analyze this usage data only in the aggregate, and discard all personally identifying data in order to protect the privacy of users. As a result, the key service quality standard recommended by Charles R. McClure and colleagues of “rate of repeat users” cannot be ascertained[26]. John V. Richardson, Jr. also pointed to users’ returning with another question as a positive service indicator[27]. Asking assessment questions regarding users’ “willingness to return” has been used as an alternative[28], but a user’s stated willingness to return may not necessarily lead to the user actually returning.

While there is no question about the importance of protecting users’ privacy, it would be useful for reference services to know more about patterns of one-time and repeat use, such as how implemented service changes have impacted users’ rate of return. Protecting users’ privacy and collecting more data about patterns of repeat use are not mutually exclusive as it might seem: Scott Nicholson and Catherine Arnott Smith describe a promising method for “deidentifying” user data based on the guidelines from the Health Insurance Portability and Accountability Act (HIPAA), while still retaining enough data to enable library evaluation[29].

Non-users

The inverse of repeat users of a service is non-users. Understanding why users choose *not* to use a remote reference service is an important aspect of assessment that is often overlooked. Non-users include those who have never used the service, as well as those who have used the service but then choose not to return. Studies of non-users in the library reference literature are relatively rare[30], but have the potential to yield valuable insights into users' motivations for choosing to use the service, and users' information seeking behavior more broadly.

One known group of non-users of library services are those individuals who make use of services indirectly, through proxies. Melissa Gross refers to these library users as "imposers," who send others to obtain library materials or ask reference questions on their behalf[31]. Gross delineates a variety of different types of audiences for indirect use including the homebound (elderly, ill, or disabled), those facing other barriers such as inability to speak English, employees seeking help on behalf of employers, and computer users making requests on behalf of non-computer users. While it is well known that indirect use of library services occurs, it is not always obvious when it is occurring, so data on indirect uses may not be accurately collected or counted.

Another group of non-users – or more accurately, partial users – are blocked or dropped users. Pascal Lupien details a variety of technical problems that can interfere with chat software operation, blocking chat users from accessing the service or dropping them out during the interaction: these include popup blockers, firewalls, incompatible browsers, and operating systems that may not be supported by vendors[32]. Alice Kawakami and Pauline Swartz describe chat technical errors by the librarian that may result in chat software problems[33]. Technical problems can also potentially be caused by rollouts of new versions of browsers or operating systems if incompatibilities exist with the chat software already in place at the service.

Missing data

One final problem that can occur in any evaluation is missing or omitted data, either through failure to include data that have been collected, or failure to collect data from all parts of the user population.

In some cases, services may "throw out" collected data perceived as not important enough for analysis, as for example with chat transactions that fail due to a technical problem. Services may decide that these technical failures are not important enough to count or analyze, focusing assessment only on interactions that were successfully completed. Matthew R. Marsteller and Danianne Mizzy, for example, removed "Technical Problems," which was their largest category, accounting for 32% of transcripts, from their sample before analysis, stating that "although the level of Technical Problems was a concern for the service, its only effect on the study was to reduce the sample size."[34] If done as a regular practice, this could potentially obscure an important assessment issue: how often do users encounter technical problems that obstruct their ability to access the service? There are indications that chat technical problems, blocked access, and sudden disconnects are common across services[35], making this an important issue for services to track both as a metric of the quality of the service provided and as evidence for service planning, software selection, and budgeting decisions.

Missing data may also be an issue when small sub-groups within user populations are not considered during the planning of survey and interview assessment. Patricia Katopol advocates greater use of “inclusive research” techniques in LIS such as snowball sampling for gaining participation from underrepresented populations, pointing out that random sampling may omit ethnic and other minority groups[36]. Lorri Mon, by making participation available to all chat and email users during the study period instead of using random sampling, was able to obtain interviews with members of various minority groups in a university’s remote reference user population including senior citizens, a disabled user, and speakers of English as a second language[3]. For particular evaluation efforts, it may be important for a service to hear from diverse user groups beyond the typical “majority user,” thus necessitating extra efforts in sampling and recruitment.

Improving remote reference service evaluation

Although there are a number of strategies to improve remote reference services, conducting better evaluations with better use of method and data collection techniques related to the service is certainly one of best places to start. This section identifies a number of key issues and factors that should be considered as a basis for improved method and data collection when conducting evaluations of remote reference services.

