Applying Systems Design and Item Response Theory to the Problem of Measuring Information Literacy Skills

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

This article reports on a project to develop an instrument for programmatic-level assessment of information literacy skills that is valid—and thus credible—to university administrators and other academic personnel. Using a systems approach for test development and an item response theory for data analysis, researchers have undertaken a rigorous and replicable process. Once validated, this instrument will be administered to students to assess entry skills upon admission to the university and longitudinally to ascertain whether there is significant change in skill levels from admission to graduation.

A biblical parable on the virtue of tenacity tells of a widow who repeatedly beseeches a judge to grant her request. Finally, the judge, although not sympathetic to her cause, grants her request lest she eventually exhaust [him] with her coming. In the golden age of higher education, when expansion was rapid and funding abundant, the widow's technique may have been effective. In the current era of finite resources and increased fiscal accountability, however, when libraries plead their cases for resources to support their information literacy programs, persistence alone does not suffice.

Purpose

Are libraries able to provide evidence that information literacy skills affect student learning and success? A thorough search of the library literature reveals that our profession is not yet in a position to agree on the best method for assessing those skills, let alone assert that they make a difference. The purpose of the Project for the Standardized Assessment of Information Literacy Skills (SAILS) is to develop an instrument for programmatic-level assessment of information literacy skills that is valid—and thus credible—to university administrators and other academic personnel. Once validated, this instrument will be administered to students to assess entry skills upon admission to the university and longitudinally to ascertain whether there is significant change in skill levels from admission to graduation. After information literacy skills have been measured and any

Lisa G. O'Connor is Instructional Services Coordinator at Kent State University; e-mail: loconnor@lms.kent.edu. Carolyn J. Radcliff is Head of Reference Services at Kent State University; e-mail: radcliff@kent.edu. Julie A. Gedeon is Manager of Academic Technology Services Evaluation at Kent State University; e-mail: jgedeon@kent.edu.

changes in skill levels over time identified, whether those skills have any relationship to students' academic success and retention must be determined.

The authors of this study were inspired by the Wisconsin Ohio Reference Evaluation Project (WOREP), a tool for evaluating reference services. The attributes of WOREP that were appealing were that it is standardized, contains items not specific to a particular institution or library, is easily administered, has been proven valid and reliable, assesses at an institutional level, and provides for both external and internal benchmarking. These laudable characteristics are ones the authors sought to emulate as they worked toward creating an instrument for measuring information literacy skills.

The authors of this article believe that in order to measure information literacy as a campuswide learning outcome, a more rigorous process must be demonstrated than currently exists.

Literature Review

The library literature on assessment practice for the past twenty years demonstrates little experience in formalized evaluation in general and does not contain or make reference to an instrument that is suitable for standardized, longitudinal, and cross-institutionally administered assessment. "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."[1] According to Teresa B. Mensching's 1987 survey of LOEX-participating libraries, only 23 percent of respondents who evaluated BI were using an assessment mechanism.[2] As Jill Coupe summed up, "Perhaps one reason that librarians have neglected the measurement of basic library skills is the lack of an adequate survey instrument."[3]

Another characteristic of current 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. Thus, instruments are developed quickly and the gathered data, not the development process, are the main focus of the research reports. According to Bonnie G. Lindauer, "almost none of these publications provide 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."[4] Lois M. Pausch and Mary Pagliero 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."[5] The authors of this article believe that in order to measure information literacy as a campuswide learning outcome, a more rigorous process must be demonstrated than currently exists.

Having failed to locate an instrument that could be used to assess the information literacy skills of students longitudinally and across institutions, literature review was performed for articles on the process of developing an instrument to measure this construct. What follows is a summary of the most significant models. For a more thorough literature review, refer to the authors' paper on the initial phase of this project.[6]

Eight articles reported creating a "paper-and-pencil" test to assess information literacy skills.[7] With the exception of Lilith R. Kunkel, Susan M. Weaver, and Kim N. Cook, studies included all or parts of their instruments. All the tests included questions on basic library skills, such as OPAC usage, call number comprehension, basic search construction, Boolean operators, citation interpretation, and locations of various services or resources within the libraries. Most also included items to assess library-related attitudes and behaviors, to allow for student self-assessment of skills, and to gather basic demographic information. Instruments contained between nine and twenty-eight 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 the studies used pre- and posttesting; three of the pre- and posttests were identical instruments. All of the instruments contained questions specific to the researchers' libraries, and three of them were subject specific. Most of the instruments were administered within two to four weeks of instruction, with the exception of a follow-up study by Thomas K. Fry and Joan Kaplowitz.[8]

The most common formal data analysis employed in these studies was Classical Test Theory (CTT). CTT is a measurement model based on information provided at the test score level, which assumes that a determination about examinees may be made based on total test score. It is appropriately used with fixed-length tests with the same set of items administered to all respondents. CTT is best suited for traditional testing situations in which all members of the target population are administered the same or parallel sets of test items (e.g., classroom testing). This model is acceptable if the test takers are homogeneous on the trait being measured. Three of these studies were of particular interest to this project.

The first of these tests was Virginia Tiefel's report describing a 1986 project at Ohio State University. The purpose of this project was to develop an instrument that could be used to measure the effectiveness of library instruction. The resulting test was administered to 1,702 students two weeks after they received instruction. Item-by-item analysis for validity occurred only after large-scale testing had revealed significant weaknesses in the instrument. Tiefel asserted that, despite its flaws, the results of her study suggested 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."[9] No follow-up assessment was reported.

