Questionnaire Ambiguity: A Rasch Scaling Model Analysis

A. Bookstein and A. Lindsay

Abstract

One of the most important means of gathering information about libraries has been by the use of questionnaires. Yet many studies show the questionnaire methodology to be an imperfect means of generating reliable information. This paper reviews the types of problems that have been associated with questionnaire based surveys and focusses upon one, the ambiguity of questions. A mathematical model is proposed to explain a type of ambiguity that often occurs in questionnaires and data presented that is consistent with model predictions.

Introduction

It is fitting that in a tribute to Herbert Goldhor so many articles have an emphasis on research methodology. How one properly carries out research has been a lifelong concern of Goldhor as a scholar, as a teacher, and as director of the University of Illinois' Library Research Center. It is significant that he is the author of an early and still respected book on research methods written for a library audience (Goldhor, 1972).

This interest of Goldhor is reflected in the theme of this collection, problem solving, for problem solving begins with a search for reliable, pertinent information, and research methodology deals with how one gathers information in which one can have confidence. That issues such as bias, validity and reliability, proper sampling technique, and instrument construction are still matters that can generate controversy is demonstrated by the recent exchange in *Library Quarterly* (Bookstein & Biggs, 1987; White, 1987).

The position taken in this article is that the types of problems with which researchers in the information sciences deal are complex and that

A. Bookstein, Center of Information and Language Studies, University of Chicago, 1100 E. 57th Street, Chicago, IL 60637
A. Lindsay, The Center for the Great Lakes, 435 N. Michigan Avenue, Chicago, IL 60611
LIBRARY TRENDS, Vol. 38, No. 2, Fall 1989, pp. 215-36

a casual approach to how we obtain the data on which we base our conclusions can result in very serious errors. The tools that we use to get information are deceiving in their simplicity. Selection of objects to examine and formulation of questions all seem natural human activities. Yet when we seek detailed and subtle information, how these tasks are implemented can strongly influence the results we ultimately obtain and our responses to the problems that motivated our investigation. In such a situation it becomes important that the tools by which we gather information become objects of inquiry, and that we undertake the effort to learn how these behave in the contexts in which we use them.

In this article we will report on the results of one such effort in the area of questionnaire design, one of the most heavily used techniques for getting data about libraries. The popularity of questionnaire research is easy to understand—it is both direct and conceptually simple. To conduct a questionnaire survey, we formulate what interests us about an area as a sequence of questions to be presented to the relevant population, and then analyze the responses of those who cooperate, much as we gather information in everyday life. Further, the manner in which the final data are often accumulated makes it very difficult to detect error even when it is present. The scholarly apparatus of coding, tabulation, and statistical testing provides a sense of propriety and security in the results. But these analytical techniques are adjuncts to proper methods for collecting data, not substitutes for them. Even the most sophisticated of statistical packages will digest misleading sets of data as comfortably as they will valid data (perhaps even more so, since poor data collection methods will often impose a degree of regularity on the data not present in the material being studied) (Campbell, 1959).

Responding to a question involves at least four stages of activity:

1. a question is presented to and interpreted by a respondent;
2. the respondent must rely on memory and a variety of cognitive processes to construct his own understanding of what information is needed to answer the question;
3. the respondent must decide whether to answer honestly or at all and what aspects of the information educed to share with the researcher; and
4. the response must be transformed into words or categories understandable by the researcher.

Each of these activities is complex and subject to error.

An awareness that people often do not provide good answers to questionnaires, and that the process of answering questionnaire surveys was deserving of systematic investigation, existed at least by the mid 1950s (Hyman, 1956). Many sources of error in questionnaire response have subsequently been identified and investigated. For example, Bradburn and his associates (1979) have carefully examined the degree to which people respond honestly to questions about socially unaccept-

able behavior; they and others (e.g., Kolata, 1987) consider a range of approaches for compensating for the tendency to distort reality in such situations.

