

Assessing Information Literacy Skills: Developing a Standardized Instrument for Institutional and Longitudinal Measurement

Lisa G. O'Connor, Carolyn J. Radcliff, and Julie A. Gedeon

Introduction

Academic librarians are embracing information literacy as an instructional framework. Recently developed national standards for information literacy have the potential to unify efforts around the country and to clarify, for librarians, administrators, and faculty, the desired outcomes of library instruction. At a time when universities and colleges are being held to high levels of fiscal responsibility and as libraries increasingly funnel funds and staff resources into information literacy programs, there is a pressing need to answer the question “are these programs making a difference?” Libraries need a valid and reliable method of assessing the impact of their instructional programs on their institutions over time.

Although many researchers have developed tools for measuring students’ knowledge of a specific library system or database, and for determining affective responses to library instruction (e.g., degree of confidence felt by students), there is not yet a standardized method for measuring information literacy that is easily administered and ap-

plied across institutions.

Our research program tackles this need along two lines. First, we are currently developing an instrument for basic evaluation of information literacy skills. Inspired by the Wisconsin-Ohio Reference Evaluation Program (WOREP) developed by Marjorie Murfin and Charles Bunge (American Library Association), the key desired characteristics of this assessment instrument include:

- assesses at the institutional level, not the instructor level
- allows for longitudinal data gathering
- allows for pre-testing and post-testing
- is quick and easy to administer
- is a standardized instrument for use at any institution
- is geared toward national standards for information literacy
- assesses both cognitive and affective dimensions

We envision combining pre- and post-testing with experimental and control conditions to answer the questions, “Does library instruction make a difference on campus? And,

Lisa G. O'Connor is a business reference librarian at Kent State University. Carolyn J. Radcliff is a reference librarian at Kent State University, and Julie A. Gedeon is a senior institutional research officer at Kent State University

"Does library instruction lead to acquisition of information literacy skills?"

Second, teaching of information literacy skills will be assessed through session-specific evaluation. A companion instrument will be developed to evaluate alternative methods and approaches to library instruction. This second instrument will gather information from both students and instructors, and, like the first instrument, will be geared to national information literacy standards. Both instruments will use similar items and scoring procedures so that their results can be compared. The use of the second instrument will answer questions such as, "What instructional method is most effective for teaching information literacy skills?"

Our work on the project coincides with the development of national information literacy standards and guidelines by the American Association of School Librarians (AASL) and the Association of College and Research Libraries (ACRL). The AASL standards were published in 1998 in the book *Information Power: Building Partnerships for Learning*. There are nine standards grouped into the three categories of information literacy, independent learning, and social responsibility. ACRL developed the "Information Literacy Competency Standards for Higher Education," consisting of five standards with performance indicators. Before the ACRL standards were published, the Kent State University libraries adapted the AASL work to our environment. As a result, we have at Kent State a set of information literacy standards with learning objectives that closely resemble a mix of AASL and ACRL standards.

This paper reports on the early stages of the development of the first instrument as described above, a standardized tool for assessing information literacy skills.

Literature Review

Hoping to find an existing instrument, we searched the literature since 1980 for an instrument that could be used to assess the information literacy skills of students longitudinally and across institutions. Upon finding no adequate existing instrument, we reviewed the literature for information that might assist us in the process of creating a new instrument. The literature on information literacy assessment falls into distinct categories: literature review articles that tend to describe the need for assessment and discuss the political and pragmatic barriers to library assessment; theoretical articles that discuss the various types of assessment and describe the strengths of each type for evaluating library instruction; and reports of assessment projects. Ragains characterizes the literature of instructional assess-

ment in two categories. He writes, "much of the literature on evaluation tends either to be based in practice (i.e., 'this is how we did it') or prescriptive (i.e., 'this is what you can/should do')" (Ragains 160).

Descriptive Literature

Descriptive library literature demonstrates little experience in formalized evaluation in general and contains nothing that might be used for a standardized, longitudinal, and cross-institutionally administered assessment instrument. "Surveys published since 1980 reflect an increase in the number of institutions implementing evaluation as part of their BI programs. However, formal evaluative methodologies are still not being applied to any significant degree" (Bober, Poulin, and Vilen 57). According to Mensching's 1987 survey of LOEX-participating libraries, only 23% of respondents evaluated BI were using an assessment mechanism. As Coupe summed up, "Perhaps one reason that librarians have neglected the measurement of basic library skills is the lack of an adequate survey instrument" (189).

Another trait of existing assessment programs is that they emphasize measuring the efficacy of individual components of instruction in order to plan for improvement, rather than assessing whether library instruction forwards the instructional goals of the institution. According to Lindauer, "Almost none of these publications provides measures or methods for assessing the impact of academic libraries on campuswide educational outcomes. Overwhelmingly, the literature is internally focused, looking at the academic library as an overall organization or at one or more of its components or services" (548). Pausch and Popp agree: "Review of the recent literature on assessment of library instruction reveals few changes in the formal evaluation methodologies employed by librarians. In fact, evaluation of any kind is more likely to be informal in nature, as is noted. . . . Where formal evaluation is being carried out, little full program assessment is being done. . . . These studies also reveal that no control groups are used."