Nancy Wootton Colborn and Roseanne M. Cordell developed an instrument to measure knowledge in five fundamental areas. Their test was administered prior to and after library instruction. Of all the studies examined, this one detailed the most rigorous development process, in which the authors used both a difficulty and a discrimination index to examine all items and revised their instrument accordingly. Data from 129 students showed no significant difference between pre- and posttest results. No explanations were given for the disappointing results. The authors asserted that the test itself was not the weak link; however, inadequate data were collected to rule that out. Their experience shows how difficult and unpredictable the test development process can be.

Larry Hardesty, Nicholas P. Lovrich Jr., and James Mannon reported on the development of an instrument containing twenty-six items to test library use skills and ten attitudinal items. The instrument was administered to 162 freshmen prior to instruction and to the same group eight weeks after instruction. Skill items were considered reliable if more than 50 percent and less than 90 percent of respondents answered items correctly. Results indicated a 35 percent increase in correct responses from the first to the second administration. Researchers used a control group in the testing phase to ensure the legitimacy of any significant differences discovered in their study. This research project, which was later replicated by the same authors, achieved 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."[10]

The body of literature on library instruction assessment reveals weaknesses that the authors of this article were determined to avoid. It also affirmed the importance of many of their original project goals. For example, it was clear that the very narrow skill sets tested by most of the instruments did not measure the full range of the information literacy trait. Also, a thorough trial process with small, easily assembled groups will identify problem items prior to large group trials, saving time and effort. Instrument development should be undertaken with thorough consideration for how assessment will ultimately be administered. Because the aim of this project was to conduct programmatic-level assessment, participant samples had to be large enough to enable results to be generalized to the institutional population; therefore, the instrument had to be easily administered and scored. Although this requirement indicates assessment that is not authentic (that is, assessment requiring learners to actually perform tasks associated with learning outcomes), it is nonetheless vital to the project's long-term goals. In addition, assessing long-term skill acquisition is ultimately more valuable than measuring short-term gains; thus, longitudinal testing is imperative. If identical pre- and posttests are used, the effect the test-taking experience has on results must be evaluated. The implication for the authors' project was that a large test bank should be available so that different items testing the same outcomes can be generated randomly to accommodate repeated testing on the same samples without test experience effects. Because measuring information literacy alone does not address its relevance to the university's mission, data must be gathered and analyzed in such a way as to assess the effect of information literacy on student success and retention. Finally, the process of instrument development and testing should be reported at the level of detail needed to replicate the study. The rigor should be evident to anyone who questions its validity. As Ralph Catts wrote, "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."[11]

Methodology

The developmental phase of this project was based on a systems model popularized by Walter Dick and Lou Carey.[12] This model theorizes that instruction is a systematic process in which every component, including assessment, is crucial for learning to occur. It asserts that legitimate assessment is inexorably linked to instructional goals and performance outcomes. The systems process is thorough and time-consuming, so it is best suited to programs of instruction rather than to individual instructional sessions.

There are five phases of systematic instructional design that lead to effective instruction and assessment. The first phase is to determine what, specifically, the researcher wants learners to be able to do as the result of instruction. The second phase is to analyze the instructional goal, which means describing precisely the behaviors that learners will engage in when they have achieved that goal. Fortunately, when the authors first began analyzing information literacy in 1998, the first two of these phases had been essentially accomplished with publication of the nine standards of student learning from the Association of School Librarians.[13] These standards and their corresponding performance indicators delineated information literacy and clearly defined its goals.

The third phase is to analyze learners and contexts to gain a more complete understanding of the learning environment. The authors' analysis was based on national trends in college student expectations and experiences, data provided by Kent State University's (KSU) Office of Institutional Research (e.g., graduation rates, ACT scores, high school grade point averages, etc.) and personal experience in working with students over the years. This phase is an opportunity to shift away from a narrow focus on the specific task and to think more broadly about the population to be studied.

After analyzing learners and contexts, the fourth phase is to write performance objectives. These objectives are specific statements that identify and describe skills to be attained, conditions under which they must be performed, and criteria for successful performance. Because the objec-

FIGURE 1
Example of Item Development

ACRL Standard Two:
The information-literate student accesses needed information effectively and efficiently.

Performance Indicator 1:
The information-literate student selects the most appropriate investigative methods or information retrieval systems for accessing the needed information.

Outcome:
Selects appropriate tools (e.g., indexes, online databases) for research on a particular topic

ITEM:
If you are required to write a paper on teenage pregnancy, which of the following types of databases might have articles on this topic?
□ architecture database
□ education database
□ health database
□ mathematics database
□ physics database
□ psychology database

tives must be written for every skill and subskill required to meet the overall instructional goal, this phase can be both tedious and time-consuming. In fact, it took a KSU team of four instructional services librarians several months of concentrated work to complete this phase. For these reasons, this is the phase researchers may be most likely to skip or rush through. If this phase is done well, the next phase can be accomplished more efficiently and effectively. Although the writing of objectives was completed in 1999 by KSU's Instructional Services Team, the objectives have since been replaced with the model learning outcomes written and adopted by ACRL in order to achieve consistency with other institutions around the country.