A well-known example from the library literature illustrates the importance of appropriate methods, and reliable and valid measures. In an evaluation of the accuracy of answers provided by reference services, Peter Hernon and Charles R. McClure found that approximately 55% of answers were accurate[37]. This “55% Rule” has been controversial, and other measures have been proposed as more appropriate for evaluating reference services, such as various aspects of user satisfaction[38]. Additionally, other methods for evaluating the accuracy of answers provided by reference services have found significantly higher accuracy rates[39].

This example illustrates the critical importance of careful planning of evaluations of remote reference services, and indeed, any library service. A range of measures may be useful, but it is critical for the evaluator to decide which measures are the most appropriate for the context of the evaluation. Peter Hernon and Charles R. McClure’s and Neal Kaske and Julie Arnold’s methods for evaluating answer accuracy are mutually exclusive[40]; if the evaluation of answer accuracy is important, then the method most appropriate for the context must be selected. Both of these methods, however, may be used alongside Joan Durrance’s satisfaction metrics[41], and by using multiple methods a fuller picture of the service will emerge.

Experienced evaluators realize that oftentimes the degree to which quality data can be collected for use in an evaluation is a trade-off. For example, if the evaluation had more time, more resources, and if the evaluators were more skilled, the quality of data could be improved. But in fact, there never seems to be enough time or resources for conducting any kind of evaluation, and this includes evaluation of remote reference services. Thus, the evaluator is constantly faced with a competing array of difficult decisions about how best to maximize the quality of the data yet still complete the evaluation and produce findings that assist in decision making.

Clearly, error cannot be eliminated in the conduct of evaluation, but error can be reduced and quality of data increased by any number of techniques – a number of which are described in this paper. It may be less important which types of errors are reduced and which types of techniques to improve quality are used by the evaluator, than that some techniques are employed. Some increase in the quality of data is better than no increase in the quality of data. And "good enough" data is better than no data when one is aware of the limitations of the data.

Providing adequate training to those evaluating remote reference is critical to the success of the evaluation. One of the best investments that a library can make is to have the evaluators attend classes on evaluation methods and data collection techniques, attend conferences, read basic texts on evaluation, and conduct evaluations under the tutelage of someone with significant experience in conducting such evaluations. Evaluation knowledge and skills are an important factor in producing evaluation results that can improve remote reference services.

While there certainly is a need to conduct evaluations that are comparable across different types of libraries and different types of remote reference delivery systems, obtaining comparable data is extremely difficult. A nearly infinite number of situational factors may affect evaluation at one location as opposed to another location. As a result, the likelihood that data can, in fact, be compared meaningfully across locations is unlikely. Thus, the authors argue that those engaged in evaluation of remote reference services strive to do outstanding evaluations in their particular library or situation with the objective of improving the quality and impact of that remote reference service – and not worry about producing data that are comparable across different libraries.

Some attention should be given to additional research on how best to conduct evaluation of remote reference services. Researchers should do direct comparisons among different methods and approaches to better understand what types of evaluation methods and data collection techniques are best for various types of remote reference services and under specific situational factors. For example, there may be great potential in server log evaluation techniques, but little research in this area has been done. Such research may be able to suggest evaluation strategies that are more efficient and produce better findings than the traditional techniques currently in use.

Evaluation, in general, is not for the faint of heart. It is especially not for the faint of heart if the evaluation is of remote reference services. Not only does the evaluator need to have a range of skills and knowledge related to evaluation, she needs to understand the complex technological aspects of remote reference services; needs to have excellent political skills to be able to work effectively with a broad range of stakeholders; needs to understand internal policies and procedures as well as vendor operations and procedures; and needs to remain objective and aware of numerous factors that could affect the evaluation, controlling those factors as best as possible. Nevertheless, ongoing evaluation of remote reference services is essential to improving those services.

The issues and strategies identified in this paper to improve the evaluation of remote reference services have come from many years of experience and research on the part of the authors. Being aware of these issues and developing strategies to improve the quality of remote reference service is an important step to improving the usefulness and impact of such evaluations, and ultimately improving the quality and usefulness of these services to users.