We reviewed eight articles that reported using a "paper and pencil" test to assess information literacy skills: Hardesty, Lovrich and Mannon (*Evaluating Library-Use Instruction* 1979), Kaplowitz (1986); Tiefel (1989); Coupe (1993); Franklin and Toifel (1994); Kunkel, Weaver and Cook (1996); Colborn and Cordell (1998); Lawson (1999); and Rabine and Cardwell (in press). With the exception of Kunkel, Weaver and Cook, all studies appended their instruments either in their entirety or with a significant sampling of the items used. All of the instruments included

questions on basic library skills, such as Library of Congress Subject Headings, call number comprehension, locations of various services or resources within the libraries, the purpose of Boolean operators, citation interpretation, OPAC usage, and basic search construction. Most also included items to (1) assess library related attitudes and behaviors; (2) allow for student self-assessment of skills, and (3) gather basic demographic information. Instruments contained between 9 and 28 items and were administered to as few as 111 students and as many as 1,702, with most studies including between 200 and 400 students. Five of these studies used a pre- and post-testing process; three of the pre- and post-tests were identical instruments, with the exception of additional affective questions added to the post-test. Only Tiefel's study reported that researchers requested student ID numbers. Three of the instruments were subject specific and all of them contained questions specific to the investigators' libraries. Control groups were not utilized in any of the studies. Although Kaplowitz was able to use a naturally occurring group who did not participate in the study, that group was too small to be used as a formal control group. Most of the instruments were administered within 2 to 4 weeks of instruction, with the exception of a follow-up study by Fry and Kaplowitz which will be discussed in the longitudinal portion of this literature review.

Although many of these instruments provided us with ideas in planning the design of our own test, none of them were considered appropriate for the purposes of this program. Aside from being either subject specific or oriented to a specific institution, most of these instruments could not be considered scientifically valid and reliable. In fact, with the exception of Tiefel and Colborn and Cordel, there was very little discussion of instrument development. Understandably, authors seem to be primarily concerned with creating an instrument that can be used to improve instruction and provide internal accountability for their instructional time and effort. Thus, the instruments are developed quickly and the gathered data is the main focus of the research reports. We believe, however, that in order to measure campuswide learning outcomes, a more rigorous process must be demonstrated.

Tiefel's project at Ohio State University incorporates a careful development process. In 1986, Tiefel's instrument was administered to a large group of 1,702 students. Item-by-item analysis occurred only with this large-scale testing, revealing notable weaknesses with the instrument. A more thorough trial process might have identified weak questions previous to large group trials. Tiefel writes that nine

years of experience in instructional evaluation have taught librarians at Ohio State University that it is critical to "ensure that evaluations reflect accurate measurements. Evaluations of LIP [Library Instruction Program] have varied in validity and reliability over the years, but in the future, careful attention will continue to be given to planning and implementing only the most credible studies possible" (258). Aside from these development problems, this project suggests that "Ohio State's Library Instruction Program has brought about a statistically significant improvement in students' knowledge about the library, their ability to use libraries, and their attitudes toward libraries and librarians" (258). This difference was measured two weeks after the students received library instruction, leaving open the question of long-term effects. As desirable as this situation is from a library point of view, what is lacking from the University administration perspective is any evidence that these skills affect student success and retention over the course of time. Other shortcomings of this instrument for use in our project include a lack of control group and the use of identical pre- and post-tests, which makes the student improvement of 7% on average suspect when one factors test experience into the equation.

Colborn and Cordell developed an instrument to measure knowledge in five fundamental areas. The instrument was administered to 131 students who returned 129 completed and usable tests. Their resulting data showed no significant difference between pre- and post-test results. Although this study details the most rigorous development process, in which the authors used both a difficulty and discrimination index to examine all items and revising their instrument accordingly, no definite conclusions are drawn about the disappointing results. The authors are fairly certain that the test itself was not the weak link, however inadequate data was collected to rule that out. Their experiences at the very least are instructive about how difficult and unpredictable the test development process can be. Their instrument, even if it had proven successful, would be difficult to administer institutionally or across institutions because it contained fill-in-the-answer types of questions that would be cumbersome to score in a very large group setting. Additionally, their results were not reported at the level of detail needed to replicate their study. As Catts writes, "for assessment of information literacy to be accepted, all stakeholders must have confidence in the reliability of the assessments. This means that assessment must be internally consistent, and reproducible" (280).