Only after following the first four steps carefully were the authors prepared to begin instrument development. In this phase, items flow naturally, although not easily, from objectives. Writing items is only a question of determining how the learner's ability to perform in the manner already described can be measured. Figure 1 shows an item that was developed based on the original instructional goal and learning outcome. It is important to note that this is not the only item developed for that performance indicator or even for the specific outcome.

When items have been developed, several iterations of evaluation are necessary to ensure that the items function as they were designed to function. Items are put to the test in a series of three trial phases. In the first trial phase, the items are tested in one-on-one trials. The purpose of this phase is to identify and remove the most obvious errors and is accomplished through direct interaction between designers and individual learners. Dick and Cary recommended three or more learners drawn from the target population. This study used six learners. In this phase, learners engage in in-depth communication and analysis with designers as they complete the items. Researchers try to determine, among other things, what is clear, what is unclear, how learners interpret questions, and why learners select specific responses. Items are revised on the basis of data gathered from this phase.

In the next phase, items are tested in small group trials. These trials are an extension of one-on-one trials and provide opportunities for further revision. In this

FIGURE 2
Example of Content Changes Resulting from Trial Process
(Changes indicated by highlighted text)

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

REVISION 1 (After one-on-one trials)
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 one 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 that have been purchased on behalf of students.
____Information must be deemed authoritative to be included.
____Is organized systematically with a classification scheme.
____Offers online option to ask questions.

project, items were administered to a class of twenty students in a manner that approximated a normal test-taking situation, except that learners were asked to make notes of questions, problems, and thoughts about items. When all the instruments were completed and returned, the authors engaged in a dialog with students. Students' interpretations of questions and their answers enabled the authors to make substantial improvements to the instrument.

Finally, the items were tested in field trials. Field trials most closely emulate the intended context for instrument administration. Designers become observers only and do not interact with learners. Feedback for revision is taken exclusively from the data gathered. In the authors' spring 2001 study, the instrument was administered to 554 students in this phase of the project. Figure 2 provides an example of how one item changed throughout the process.

FIGURE 2 (CONTINUED)
Example of Content Changes Resulting from Trial Process
(Changes indicated by highlighted text)

REVISION 2 (After small group trials)

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 one 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 academic library.
Put a **B** if the statement is true about both the academic 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 shoppers, support groups, scholars, students, hobbyists, businesses.
_____Has materials that 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.

REVISION 3 (After field trials)

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 one 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 academic library.
Put a **B** if the statement is true about both the academic library and the Web.
Put a **N** if the statement is **not** true about either the academic library or the Web.

_____All its resources are free and accessible to students.
_____Anyone can add information to it.
_____Targets all audiences, including shoppers, support groups, scholars, students, hobbyists, businesses.
_____Has materials that 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.

A systems approach has worked especially well for this project because it is an integrative and reiterative process. When items are developed to measure specific behaviors, they are easier to write and more authentic. Because the work is based on thoroughly analyzed goals and objectives, the process is empirical and easily replicable. Designers who follow the same procedures would theoretically produce a similar instrument. The systematic process also facilitates a high degree of collaboration between content and measurement experts. Because the systematic approach was created particularly for programmatic-level instructional design, it also works particularly well for programmatic-level assessment.

Participants in the Field Trials

Participants in this phase of the project were undergraduates enrolled during the spring semester 2001 at KSU. Data were collected from freshmen orientation, nursing, and journalism and mass communications classes. The classes were selected based on several factors, including number of students enrolled and faculty willingness to allow class time for participation. Respondents ranged from freshmen to seniors.

The authors examined results to determine whether the latent trait of information literacy could be measured adequately based on six criteria outlined by Benjamin D. Wright and Mark H. Stone.[14]

Participants were given a consent letter detailing the purpose of the study, which they were required to sign before participating. Rather than self-reporting demographic and academic data such as grade point average (GPA), major, and class, the students were asked to provide their student ID numbers so that more accurate information could be gathered. The authors remained in the room while the students completed the instrument and collected them immediately thereafter.

Of the 554 instruments administered, 537 were completed. These were used for the item analysis described below. Three hundred and ninety-eight students provided valid ID numbers, which were used to obtain demographic and academic data.

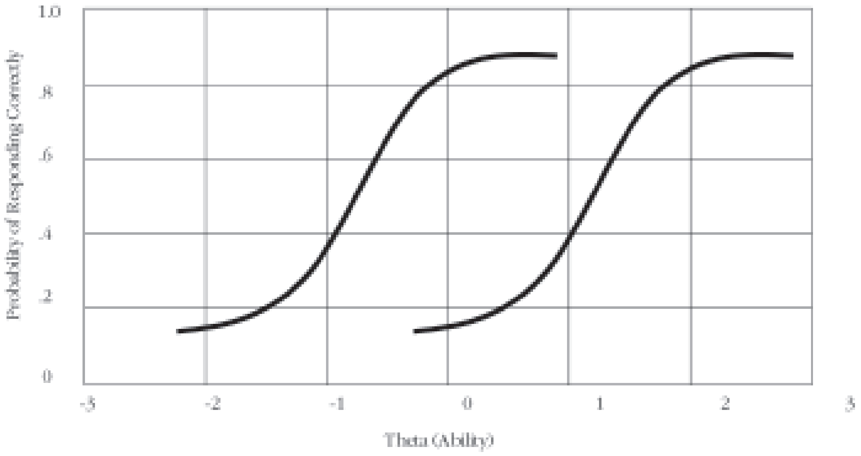
Measurement Model

The authors opted to use item response theory (IRT) as the measurement model for this project. IRT, also known as latent trait theory, focuses on latent, unobservable traits such as knowledge or ability level. In this project, the latent trait is information literacy ability. To measure this ability, the authors devised a set of questions of varying difficulty levels. Difficulty level of the questions was verified by presenting the items to content experts (experienced reference and instruction librarians) and having them rate each item as easy, medium, or difficult. This approach gives the ability to differentiate between people who get easy items correct and those who also get difficult items correct. The analysis of responses is based on a mathematical model of the probability of how people at different ability levels respond to an item.