Errors are sometimes introduced in questionnaire responses because of the cognitive processes involved in storing and retrieving information from memory, even when the respondent intends to cooperate fully with the research. The format of a question, for example, whether it is open- or closed-ended, has been found to affect responses (Schuman & Scott, 1987) partially because the framework it establishes for the response categories influences what information is retrieved from a respondent's memory. The ability to retrieve information from memory and mechanisms that might distort the results of such a retrieval have often been studied. For example, very pronounced and systematic effects are present when time related information is requested (Neter & Waksberg, 1964; Sudman & Bradburn, 1973; Bradburn, et al., 1987). But concern about how people respond to questions demanding quantitative information, and in particular, how they use words denoting quantities, has been evident for some time (Simpson, 1944; Hoyt, 1972; Pepper, 1974), including one study taking place within a library context (Kidston, 1985).

Of particular interest to us are problems of ambiguity in questions about library activity. Although it has long been recognized that people understand the same words in different ways, and that this affects their responses to questionnaires (Payne, 1951), it is only relatively recently that the implications of this for research in libraries have been probed. In this article we will examine one study (Bookstein, 1985; Kidston, 1985) that did find differences in how people understand words that occur frequently in library surveys.

The term *ambiguity* refers to the problem that different people understand the same term or expression in different ways. But though the single term, *ambiguity,* is used, there are many reasons why the phenomenon it refers to might happen. We believe that ambiguity is a serious and easily overlooked problem in questionnaire design, and that understanding more precisely why it is that two people might disagree on the meaning of a commonly used term is an important first step in learning how to control this problem. In the papers by Bookstein (1985) and Kidston (1985) noted earlier, a specific and very interesting mechanism for such disagreement suggests itself that might apply to a variety of terms occurring in library questionnaires. Specifically, we argue that a source of ambiguity of some terms—for example, the word *use*—is a scaling phenomenon. Different activities that take place in libraries are associated with libraries to different degrees, while at the same time individuals differ in their willingness to accept a degree of "librariness" as constituting a library activity. When such words occur in the questions people are asked about libraries, how people respond can be influenced by their location on this scale. If this phenomena is in affect,

the disagreements it produces should exhibit a great deal of regularity; studying data designed to bring out these regularities should both reveal the existence of such a scale and allow us to place both library activities and respondents on the scale. In this article we will explore the ability of a scaling technique, the Rasch Scaling Model, to fit and explain data we have collected to display disagreement on whether specific activities constitute library uses.

In the following sections we will describe the experiment that was carried out to explore the problem of question ambiguity and apply the Rasch Model to the resulting data. However, as not all readers are familiar with scaling methodology, we will first offer a quick overview of what scaling is and, specifically, describe the Rasch approach to scaling.

Background

The measurement of attitudes and perceptions is common in social science fields like education, psychology, and sociology. The measurement procedure often involves a series of items on a questionnaire. When successful, the process results in a well defined variable, on which both the items used to make the measurement and the subjects being measured are assigned values, depending on the extent to which they exhibit the quality in question. This is essentially a scaling process.

In this study, we apply a method of scaling to the field of library and information science. The attitude under investigation is a somewhat abstract concept that we refer to as "library sensitivity." By this we mean the propensity to identify activities occurring in libraries as inherently library activities.

People vary in their use of libraries and their attitudes about libraries. Attempts have been made to explain such variability in terms of demographic and social variables. Such efforts always leave a substantial amount of variance unexplained. We are suggesting the existence of library sensitivity as an intrinsic personality variable that may contribute to explaining user behavior; we also describe a means for measuring this quantity by using the Rasch Psychometric Scaling Model to develop a scale of library sensitivity. If this personality trait is in fact a definable and scalable phenomenon, the establishment of a formal measurement tool would provide a means of measuring this trait, which in turn would permit us to observe correlations between this characteristic and other personal qualities.

The possibility that library sensitivity might exist as a personality characteristic was suggested by the results of previous research carried out by Bookstein (1985) and Kidston (1985) on questionnaire design in a library context. In these studies, several problems in questionnaire research were examined. Particularly interesting was the problem of question interpretation—that is, whether different people share a common understanding of the phrases used in questionnaires to describe

basic library activity. The concept of *library use* was found to be ambiguous. On the basis of these studies, it is reasonable to expect that when subjects are asked how often they *used* the library or a library's material (as opposed to a more specific question such as, How many times did you check out a book?), the resulting data are probably inaccurate. This is because, as these studies show, interpretations of the term *use* vary widely—two people, having performed the same library activities, may very well respond differently to the library use question although both are trying to respond honestly.