Hardesty, Lovrich, and Mannon (*Evaluating Library-Use Instruction*) provide the most helpful model for this assess-

ment project. Their instrument, which contains ten attitudinal items and 26 items to test library use skills, underwent rigorous pre-testing for reliability and validity. They also utilized a control group in the testing phase to ensure the legitimacy of any significant differences discovered in their study. The pre-test was administered to 162 freshman prior to instruction and to the same group of students eight weeks following instruction. The pre-test was identical to the post-test, and no assessment was made of the effect the test-taking experience may have had on post-test results. Larger numbers of subjects would also have strengthened the authors' ability to generalize their results to their institutional population. Although the resulting instrument did not cover the wide range of information literacy skills that we are interested in assessing and thus was not sufficient to meet the goals of our study, this research project, complemented effectively by follow-up project by the same authors, achieves its stated purpose of providing, "a model of evaluation and its application, which may be of use to others interested in systematic assessment of instructional programs" (*Evaluating Library-Use Instruction* 315).

Longitudinal Studies

Another limitation with studies described in the literature is that most provide for little sense of longitudinal change. We examined four studies that provide an exception to this trend: Person (1981), Kaplowitz (1986) and its follow-up study Fry and Kaplowitz (1988), Selegean, Thomas, and Richman (1983), and two studies by Hardesty, Lovrich, and Mannon (*Evaluating Library-Use Instruction*, 1979 and *Library-Use Instruction*, 1982).

Person evaluated the perceived effects of a semester-long library skills credit course six years following its delivery. Person's 26-question survey did not attempt to measure skills, but instead assessed students' attitudes about the course in retrospect. Although Person's study showed that student appreciation for bibliographic instruction "frequently increases during the years after the course has been taken" (19), his results are somewhat questionable due to the low return rate on his survey. Of 730 distributed surveys, only 169, a little more than 25%, were returned.

Kaplowitz, whose study has been discussed previously in this literature review, partnered with Fry in 1982 to re-evaluate students three years after initial instruction and assessment. Their study suffers from the same low response rate reported in Person. Of 500 distributed instruments, only 98 or 19.6% were returned. One could reasonably suggest that the students who responded were more motivated

by positive feelings about library instruction than those who did not return the survey.

Selegean, Thomas, and Richman took a completely different approach, comparing the grade point averages and retention and graduation rates of 512 undergraduate students enrolled in a library skills course to students in a control group of students who had not taken the course. The comparison was conducted three to seven years after completion of the course. Statistical analysis of results indicated significant positive differences for GPA and retention rates, but no significant differences for graduation rates. Although this study does not link actual library skills to these variables in the same manner we hope to, it provides a precedent for the statistical analysis we would like to employ longitudinally for our instrument.

Hardesty, Lovrich, and Mannon, whose first study has been discussed previously in this literature review, used their instrument in a second, longitudinal study. A panel study was conducted of 82 students over the course of a three-year period to compare their scores on the skills test prior to and eight weeks after instruction and also later, when the students were seniors. Multiple regression analysis was conducted to determine the relative importance of the factors intellectual capacity, academic diligence, and the amount and types of instruction received throughout their careers "in the determination of the variation of library-use skills possession" (43). The study concluded that "library-use instruction is much more highly correlated with skill possession than either inherent intellectual ability or academic diligence" (43). This study is groundbreaking both for its thoroughness and scope. The level of detail provided about the development of the instrument makes the study highly replicable and unique in library literature. Lessons learned from Hardesty, Lovrich, and Mannon have provided us with a valuable model for own development process.

Prescriptive Literature

The prescriptive literature tends to bemoan the barriers to effective assessment programs and outlines what needs to be improved or implemented to create them. For example, in their literature review Bober, Poulin, and Vileno conclude, "Despite the growing interest in the evaluation of bibliographic instruction, the same pattern of limited systematic evaluation identified in 1980 continues to exist" (53). The most frequently cited reasons for the lack of instructional evaluation are shortage of staff time, limited financial resources, and lack of a well-developed standardized instrument. Other reasons cited were more political in

nature. Some authors assert that many libraries may not be ready to assess at their institutions, because they may not have “a formal process to enable information literacy goals to be realised” (Catts 272). Lindauer also recognizes this phenomenon, writing “perhaps not surprisingly, the author found that in most cases empirical research connecting college students’ experiences and outcomes with specific campus services and resources did not include any mention of libraries/learning resource centers” (553).

Institutionally, there also must be value attached to information literacy goals and some incentive for achieving them. Catts writes, “Rewards must also exist at the institutional level for achieving the learning outcomes associated with information literacy” (273). Literature prior to the 1990s also cites the absence of “generally accepted criteria” or clearly defined learning objectives as a barrier to effective assessment (Selegan, Thomas, and Richman 476); however with the development and approval of objectives through AASL and ACRL this obstacle has been removed. Colborn and Cordell are unique in appending the framework from which their instrument was developed to their article, and as a result their work provides a useful model for this project.