Plotting the probability of a correct response against a continuum of ability levels results in the item characteristic curve (ICC), a key construct of IRT. Each item will have its own ICC. Figure 3 shows an example of two ICCs representing two items of varying difficulty level with probability on the vertical axis and theta (ability) on the horizontal axis. The curve on the left represents an easier item because it appears at the lower end of the ability continuum. Moving higher on this continuum, respondents with higher levels of ability have a higher probability of responding correctly to this item. For the more difficult item, the one on the right, lower-ability respondents have a lower probability of responding correctly than do higher-ability respondents.

There are different models of item response theory, based on the number of parameters to be estimated for the items. The authors used a one-parameter model, the Rasch rating scale. In this one-param-

FIGURE 3
Item Characteristic Curves



eter model, the item characteristic curves for all items on the instrument vary only in their location along the ability continuum. The Rasch model is represented by P_{ni} (x = 1) = f (B_n - D_i), which means that the probability of person n responding correctly to item i is a function of the difference between his or her ability (B_n) and the item's difficulty (D_i) (x is any given score and 1 is a correct response).

The Rasch model uses a logistic function to describe the interaction between items and persons. A logit is the natural logarithm of the odds of an event's occurring. The logit transformation linearizes the inherent nonlinear relationship between the items and the persons. The distribution of logits is symmetric around the midpoint probability of 0.5 (logit = 0) and ranges from negative infinity to positive infinity.

In the Rasch model, the items and the respondents are calibrated along the same continuum. Items that are more difficult to answer are calibrated toward the high end of the continuum, and items that are easier to answer appear at the low end. Respondents are placed along the same continuum in a similar manner—those who are able to respond correctly to more items are placed at the higher end of the continuum, and those who have less success responding appear at the low end.

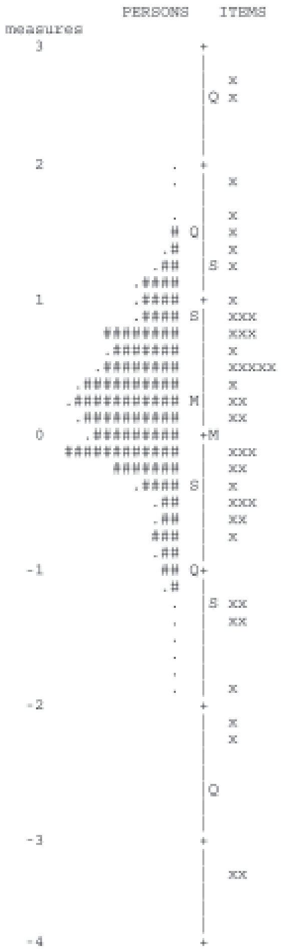
The use of item response theory, and specifically the Rasch model, guides the data analysis. The following section discusses what the data say about the items developed and whether they accurately measure the information literacy trait.

Data Analysis

Data from the field trials described earlier were analyzed using WINSTEPS, a Rasch modeling program created by researchers at the Mesa Institute at the University of Chicago. The authors examined results to determine whether the latent trait of information literacy could be measured adequately based on six criteria outlined by Benjamin D. Wright and Mark H. Stone.[14]

1. *Is a discernible line of increasing intensity defined by the data?* The project authors addressed this question by looking at the extent to which item calibrations were spread out to define distinct levels of information literacy skill. One criterion was that separation of item difficulties should be between -3 and +3 on the logit scale to be adequate. In addition, the item-person map, plotting person ability against item difficulty, should show the

FIGURE 4
Map of Persons and Items



```
                      PERSONS   ITEMS
measures
   3                        +
                            |
                            |  x
                            |Q x
                            |
                            |
                            |
   2                    .   +
                        .   |  x
                            |
                        .   |  x
                        # Q|  x
                       .#   |  x
                      .##   |S x
                    .####   |
   1                .####   +  x
                    .#### S|  xxx
                 ########   |  xxx
                 .#######   |  x
                .########   |  xxxxx
              .##########   |  x
             .########### M|  xx
              .##########   |  xx
   0           .#########   +M
             ############   |  xxx
                  #######   |  xx
                    .#### S|  x
                      .##   |  xxx
                      .##   |  xx
                      ###   |  x
                      .##   |
  -1                   ## Q+
                       .#   |
                        .   |S xx
                        .   |  xx
                        .   |
                        .   |
                        .   |
                        .   |  x
  -2                        +
                            |  x
                            |  x
                            |
                            |
                            |Q
                            |
                            |
  -3                        +
                            |
                            |  xx
                            |
                            |
                            |
  -4                        +
```

Note: M = mean person ability or item difficulty, S = one standard deviation, Q = two standard deviations

items spread equally among the respondents.