In the above research, respondents were presented with descriptions of a number of activities occurring in libraries, and, for each, asked whether, if they had engaged in that activity, they would describe themselves as having used the library. Table 1 (Bookstein, 1985) shows the responses of two groups of people to questions about their interpretations of library use. The GLS group was comprised of University of Chicago Graduate Library School students and their friends while the GSB group was made up of University of Chicago Graduate School of Business students. The results show that even within fairly homogeneous groups of people, there is much variability in the perception of library use. For example, each group was split approximately evenly on accepting the action of unsuccessfully trying to find a book in the collection as a library use. Even actions which most people agreed to view as library uses (e.g., recalling a book) still showed some disagreement.

Our interest here in these results is not in the surprising variability of interpretation but rather in the regularity that appears within the variability. As seen in Table 1, both groups' overall ranking of the questionnaire items is approximately the same—recalling a book, for example, is the item most frequently seen as a library use by either group. The most striking disagreement in ranking is that the GSB group much more often saw "reading own book" as a library use. This is "probably because, for most, this is all they did in the library" (Bookstein, 1985, p. 26). On the other hand, although items were ordered similarly in both groups, the GSB group seemed nearly consistently less likely to describe an activity as a library use. It is the systematic and probabilistic character of this response pattern that we are trying to explain in terms of a Rasch-Model scale.

According to Wright and Masters (1982): "The invention of a variable begins when we notice a pattern of related experiences and have an *idea* about these experiences which helps us to remember their pattern. If the idea orients us to more successful action, we take it as an 'explanation' of the pattern and call it a theory" (p. 1). The pattern displayed by the GLS and GSB groups suggests the existence of a variable representing the willingness of individuals to see actions as library uses or, more generally, representing library sensitivity. Bookstein (1985) suggested a preliminary scaling model to represent this idea:

> people engage in a wide range of activities in a library. These activities fall along a scale of "librariness"—the extent to which people tend to associate that activity with libraries. On the other hand, people also fall along a scale according to their willingness to see an activity as a library use. The response to

TABLE 1
POPULATION AGREEING ON WHAT CONSTITUTES LIBRARY USE

Action	*Percent Considering Action as a Use of Library*	
	Bookstein (GLS) (n=43)	*Kidston (GSB) (n=90)*
Recalled a book	88%	74%
Duplicated an article	81%	65%
Checked name in card catalog	70%	53%
Read own book	58%	89%
Tried unsuccessfully to find book	51%	42%
Returned a book	38%	18%
Met a friend	19%	24%
Used restroom	19%	3%

> a particular question,then, is governed by both the position of the activity and the individual on this scale. (p. 26)

In our study, we build on this preliminary model by applying the Rasch Model to the instrument described earlier. This psychometric scaling model enables us to determine formally whether the phenomenon observed is indeed a measurable attitude variable.

SCALING METHODS

Background

Scales are created to compare characteristics of objects along a common unit of measure. While scaling methods share this objective, their procedures vary greatly. (The characteristic being measured in our study is an attitude or personality trait; however, the same scaling methods can be used for other variables such as amount of knowledge on a specific topic or personality traits such as introversion.)

The simplest and most direct scaling method consists of a subject marking his own location on a graphical representation of the scale. This method requires a detailed description of the concept being measured in order to indicate clearly what values of the concept each location on the scale represents. The method also depends on an honest and objective self-evaluation on the part of the subject. More often, scaling methods act indirectly, combining a subject's responses to several items into one score that is then used to locate the subject's position on a scale. Instead of directly asking a subject to identify his location on the scale, they solicit related information that allows a scaling procedure to calculate the location. Item inventories are a popular means of obtaining the multiple responses needed for this method. In this section, we will review scaling techniques used in conjunction with questionnaire-like item inventories.