Prescriptive literature also clearly identifies the difficulties associated with developing a standardized measure of learning. According to Catts, “The difficulty of developing unbiased, valid and reliable tests is well recognized and described” (277). Catts identifies common errors in formalized testing methodology using three studies of standardized measurement as examples. In fact, the ability to measure the higher level concepts of information literacy at all is also questioned. For example, Tiefel asserts that “it is not possible to test the program at the level of concept mastery and transferability” (250). This manner of circumspection may explain in part the concentration of available instruments covering the most basic library skills, rather than on the more difficult conceptual knowledge.

Measurement Issues—Item Response Theory

Our review of the literature showed that although there are numerous studies assessing learning and knowledge in libraries, there are methodological flaws in those studies, especially in the analysis of data collected by these instruments and the assumptions made about the instruments. We turn our attention now to the issues of measurement and how they affect instrument development.

Traditional methods of assessing validity and reliability are limited, in that they are sample specific as regards

the validity and reliability of the test items and the subjects’ scores. Reliance on classical test theory does not allow one to accurately *measure* respondents’ knowledge, as the score is based on merely counting the number of items answered correctly. In this study we have elected to use a method for more efficiently measuring students’ information literacy knowledge. This alternative to classical test theory, item response theory (IRT), goes beyond merely counting the number of correct items and permits an examination of not only item performance, but examinee performance together with item performance, and provides sample-free statistics for both items and examinees.

Item response theory is sometimes referred to as latent trait theory because it measures the latent trait underlying a set of exam items. It is assumed that responses to items can be accounted for by (usually one) underlying trait, in our case, information literacy. The theory is built on a mathematical model of how examinees at different ability levels for the trait should respond to an item. In contrast, classical item analysis statistics do not provide information about how examinees at different ability levels on the trait have performed on the item. Thus, IRT allows comparison of the performance of examinees who have taken different tests.

One central concept of IRT is the item characteristic curve (ICC), which plots the probability of responding correctly to an item as function of the latent trait, or ability, underlying performance on the items on the test. As the ability level of the respondent increases, so does the probability of answering correctly. The difference between the ICC and the item difficulty and discrimination statistics provided in classical test theory is that the ICC allows us to see how the probability of answering correctly depends on the underlying ability level.

The ICC represents the proportion of a subgroup of examinees at a specific ability level who would be predicted to answer the item correctly. The probability of responding correctly is the probability of a randomly chosen member of a homogeneous subgroup responding correctly to the item.

Another important concept of IRT is that it permits examinees to be compared even when they have not been given the same items—this is referred to this as test-free measurement. As long as the items are written to test the same latent trait, the model allows us to make this comparison.

There are three general models of IRT in use. The one-parameter model requires estimation of item difficulty only, assuming item discrimination is constant across all items. The two-parameter model requires estimation of item dis-

crimination as well as item difficulty. The three-parameter model adds a guessing parameter.

The model to be used in the analysis of data for this study is the one-parameter model developed by Georg Rasch, which is called the Rasch model (Wright and Stone). It requires us to assume that item discrimination is the same across all items and that guessing is not a serious consideration. One of the advantages of using a one-parameter model is that it is legitimate to use with smaller groups of examinees, because we are estimating fewer parameters. Also, the more complex the model, the less accurate the estimations are. The Rasch model has been shown to be robust to violations of the assumptions of constant item discrimination and no guessing.

To demonstrate this model further, Figure 1 shows a hypothetical example of a pattern of responses to a ten-item test. Although examinees 1, 2, 3, and 4 responded correctly to the same *number* of items (five), it can be seen that their patterns of response differ. For example, examinee 1 correctly answered only the five easiest items, while examinee 2 missed the easy items and responded correctly to the five most difficult items. Examinees 3 and 4 each answered some of the easier items and some of the more difficult items correctly. Although it might be tempting to give all four examinees the same score, it would be misleading because they responded differently to items of differing difficulty levels. Item response theory allows us to (1) determine the difficulty levels of the items and place them along a difficulty continuum and (2) analyze the response patterns of the examinees and place them along an ability continuum on the latent trait based on their individual response patterns.

Examinee 5 was not well served by this 10-item exam. Because he responded correctly to all the items, we do not know his true ability level on the latent trait. This individual should be given more difficult items until he is un-

able to respond correctly to them. Thus it can be seen that there is no “100% correct” when measuring with IRT.

Methodology

We were greatly assisted in this part of the project by material from *Standards for Educational and Psychological Testing*, jointly published by the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education. This work provides objective, scholarly guidance for developing test in accordance with accepted practices.

Instrument Development

The instrument is being developed in three distinct steps. Through careful design, testing, and re-testing, we hope to develop strong, robust items during the pilot phases of one-on-one trials and small group trials. Testing the items with large groups through field trials will then give us information for refining the instrument. Though this process is time consuming, we hope to avoid the problems mentioned by Tiefel with discovering substantial problems only after field testing with large numbers of students.