Figure 4, which is the map of persons and items created by the WINSTEPS program, shows the latent trait of information literacy depicted by the vertical line. The distribution of persons is on the left of the line, and the distribution of items is on the right. Better-able persons and more-difficult items appear toward the top of the map. Persons and items are scaled in logits, with a mean of 0 and a standard deviation of 1.

Examination of the map shows item calibrations evenly spread along the latent trait, ranging from -3.29 to +2.86. Although there are some small gaps, for the most part, items cover the range of difficulty levels. The distribution of persons along the variable is approximately normal, and persons and items are targeted, that is, lined up with each other along the variable. Moreover, the persons are bounded by the items, meaning that some items are more difficult than the highest-ability level and some are easier than the lowest-ability level.

2. Is item placement along this line reasonable? A satisfactory response to this question calls for items to be ordered in a way that follows expectations. Those items measuring higher-level information literacy skills must group together at the high end of the continuum, and those measuring lower-level skills should group together at the low end of the continuum.

Upon examination of the individual items and comparison of their placement along the variable with their own intent, the content experts' input, and the previous phases of data collection, the authors found that, for the most part, the ordering from easy to difficult makes sense based on content area and knowledge level needed to respond correctly. For the few items that seemed incorrectly scaled, the authors looked at the questions themselves: the wording, the response options, and even the items' placement within the instrument.

For example, the item calibrated as the easiest on the item-person map was item 7, which asked students how to search for items written by Charlotte Brontë. This item proved easy in the earlier phases of instrument development, and most people would agree that this is a fairly easy concept for students to understand. Therefore, the authors were quite confident that this item was calibrated correctly. The next easiest item was item 13, which was one of a series of items on the theme of how a student starts on a class assignment requiring library research. When this item was reexamined after an analysis of the data, the authors realized it was really an opinion-type item allowing three correct options (which may explain why it was calibrated as easy): the item itself and the way it was written and scored may not be adequate to test the skill.

At the other extreme, looking at the most difficult items, it was found that item 18 calibrated as the most difficult. This item asked students to select which databases they would search to find information on teenage pregnancy. It was scored such that respondents had to mark all three options to get credit for responding correctly. Again, upon careful examination, it was noticed that the item itself was not extremely difficult to the content experts or the respondents in the earlier phases of data collection, so perhaps the way it was scored made it more difficult. It might be better, and more accurate, to award partial credit for each correct option selected.

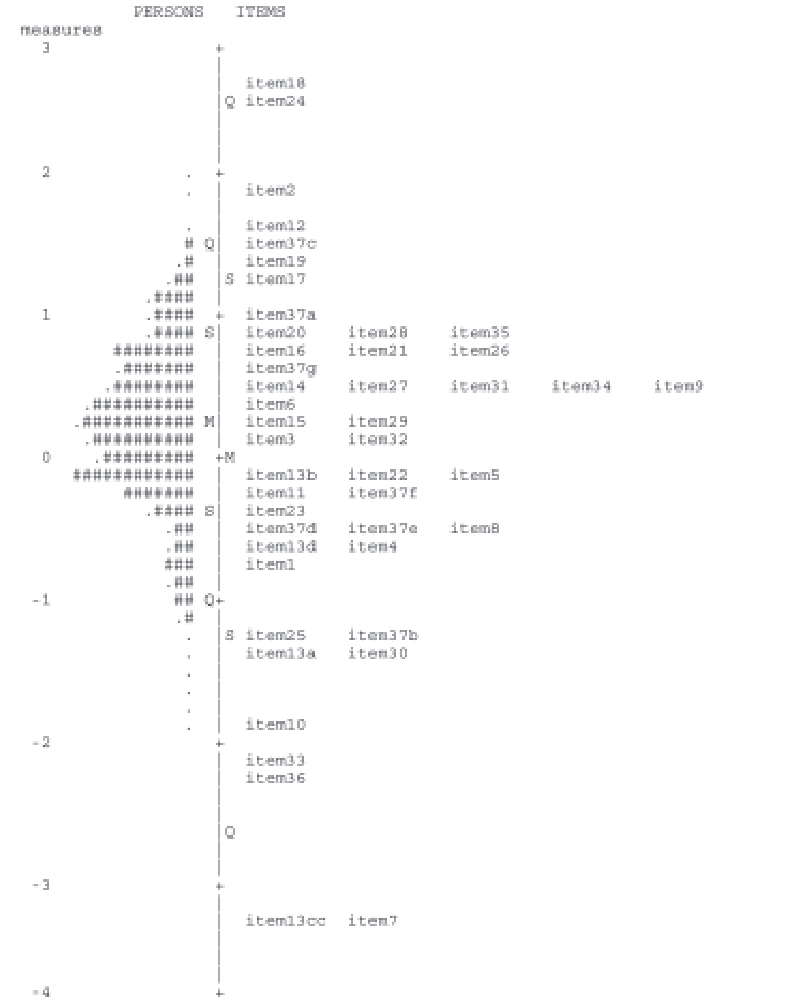
The next most-difficult item was item 24, which asked students to select the optimum search strategy to locate information on the use of color in the famous painting *The Madonna.* All the content experts and previous phases of data collection provided evidence that this item was quite difficult for the respondent population, so the authors were confident that this item was calibrated correctly.

This is the process the authors followed when examining each of the items and their placement along the continuum, and some decisions were made for revisions based on these deliberations.