Notes

[1] Jonathan D. Eldredge, "Evidence-based librarianship: An overview," *Bulletin of the Medical Library Association* 88, 4 (October 2000): 289-302.
[2] Anne Grodzins Lipow, *The Virtual Reference Librarian's Handbook* (New York: Neal-Schuman Publishers, Inc., 2002): xix.
[3] Lorri Mon, "User Perceptions of Digital Reference Services," Ph.D. dissertation, (University of Washington, 2006).
[4] Jeffrey Pomerantz, "A Conceptual Framework and Open Research Questions for Chat-Based Reference Service," *Journal of the American Society for Information Science and Technology* 56, 12 (2005b): 1288-302.
[5] Joseph Janes, *Introduction to Reference Work in the Digital Age* (New York: Neal-Schuman Publishers, Inc., 2003): 69.
[6] Marie L. Radford, "Encountering Virtual Users: A Qualitative Investigation of Interpersonal Communication in Chat Reference," *Journal of the American Society for Information Science and Technology* 57, 8 (2006): 1046-59.
[7] Phillip M. Hallam-Baker and Brian Behlendorf, "Extended Log File Format: W3C Working Draft WD-logfile-960323" (N.D.), http://www.w3.org/TR/WD-logfile.html (accessed September 12, 2007).
[8] Denise Troll Covey, "Usage and Usability Assessment: Library Practices and Concerns" (Washington, DC: Council on Library and Information Resources, 2002): 41. http://www.clir.org/PUBS/reports/pub105/contents.html (accessed September 12, 2007).
[9] John Carlo Bertot, Charles R. McClure and Joe Ryan, *Statistics and Performance Measures for Public Library Networked Services* (Chicago: American Library Association, 2001).
[10] Claes Fornell, Michael D. Johnson, Eugene W. Anderson, Jaesung Cha, Barbara Everitt Bryant, "The American customer satisfaction index: Nature, purpose, and findings," *Journal of Marketing* 60, 4 (October 1996): 7-18.
[11] Donald S. Elliott, Glen E. Holt, Sterling W. Hayden and Leslie Edmonds Holt, *Measuring Your Library's Value: How to Do a Cost-Benefit Analysis for Your Public Library* (Chicago: American Library Association, 2006).
[12] David S. Carter and Joseph Janes, "Unobtrusive Data Analysis of Digital Reference Questions and Service at the Internet Public Library: An Exploratory Study," *Library Trends* 49, 2 (Fall 2000): 251-65.