So far, we have completed two phases of instrument development. First, using our information literacy standards, which are based on the AASL and ACRL standards, we identified specific skills to demonstrate understanding of each of the sub-areas. We focussed on two standards, dealing with efficiently and effectively accessing information (AASL standard number 1; ACRL standard number 2), and evaluating information critically and competently (AASL standard number 2; ACRL standard number 3). We then began the task of crafting items which could be easily depicted and understood in a paper and pencil format. Refinement of the items took place over several meetings as each of us reacted to what the others had written. After reasonable items had been written, refined, and agreed upon, we conducted one-on-one

Figure 1: Hypothetical Example of a Pattern of Responses to a Ten-Item Test

		TEST ITEMS										# CORRECT	
		Easier					More difficult						
EXAMINEES		1	2	3	4	5	6	7	8	9	10		
	EX 1	1	1	1	1	1	0	0	0	0	0	0	5
	EX 2	0	0	0	0	0	1	1	1	1	1	1	5
	EX 3	0	1	0	1	0	1	0	1	0	1	1	5
	EX 4	0	1	1	0	0	1	1	0	0	1	1	5
	EX 5	1	1	1	1	1	1	1	1	1	1	1	10

Adapted from Crocker and Algina.

trials of the instrument with selected members of the target population. This phase provided feedback which helped to further refine some of the items as well as the layout of the instrument as a whole. Additionally, items to be used as criteria for further analyses were added to the knowledge items.

The next phase of this study, small group trials, involved collecting data in classrooms from subjects in the target population. Students in this section were asked to provide any feedback they wished, in addition to answering the questions. We were also able to have general group discussions about library skills and the instrument with some of these students. These data were used to more closely examine the items for further refinement for the next phase, which will be a large-scale administration of the instrument and very detailed analysis of how well the items are measuring the latent trait of information literacy.

One of our goals here is to create items of varying level of difficulty. In this way, we will be able to determine whether or not a student's skill level increases as they progress through college. We examined the difficulty level of items as we created them by using individual judgments of a panel of experts (Kent State librarians). An overall level of difficulty was assigned based on the predominance of responses. Where there was little agreement, we made a determination based on our own knowledge and the intent of the item. We expected that more able students (i.e., those with higher high school grade point averages (HSGPA) and entry test scores) would not only respond correctly to more items overall, but would be more likely to respond correctly to the more difficult items.

Our long-term goal is to create a web-based instrument instead of paper and pencil. In this first study, however, we concentrated on creating a paper-and-pencil version so that items could be written and tested within the time constraints of an academic semester. We wanted to concentrate on writing and validating items before engaging in the task of creating an electronic version.

Subjects

The target population for the final instrument to measure information literacy is college and university undergraduates. Therefore, all subjects used in this study were selected from the undergraduate population at Kent State University. We obtained permission to conduct our study from the University's Human Subjects Review Board.

Subjects for the one-on-one trials were six undergraduate students who worked in the Kent State Main Library.

Although we suspected that these students would be more knowledgeable than the general undergraduate population, one of the goals for this phase was to make sure the items were asking what we intended them to ask. We were also interested in readability of not only the individual items, but the instrument as a whole.

Subjects for the small group trials were students in five fall 2000 freshmen orientation classes. Subjects were selected without sampling and participation was voluntary. Of approximately 109 potential subjects, 91 completed and returned the instrument.

Data collection

For the one-on-one trials, students were asked to meet with us individually. They were given the instrument and asked to complete the knowledge items, to write any comments or questions on the instrument, and to provide verbal feedback while completing the items regarding ambiguities or their thought processes as they responded to items. We took notes on the oral feedback and asked for further information when necessary. We also noted the time it took subjects to complete the instrument.

For the small group trials, orientation instructors generously donated class time to the study. We explained the study, asked for participation, and distributed an informed consent letter with the instrument. Subjects were asked to complete the instrument to the best of their ability and to write any comments or questions on the instrument itself. We collected the completed instruments.

Analyses

Responses to the one-on-one trials were examined individually to determine whether there were any obvious misunderstandings. Comments and questions from the subjects were used to refine the items and improve the layout and readability of the entire instrument.

Responses for 90 of the 91 subjects collected during the small group phase were coded for data entry and analyzed using SPSS PC. One subject was eliminated because his response pattern was suspect. Several different analyses were done to determine how well the items were functioning. However, because the number of respondents was low, detailed item analyses were not performed on these data.

First, an examination of straight frequencies of responses was done to determine if there were any problems with data entry or coding. Additionally, the frequencies of responses were examined to provide evidence of the difficulty level of the items. For example, if 70% or more of the

respondents answered correctly, this was one indication that the item might be easy. Additionally, these frequencies were compared to the categorizations by the subject experts to determine whether the items deemed easy by them were indeed responded to correctly by a majority of the subjects.