3. 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. In other words, people with higher-ability level should have answered most items correctly, particularly the easier items; and people with less ability should have answered most easy items correctly but missed more of the difficult items. This can be analyzed by examining "item fit," in particular, the number of item misfits. A lack of (or few) item misfits indicates that the variable can be reasonably ordered from easy to difficult without too many persons violating this pattern.

Item fit is measured in two ways in the Rasch model. The first is referred to as infit, which indicates how well an item works for persons close to it in ability level. The second way to look at item fit is outfit, which indicates how well an item works for persons far from it in ability level. Figure 5 illustrates the concepts of infit and outfit. Looking at item 3 on the map, one would expect that the people near that item would have a 50 percent

FIGURE 5
Map of Persons and Items with Item Names



```
           PERSONS   ITEMS
measures
  3                 +
                    |
                    |  item18
                    |Q item24
                    |
                    |
                    |
  2              .  +
                 .  |  item2
                    |
                 .  |  item12
                 # Q|  item37c
                .#  |  item19
               .##  |S item17
             .####  |
  1          .####  +  item37a
             .#### S|  item20    item28    item35
          ########  |  item16    item21    item26
          .#######  |  item37g
         .########  |  item14    item27    item31    item34    item9
       .##########  |  item6
      .########### M|  item15    item29
       .##########  |  item3     item32
  0     .#########  +M
      ############  |  item13b   item22    item5
           #######  |  item11    item37f
             .#### S|  item23
               .##  |  item37d   item37e   item8
               .##  |  item13d   item4
               ###  |  item1
               .##  |
 -1             ## Q+
                .#  |
                 .  |S item25    item37b
                 .  |  item13a   item30
                 .  |
                 .  |
                 .  |
                 .  |  item10
 -2                 +
                    |  item33
                    |  item36
                    |
                    |
                    |Q
                    |
                    |
 -3                 +
                    |
                    |  item13cc  item7
                    |
                    |
                    |
 -4                 +
```

Note: M = mean person ability or item difficulty, S = one standard deviation, Q = two standard deviations

probability of responding correctly to it (infit). If, in fact, everyone responds correctly or everyone misses this item, the item is not working as expected. When considering the people at the top of the ability scale in relation to item 3, one would expect them to answer correctly (outfit). If they all miss this item, it provides information that the item does not work for higher-ability people.

The pilot instrument had four misfitting items (9% misfits) out of the forty-six included. It is important to examine the misfits individually to determine why they are occurring. The authors considered the wording of the items and the response options to determine whether there was anything tricky or misleading. The frequencies of responses across options also were examined to determine whether some of the incorrect options were too attractive to respondents. An example of this was item 16, which asked respondents why searching a database using the term "skin cancer" would retrieve an article with the word "melanoma" in the title. One of the incorrect response options was "relevancy ranking," which proved to be very attractive, even to higher-ability subjects, throughout each phase of instrument development.

The next three criteria dealt more closely with the adequate measurement of persons using the information literacy pilot instrument. The authors are doing additional work to analyze the results in relation to the person measures; this effort will guide further refinement of the instrument.

4. *Are persons adequately separated along the line defined by the items?* It should be possible to separate persons measured into distinct levels of ability in information literacy. The measure of person separation reliability will provide an indication of whether this has been achieved. Person separation reliability is an estimate of how well persons can be differentiated on information literacy. Its values are analogous to Cronbach's alpha, with values ranging from 0 to 1.

The person separation reliability for this pilot data collection was .64, indicating a moderate level of confidence in the ability to separate students into levels of information literacy.

5. *Do individual placements along the variable make sense?* This can be judged based on other information available about the persons being tested. A comparison of the position of subjects with higher ACT scores or higher KSU cumulative GPAs against those with lower ACT scores or GPAs would give an indication of the reasonableness of the placements along this line. ACT scores generally are more attractive for this purpose because of the standardized nature of the scores. GPAs are more variable depending on the level of courses and subject matter.

This question was addressed by looking at the 398 respondents who provided valid student ID numbers, allowing the authors to obtain additional data from the university's student information system. Correlations were computed between the Rasch person measures and ACT scores and the Rasch person measures and KSU cumulative GPAs. A small correlation (.263) was found between the Rasch measures and the KSU GPAs. A moderate correlation (.482) was found between the Rasch measures and the ACT scores. This moderate correlation provides some indication that the information literacy instrument is adequately measuring students with different ability levels.

6. *How valid is each person's measure?* Each person's responses can be examined for consistency. The order of the item difficulties should be similar for everyone. If there are wide discrepancies, the validity of that person's measure is suspect. The measure used to determine this is person fit. Person misfit is determined by infit and outfit, similarly to item misfit.

There were 44 misfitting persons out of 537 (8.2%). Misfitting persons' responses were examined individually to determine why the instrument was not working for them. Several aspects of their responses were considered, including any patterns (selecting the first option on the first page, the second option on the second page, etc.), use of extreme categories (marking all ones or sixes), the number of items to which the individual responded (if less than half were answered, there may not be enough information to accurately measure the person), and specific items causing problems for many of the misfitting persons. There were no apparent patterns to the responses of misfitting persons.