[13] Jeffrey Pomerantz, “A Linguistic Analysis of Question Taxonomies,” *Journal of the American Society for Information Science and Technology* 56, 7 (May 2005): 715-728.
[14] Neal Kaske and Julie Arnold, “Evaluating the Quality of a Chat Service,” *portal: Libraries and the Academy* 5, 2 (April 2005): 177-93.
[15] Lorri Mon and Joseph Janes, “The Thank You Study: User Feedback in Email ‘Thank You’ Messages,” *Reference and User Services Quarterly* 46, 4 (2007): 53-59.; John V. Richardson, Jr., “Understanding the Reference Transaction: A Systems Analysis Perspective,” *College & Research Libraries* 60, 3 (May 1999): 211-22.
[16] Jeffrey Pomerantz and Lili Luo, “Motivations and Uses: Evaluating Virtual Reference Service from the Users’ Perspective,” *Library & Information Science Research* 28, 3 (2006): 350-73.
[17] J. B. Hill, Cherie Madarash-Hill and Ngoc Pham Thi Bich, “Digital Reference Evaluation: Assessing the Past to Plan for the Future,” *Electronic Journal of Academic and Special Librarianship* 4, 2-3 (Fall 2003).
[18] Margie Ruppel and Jody Condit Fagan, Instant Messaging Reference: Users’ Evaluation of Library Chat,” *Reference Services Review* 30, 3 (August 2002): 183-97.
[19] Robert M. Groves, Don A. Dillman, John L. Eltinge, Roderick J. A. Little, *Survey Nonresponse* (New York: Wiley InterScience, 2001).
[20] Terry G. Vavra, *Improving Your Measurement of Customer Satisfaction: A guide to creating, conducting, analyzing, and reporting customer satisfaction measurement programs* (Milwaukee, WI: ASQ Quality Press, 1997); Peter Hernon and Danuta A. Nitecki, “Service Quality: A Concept Not Fully Explored,” *Library Trends* 49 4 (Spring 2001): 687-708.
[21] Bruce Stoffel and Toni Tucker, “Email and Chat Reference: Assessing Patron Satisfaction,” Reference Services Review 32 no. 2 (2004): 120-140.
[22] Jo Kibbee, David Ward and Wei Ma, “Virtual Service, Real Data: Results of a Pilot Study,” *Reference Services Review* 30, 1 (2002): 25-36; Stoffel and Tucker; Pomerantz and Luo.
[23] Herbert Goldhor, “The Patrons’ Side of Public Library Reference Questions,” *Public Library Quarterly* 1, 1 (Spring 1979): 45.
[24] Patricia Dewdney and Catherine Sheldrick Ross, “Flying a Light Aircraft: Reference Service Evaluation from a User’s Viewpoint,” *RQ* 34, 2 (Winter 1994): 217-29; Goldhor.
[25] Rachel Applegate, “Models of User Satisfaction: Understanding False Positives,” *RQ* 32, 4 (Summer 1993): 525-539.
[26] Charles R. McClure, R. David Lankes, Melissa Gross and Beverly Choltco-Devlin, *Statistics, Measures and Quality Standards for Assessing Digital Reference Library Services: Guidelines and Procedures* (Syracuse, NY: Information Institute of Syracuse, 2002). http://quartz.syr.edu/quality/ (accessed September 12, 2007).
[27] John V. Richardson, Jr.: 219
[28] Joan C. Durrance, “Reference Success: Does the 55 Percent Rule Tell the Whole Story?” *Library Journal* 114, 7 (April 15, 1989): 31-36.
[29] Scott Nicholson and Catherine Arnott Smith, “Using Lessons from Health Care to Protect the Privacy of Library Users: Guidelines for the De-Identification of Library Data Based on HIPAA,” *Journal of the American Society for Information Science and Technology* 58, 8 (2007): 1198-206.

[30] John Lubans Jr., “Nonuse of an Academic Library,” *College & Research Libraries* 32, 5 (September 1971): 362-367; Mary Jane Swope and Jeffrey Katzer, “Why Don’t They Ask Questions?,” RQ 12, 2 (Winter 1972): 161-66.

[31] Melissa Gross, “The Imposed Query: Implications for Library Service Evaluation,” *Reference & User Services Quarterly* 37, 3 (Spring 1998): 290-99.

[32] Pascal Lupien, “Virtual Reference in the Age of Popup Blockers, Firewalls, and Service Pack 2,” *Online* 30, 4 (July/August 2006). http://www.infotoday.com/online/jul06/Lupien.shtml (accessed September 12, 2007).

[33] Alice Kawakami and Pauline Swartz, “Digital Reference: Training and Assessment for Service Improvement,” *Reference Services Review* 31, 3 (2003): 227-36.

[34] Matthew R. Marsteller and Danianne Mizzy, “Exploring the Synchronous Digital Reference Interaction for Query Types, Question Negotiation, and Patron Response,” Internet Reference Services Quarterly 8, 1/2 (2003): 156.

[35] Jo Kibbee, David Ward and Wei Ma; Ian J. Lee, “Do Virtual Reference Librarians Dream of Digital Reference Questions?: A Qualitative and Quantitative Analysis of Email and Chat Reference,” *Australian Academic & Research Libraries* 35 2 (June 2004): 95-110; Pascal Lupien; Bruce Stoffel and Toni Tucker.

[36] Patricia Katopol, “Inclusive Research: Including Everyone in LIS Research,” Unpublished paper in progress (in process).

[37] Peter Hernon and Charles R. McClure, “Unobtrusive Reference Testing: The 55 Percent Rule,” *Library Journal* 111, 7 (April 15, 1986): 37-41.

[38] Joan C. Durrance.

[39] Matthew L. Saxton and John V. Richardson Jr., *Understanding Reference Transactions: Transforming an Art into a Science* (Amsterdam, NY: Academic Press, 2002); Neal Kaske and Julie Arnold.

[40] Peter Hernon and Charles R. McClure; Neal Kaske and Julie Arnold.

[41] Joan C. Durrance.