Crosstabulations of the responses to the knowledge items with several perception and affective items (covering self-perceived levels of knowledge, self-perceived level of success completing the instrument, library use, level of comfort in using the library, amount of previous library instruction, and class status) were run to determine if students who felt more knowledgeable, more comfortable, etc. responded correctly more frequently.

Student records on the university's Student Information System were pulled for the subjects who had provided valid student identification numbers (SIDs). Information regarding their HSGPA and ACT composite or SAT combined scores were gathered. One student provided no SID, seven students provided invalid SIDs, and one student was a transfer student with no HSGPA in the system. Additionally, 4 students with valid SIDs had no test scores in the system (Kent does not require ACT or SAT scores for general admission to the university). Because most Kent State students submit ACT scores instead of SAT scores, in those few instances where we had SAT scores with no ACT score we converted the SAT score to ACT equivalents, using a model developed at Kent State University (Kuhn). HSGPAs and ACT scores were then recorded into two groups, high and low, based on the arbitrary cut-off value of the means for fall 2000 new freshmen at Kent State. Test scores at and below the mean of 21 were grouped into "low" and those above 21 were grouped into "high." HSGPAs at and below the mean of 3.00 were grouped into "low," while those above 3.00 were grouped into "high."

Crosstabulations were run with knowledge items versus HSGPA and ACT to determine if higher-ability students responded correctly more often, especially to the more difficult items as judged by the experts and as compared to the straight frequencies.

Results

For the majority of items reviewed by the subject experts, there was agreement of 60% or more as to the difficulty level of the item. For 10 items, we reviewed the categorizations made by the experts and made an overall classification based on the specific item as well as the original intent of the item as we wrote them. For those items for which there was greater discrepancy among the subject experts, it

was agreed that closer examination should be done of the responses by subjects.

The subjects who responded to the one-on-one instrument pointed out several ambiguities or questions. See Figure 2 for an example of a re-worked set of questions. One potential problem identified was the number of responses allowed for each item: some items required only one response, whereas others allowed respondents to mark all relevant answers. Numerous items were rewritten and the overall layout of the instrument was improved. For example, instructions were enlarged and set in bold type so that the number of acceptable responses (only one versus all that apply) was more clear. Additionally, four of the items regarding the process of gathering information for a term paper, which in our minds belonged together, were enclosed in a box and put on one page to make it more clear that these questions all referred to the same scenario. The instrument was also altered to include a "please proceed to the next page" statement at the bottom of each page. At the end of this phase, the instrument had 49 items.

The results of the pilot provided useful information regarding the items. The difficulty level as judged by the experts was compared to the frequency of correct responses by subjects. For 18 of the items, the actual results differed from the overall expert rating. Eight of these items merit future close scrutiny because of the degree to which the responses differed from the expert ratings. Where there was discrepancy we looked more closely at the items in an attempt to determine why. Several items proved to be easier than expected because of the format. For example, the item asking students to match primary, secondary, and tertiary sources with their definitions gave three terms and three definitions, thus allowing students to pick the correct response for the third based on elimination of the other two responses. Therefore, it was not necessary for a student to actually know what a tertiary source was if he or she knew primary and secondary. Another potential problem item concerned the concept of truncation. It was discovered in the one-on-one phase, and based on the comments and questions subjects wrote during the small group trials, that using the word "truncation" was causing many more students than expected to answer incorrectly or indeed not respond at all to the item. This particular case raises an interesting question. Are we interested in know whether students can perform truncation, can identify the term "truncation," or both? In either case, we want to make sure we are measuring what we think we are.

Examination of the crosstabulations of knowledge items with the perception and affective items provided no addi-

Figure 2: Example of Content Changes Resulting from One-on-One Trials

ORIGINAL VERSION

Each of the following statements is true about the library or the World Wide Web. Identify which statements describe the library or the Web.

Use **W** if the statement is true about the Web

Use **L** if the statement is true about the library

Use **B** if the statement is true about both the library and the Web

- Has information that has been through traditional publishing process.
- Has information that is sold by publishers.
- Has a classification system.
- Has information provided by organizations, individuals, companies, and governments.
- Is available 24 hours a day.

REVISED VERSION

Academic libraries are generally thought of as collections of materials in print and electronic formats. Some of these materials are made available to users through the Web, but are not included in what we traditionally think of as the Web.

The World Wide Web is a means of communication. Computers all over the world network with on another by using a common language.

Which of the following statements are generally true about academic libraries and/or the Web?