The data analysis centered on addressing the six accepted criteria for adequate measurement of information literacy using the Rasch model of item response theory. The analysis showed that most items developed to date were reliable and valid and that the items worked together to measure at least some portion of the trait of information literacy. After items have been developed and validated for all learning outcomes, the authors will be able to assert that they can measure the trait to the extent that it can be measured by a standardized multiple-choice test format. Similar analyses will be conducted as new items are developed and during new rounds of data collection through field trials.

Current Phase of Project

As of the time of this writing (March 2002) and since the last round of trials, the authors have revised the problem items. They also have created forty-six additional items, covering approximately 70 percent of the unique learning outcomes deemed testable. Both the revised and the newly created items have been tested in one-on-one and small group trials. The instrument has been converted from a paper-and-pencil format to a Web-based format. The Web format offers advantages over paper, primarily in terms of data collection. Web responses going directly into a database cut out the tedious and expensive step of data entry and substantially reduce the potential for errors. It also is possible to work with instructors to have students complete the questionnaire outside class time without the additional time and staffing demands of getting the proper forms to the students, collecting the completed responses, and meeting other administrative challenges. One significant disadvantage, however, is that networked computer workstations must be available to students completing the instrument, a requirement not met by all venues. The authors' previous work has entailed going into classrooms that do not have computers. On balance, however, the Web-based format has proved preferable to the paper format for the purposes of this study.

Both old and new items are now being retested in new iterations of the instrument to ensure that the new versions are in fact measuring accurately the skills they are attempting to measure. There are multiple versions of the instrument in which item order varies so that any uneven effects of test fatigue or distraction will be reduced or eliminated. The authors have administered the instrument to a group of freshmen from university orientation at KSU and are beginning to analyze the results. In addition, they have identified other institutions that are willing to participate in the next phases of the study. Items were written to be institution neutral by avoiding reference to the name of the library catalog or locations specific to the KSU libraries. However, the best test of whether the project has been successful is to have students at other institutions complete the instrument and then to compare their results with those of KSU students with appropriate controls. Pilot testing with other colleges and universities has begun and will run through the fall of 2002. The authors have selected a variety of institution types, including small private colleges, community colleges, other universities, and institutions with more heterogeneous populations.

Future Phases

The next step is to develop and test items that correspond to the ACRL behavioral objectives that have not yet been addressed and are appropriate for an undergraduate learner. Ultimately, before administering the test longitudinally, the authors will develop a test bank of items that permits repeat testing of students while avoiding the problem of test-taker familiarity with the test items. Criticism of pre- and posttesting is often prompted (justifiably) when the pretest items are the same as the posttest items. Developing several items that measure the same construct allows for acceptable substitution of one of those items for another.

The authors will further investigate the validity of the instrument by interviewing and administering performance tests to both high- and low-scoring students from the field trial phase. That is, students will be asked to perform a task deemed to be based on knowledge represented by an item. For example, one item asks, "To find material about the poet Maya Angelou, which type of search is most effective?" The possible answers are author, subject, and title; the correct answer is subject. During the performance interviews, students will be asked to use the library catalog to find materials about a person. Their actions will be observed, anticipating that students who scored high on the instrument will use the "subject" option more often than those who scored low on the instrument. This phase of the project has been funded by a research grant from the Academic Library Association of Ohio.

The Web-based format will be expanded in the future to allow for actual performance of activities demonstrating information literacy skills. An interactive module with a simulated database will present a test of students' skill and knowledge that is more realistic than the multiple-choice questions currently being used. For example, students will have the opportunity to demonstrate the correct use of Boolean operators by searching a limited database. Performance will replace the terminology recognition that limits the present items.

Information literacy competencies are both broad in scope and thorough in depth. Currently, it is nearly impossible to measure the entire trait of information literacy without subjecting respondents to an extremely long test. A working group of librarians from ARL universities and Ohio institutions will convene in April 2002 to cluster outcomes into skill sets and difficulty levels. This activity will not only allow the authors to report test results in a way that is most useful for librarians to identify problem areas, but it also will prepare the authors for the final phase of the initial development process in which the Web-based test will be converted to a computer adaptive format. This format will permit testing of the information literacy trait in an efficient manner. The authors are currently seeking funding for this activity.

Conclusion

Project SAILS has revealed many things about assessment. First, very early on, the authors realized the need to involve an expert in measurement and evaluation. They identified a local expert who also happened to have a solid background in library science and brought that person onto the team. This partnering was essential to the success of the project.

The authors also learned that relying on an established method of development, the systematic instructional design, was very beneficial because items could be developed that performed well in the data analysis phase. Without that process, much effort could have easily been spent on large-scale testing only to find that the students taking the test did not fully understand the items. The one-on-one and small group testing helped avoid that problem.

One painful lesson has been the need to carefully document all steps—the authors were occasionally forced to reconstruct actions taken or decisions made, a time-consuming and frustrating process. Finally, the authors realize that the tremendous effort needed to create a meticulously tested standardized tool is well worth it. Thus far, responses to the work have been overwhelmingly positive and encouraging. If Project SAILS continues to be successful in its development and implementation, perhaps it can offer a viable solution for a number of libraries without the resources to engage in this level of instrument development. For further information on Project SAILS and an update on its progress, visit the authors' Web site at www.library.kent.edu/sails.

Notes

1. Christopher Bober, Sonia Poulin, and Luigina Vileno, "Evaluating Library Instruction in Academic Libraries: A Critical Review of the Literature, 1980–1993," *Reference Librarian* 51/52 (1995): 53–71.

2. Teresa B. Mensching, "Trends in Bibliographic Instruction in the 1980s: A Comparison of Data from Two Surveys," *Research Strategies* 7 (winter 1989): 4–13.

3. Jill Coupe, "Undergraduate Library Skills: Two Surveys at Johns Hopkins University," *Research Strategies* 11 (fall 1993): 188–201.

4. Bonnie G Lindauer, "Defining and Measuring the Library's Impact on Campuswide Outcomes," *College and Research Libraries* 6 (Nov. 1998): 546–70.

5. Lois M Pausch and Mary Pagliero Popp, "Assessment of Information Literacy: Lessons from the Higher Education Assessment Movement," Association of College and Research Libraries [cited 14 March 2002]. Available online from http://www.ala.org/acrl/paperhtm/d30.html.

6. Lisa G. O'Connor, Carolyn J. Radcliff, and Julie A. Gedeon, "Assessing Information Literacy Skills: Developing a Standardized Instrument for Institutional and Longitudinal Measurement," in *Crossing the Divide: Proceedings of the Tenth National Conference of the Association of College and Research Libraries* (Chicago: ACRL 2001).

7. Larry Hardesty, Nicholas P. Lovrich Jr., and James Mannon, "Evaluating Library-Use Instruction," *College and Research Libraries* 40 (1979): 309–17; Joan Kaplowitz, "A Pre- and Post-Test Evaluation of the English 3-Library Instruction Program at UCLA," *Research Strategies* 4 (winter 1986): 11–17; Virginia Tiefel, "Evaluating a Library User Education Program: A Decade of Experience," *College and Research Libraries* 50 (Mar. 1989): 249–59; Jill Coupe, "Undergraduate Library Skills: Two Surveys at Johns Hopkins University," *Research Strategies* 11 (fall 1993): 188–201; Godfrey Franklin and Ronald C. Toifel, "The Effects of BI on Library Knowledge and Skills among Education Students," *Research Strategies* 12 (fall 1994): 224–37; Lilith R. Kunkel, Susan M. Weaver, and Kim N. Cook, "What Do They Know?: An Assessment of Undergraduate Library Skills," *Journal of Academic Librarianship* 22 (Nov. 1996): 430–34; Nancy Wootton Colborn and Roseanne M. Cordell, "Moving from Subjective to Objective Assessments of Your Instruction Program," *Reference Services Review* 26 (fall/winter 1998): 125–37; Mollie D. Lawson, "Assessment of a College Freshman Course in Information Resources," *Library Review* 48 (1999): 73–78; Julie Rabine and Catherine Cardwell, "Start Making Sense: Practical Approaches to Outcomes Assessment for Libraries," *Research Strategies* 17 (fall 2000): 319–35.

8. Thomas K. Fry and Joan Kaplowitz, "The English 3 Library Instruction Program at UCLA: A Follow-up Study," *Research Strategies* 6 (summer 1988): 100–108

9. Tiefel, "Evaluating a Library User Education Program."

10. Hardesty, Lovrich, and Mannon, "Evaluating Library-Use Instruction."

11. Ralph Catts, "Some Issues in Assessing Information Literacy," *Information Literacy around the World: Advances in Programs and Research*, ed. Christine Bruce and Philip Candy (Wagga Wagga, New South Wales: Centre for Information Studies, 2000).

12. Walter Dick and Lou Carey, *The Systematic Design of Instruction,* 4th ed. (New York City: HarperCollins, 1996).

13. American Association of School Librarians and Association for Educational Communications and Technology, *Information Power: Building Partnerships for Learning* (Chicago: ALA, 1998).

14. Benjamin D. Wright and Mark H. Stone, *Best Test Design* (Chicago: Mesa Press, 1979).

Statement of Ownership, Management, and Circulation

College & Research Libraries, ISSN 0010-0870, is published bimonthly by the Association of College and Research Libraries, American Library Association, 50 E. Huron St., Chicago, IL 60611-2795. The editor is Willliam Gray Potter, University of Georgia Libraries, Athens, GA 30602-1641. Annual subscription price, $60.00. Printed in U.S.A. with second-class postage paid at Chicago, Illinois. As a nonprofit organization authorized to mail at special rates (DMM Section 424.12 only), the purpose, function, and nonprofit status of this organization and the exempt status for federal income tax purposes have not changed during the preceding twelve months.

Extent and Nature of Circulation

(*Average* figures denote the average number of copies printed each issue during the preceding twelve months; *actual* figures denote actual number of copies of single issue published nearest filing date: November 2002 issue.) 15a. *Total number of copies (Net press run)*: average 13,448; actual 13,495. 15b(1) *Paid/Requested Outside County*: average 11,953; actual 12,017. 15c. *Total Paid and/or Requested Circulation*: average 11,953; actual 12,017. 15d(1). *Free distribution by Mail Outside-County*: average 24; actual 20. 15f. *Total free distribution*: average 24; actual 20. 15g. *Total Distribution*: average 11,977; actual 12,057. 15h. *Copies not Distributed*: average 1,471; actual 1,438. 15i. *Total*: average 13,448; actual 13,495. 15j. *Percent Paid and/or Requested Circulation*: average 99%; actual 99%.

Statement of Ownership, Management, and Circulation
(PS Form 3526, Oct. 1999) for 2002 filed with the
United States Post Office Postmaster
in Chicago, October 1, 2002.