Put a **W** if the statement is true about the Web

Put an **L** if the statement is true about the library

Put a **B** if the statement is true about both the library and the Web

- All its resources are free and accessible to students.
- Anyone can add information to it.
- Has material aimed at all audiences, including consumers, scholars, students, hobbyists, businesses.
- Has materials which have been purchased on behalf of students.
- Information must have been deemed authoritative to be included.
- Is organized systematically with a classification scheme.
- Offers online option to ask questions.

tional useful information regarding the items. We expected that students who use the library more, who have had more library instruction, or who were more confident about their responses would have responded correctly to more items, especially more of the difficult ones, than the other students did. However, there was no clear pattern that this had occurred. We surmised that items regarding self-reported levels of knowledge, experience, etc. may not be interpreted similarly among students. For example, how often a student “uses” the library could mean different things to different students. Anecdotally, one of the university orientation classes was taught by a library science student who required her students to complete “TILT,” a web-based information literacy-intensive tutorial, in addition to the standard library instruction unit. We would have expected those students to

report a fairly high level of previous library instruction; however, perceptions among these similarly-taught students varied widely. Therefore, it was felt that the information provided by these items was not useful in examining the responses to the knowledge items.

There was little useful information provided when separating the subjects into high and low HSGPAs—responses for both groups were often similar or showed no real pattern of more correct answers by more able students. We think, however, that HSGPA is a more variable measure of ability than a nationally-standardized test score, such as the ACT provides. Therefore, the crosstabs of ACT scores with the knowledge items were examined more closely to determine if further confirmation of the item difficulty levels could be discovered. It would be expected that higher ability stu-

dents (those with ACT scores above the mean) would respond correctly more often than those student with lower ACT scores, especially on the more difficult items. Although this pattern is not consistent across all items, we did find that certain groups of items did seem to fit this pattern. Specifically, the higher-ability students were better able to identify services offered by the library. For example, while the most students knew that materials could be checked out from the library, fewer knew that librarians could help them focus their research topics. However, a greater percentage of the higher-ability than lower-ability students were aware of this kind of assistance. Additionally, higher-ability students were more able to correctly distinguish among different types of publications, such as scholarly, popular, and professional, and to distinguish among various types of intended audiences.

Next Steps

The next phase in our instrument development will include making slight revisions to the items and field testing them with a large sample. We plan to administer the instrument to 500 undergraduate students to gather more complete data on how well the items measure the construct of information literacy.

The following technical criteria will be used to determine if the variable of information literacy can be adequately scaled:

a. Is a discernible line of increasing intensity defined by the data? This can be seen by the extent to which item calibrations are spread out to define distinct levels of the information literacy variable. In addition, the variable map, plotting person ability against item difficulty, should show the items spread equally among the respondents.

b. Is item placement along this line reasonable? Items must be ordered along a line in a way which follows expectations. That is, those items which are more difficult must group together at the high end of the continuum, while those that are easier should group together at the low end of the continuum.

c. Do the items work together to define a single variable? The responses to the items should be in general agreement with the ordering of persons implied by the majority of items. This can be analyzed by examining the number of item misfits. Lack of (or few) item misfits indicate that the variable can be reasonably ordered from easy to difficult without too many persons violating this pattern.

d. Are persons adequately separated along the line defined by the items? It should be possible to separate

persons measured into distinct levels of information literacy knowledge. The measure of person separation reliability provides an indication of whether this has been achieved.

e. Do individual placements along the variable make sense? This can be judged based on other information available about the persons being tested. Since we might expect better students to be more information literate, a comparison of the position of subjects with higher HSGPAs and test scores against those with lower HSGPAs and test scores along the continuum would give an indication of the reasonableness of the placements along this line.

f. How valid is each person's measure? The responses of each person can be examined for consistency. The order of the item difficulties should be similar for all persons. If there are wide discrepancies, the validity of that person's measure is suspect. The measure used to determine the validity of each person's responses is person fit. Misfitting persons should be examined individually to attempt to determine an underlying reason for the misfit (such as carelessness, guessing, use of extreme categories).

As further validation of the instrument, we plan to ask selected subjects who have completed the instrument to individually demonstrate information literacy skills. We will have their scores on the instrument as well as other measures of overall student ability (HSGPA, ACT/SAT scores, KSU GPA), so we can compare these to how well they actually demonstrate information literacy skills.

More items need to be written and tested in order to expand the instrument. Ideally, we would like an item bank of 5–10 items per objective. This would allow larger-scale testing and more adequate pre- and post-testing procedures. Once the items are scaled using IRT many different useful tests can be developed combining items with varying ability levels.

The investigators also plan to expand the coverage of the information literacy standards by developing new items. Our current effort concentrates on only two standards. In order to assess information literacy in a more comprehensive way, all of the standards need to be covered.

Once this first major phase is completed, we will turn our attention to the other goals of our project. One key desired characteristics of this assessment instrument is that it can be used across institutions. We will seek out partners at other colleges and universities to continue testing with new populations. We would also like to convert the instrument to a web-based format, a goal that will require a high degree of programming expertise.

All these steps will help us progress toward having a reliable, valid instrument that can be used anywhere to measure the construct of information literacy. Once such an instrument is available, we can use it for longitudinal testing by selecting a cohort of students upon their entry to college and following their progress throughout their academic careers. Tying this information with institution-provided data such as college GPA and knowledge of library instruction participation may allow us to answer the questions, "Does library instruction make a difference on campus?" and, "Does library instruction lead to acquisition of information literacy skills?"

Works Cited

- American Association of School Librarians. *Information Power: Building Partnerships for Learning*. Chicago: American Library Association, 1998.
- American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education. *Standards for Educational and Psychological Testing*. Washington, D.C.: American Educational Research Association, 1999.
- American Library Association, Evaluation of Reference and Adult Services Committee. *Reference Assessment Manual*. Ann Arbor: Pierian Press, 1995. Description of the WOREP on pp. 308–10; copy of the instrument on accompanying disk.
- Association of College and Research Libraries. *Information Literacy Competency Standards for Higher Education: Standards, Performance Indicators, and Outcomes*. 2000. 1 December 2000 <<http://www.ala.org/acrl/ilstandardlo.html>>.
- Bober, Christopher, Sonia Poulin, and Luigina Vilen. "Evaluating Library Instruction in Academic Libraries: A Critical Review of the Literature, 1980–1993." *Reference Librarian* 51/52 (1995): 53–71.
- Catts, Ralph. "Some Issues in Assessing Information Literacy." *Information Literacy Around the World: Advances in Programs and Research*. Eds. Christine Bruce and Philip Candy. Wagga Wagga, New South Wales: Centre for Information Studies, 2000.
- Colborn, Nancy Wootton and Roseanne M. Cordell. "Moving From Subjective to Objectives Assessments of Your Instruction Program." *Reference Services Review* 26:3–4 (1998): 125–37.
- Coupe, Jill. "Undergraduate Library Skills: Two Surveys at Johns Hopkins University." *Research Strategies* 11:4 (1993): 188–201.
- Crocker, Linda and James Algina. *Introduction to Classical and Modern Test Theory*. New York: Holt, Rinehart, and Winston, 1986.
- Franklin, Godfrey and Ronald C. Toifel. "The Effects of BI on Library Knowledge and Skills Among Education Students." *Research Strategies* 12:4 (1994): 224–37.
- Fry, Thomas K. and Joan Kaplowitz. "The English 3 Library Instruction Program at UCLA: A Follow-Up Study." *Research Strategies*. 6:3 (1988): 100–8.
- Hardesty, Larry, Nicholas P. Lovrich, Jr., and James Mannon. "Evaluating Library-Use Instruction." *College and Research Libraries* 40 (1979): 309–17.
- Hardesty, Larry, Nicholas P. Lovrich, Jr., and James Mannon. "Library-Use Instruction: Assessment of Long-Term Effects." *College and Research Libraries* 43 (1982): 38–46.
- Kaplowitz, Joan. "A Pre- and Post- Test Evaluation of the English 3-Library Instruction Program at UCLA." *Research Strategies* 4 (1986): 11–17.
- Kuhn, Terry. "Concordance Table of SAT to ACT Composite Scores." Unpublished. Kent State University, June 1995.
- Kunkel, Lilith R., Susan M. Weaver, and Kim N. Cook. "What Do They Know?: An Assessment of Undergraduate Library Skills." *Journal of Academic Librarianship*. 22 (1996): 430–34.
- Lawson, Mollie D. "Assessment of a College Freshman Course in Information Resources." *Library Review* 48:2 (1999): 73–78.
- Lindauer, Bonnie G. "Defining and Measuring the Library's Impact on Campuswide Outcomes." *College and Research Libraries* 6 (1998): 546–70.
- Menshing, Teresa B. "Trends in Bibliographic Instruction in the 1980s: A Comparison of Data From Two Surveys." *Research Strategies* 7:1 (1989): 4–13.
- Pausch, Lois M. and Mary Pagliero Popp. (1997) "Assessment of Information Literacy: Lessons from the Higher Education Assessment Movement" 1997 December 1, 2000 <<http://www.ala.org/acrl/paperhtm/d30.html>>.
- Person, Roland. "Long-Term Evaluation of Bibliographic Instruction: Lasting Encouragement." *College and Research Libraries* 42 (1981): 19–25.
- Rabine, Julie and Catherine Cardwell. "Start Making Sense: Practical Approaches to Outcomes Assessment for Libraries." *Research Strategies*. In press.
- Ragains, Patrick. "Evaluation of Academic Librarians' Instructional Performance: Report of a National Survey." *Research Strategies* 15:3 (1997): 159–75.

Selegean, John Cornell, Martha Lou Thomas, and Marie Louise Richman. "Long-Range Effectiveness of Library Use Instruction." *College and Research Libraries* 44 (1983): 476-80.

Tiefel, Virginia. "Evaluating a Library User Education Program: A Decade of Experience." *College and Research Libraries* 50 (1989): 249-59.

Wright, Benjamin D. and Mark H. Stone. *Best test design*. Chicago: Mesa Press, 1979.