In developing a scaling questionnaire, one assumes that the con-

cept being measured can be defined by a set of items that follow a single line of inquiry (Wright & Masters, 1982). This simply means that a number of statements can be made relating to the concept in question, that these statements can be placed on a scale according to the amount of this concept that they represent, and that most people would agree with this placement. This quality is known as unidimensionality. For some cases, the unidimensionality of an attitude variable has already been established. In other cases, it has not been proven but may seem likely. In the latter case, the scaling operation takes on the additional task of testing the hypothesis that individual items being combined to form a single scale score also can meaningfully be organized along a single dimension (Kidder, 1981). If this hypothesis proves false, a meaningful univariate scale cannot be devised.

Some concepts are broad enough to have many aspects (or dimensions) along which scales might be formed. "When the notion of measurement is applied to so complex a phenomenon as opinions and attitudes, we must here ... restrict ourselves to some specified or implied continuum along which the measurement is to take place" (Thurstone, 1928). Using a restricted set of items to measure a complex concept is an accepted part of the scaling process. In general, effort is made to measure only a single interesting aspect of the concept. Thurstone (1928) notes that when measuring a table, for example, usually only one attribute (e.g., height, cost, or beauty) is being measured. This attribute is used to represent the more general concept, which is the entire table, in a particular domain of investigation.

In addition to unidimensionality, scaling models have other requirements and characteristics. They must be able to describe small as well as extreme degrees of difference between objects. Therefore, the items on the questionnaire must be varied enough to represent a wide range of values of the concept.

Rasch scaling procedures begin with raw scores. "All information about a person's ability expressed in his responses to a set of items is contained in the simple unweighted count of the number of items which he answered correctly. [For the Rasch Model] raw score is a sufficient statistic for ability. For item difficulty, the sufficient statistic is the number of persons who responded correctly to that item" (Wright & Mead, 1977). While raw person and item scores provide the information necessary for scaling, they cannot directly form the scale because of several inherent problems.

Essential to scaling is the notion of a common unit of measure. We are familiar with standard measurement units such as the inch and centimeter. With regard to tests and questionnaires, appropriate units of measurement are less obvious. Students are sometimes ranked by raw test scores, where, for example, a score of twenty correct out of twenty is perfect and ten correct out of twenty is failing. One might be inclined to assume that a student achieving a score of ten possesses half as much of

the characteristic probed by the test as a student scoring twenty. However, comparison of the students' abilities cannot be made on the basis of raw scores in the same way that two height measurements could be compared. Raw scores depend strongly on the particular test or questionnaire items in use; therefore, they describe the ordering of subjects but not the distance between them. Thus, in the earlier example, the amount of the variable required to correctly answer the second ten test items may be much more or only slightly more than that which the failing student possesses. Item and subject scores must be properly placed on a common scale with respect to the amount of the variable they exhibit in order to measure distances between scores. Scaling methods, in varying ways, transform raw item and person scores into these single-scale values.

Scaling models should also attempt to free the scale from any dependence on a particular set of items or subjects. Raw scores are, of course, completely dependent on the particular questionnaire items and subjects involved. A questionnaire with most items representing small amounts of the variable would tend to produce higher raw scores than would a questionnaire in which most items represented more of the variable if both were given to subjects with the same distribution of attitudes. Our goal is to obtain equal scale values for those subjects that exhibit equal amounts of the variable being measured no matter what set of items is used on the questionnaire. Therefore, scaling methods must take into account the demands of the items and accordingly translate the raw person scores to scale values. An analogous operation must take place with the item scores and item scale values. Raw item scores depend on the attitude levels of the test subjects. This dependency must be removed if the items are to be located on a scale according to the absolute amount of the variable they exhibit.

Finally, scaling methods should provide some means of testing the fit of the data to the scaling model. One can always devise rules to place raw scores on a linear scale; however, one must be sure that the resulting item and person scale values make sense. This may be done by studying the actual detailed responses of individuals, considering the scale values they are assigned, and deciding whether this is logically acceptable. Tests of fit may also be done using more advanced statistical methods.

Some Examples

During this century many scaling models have been developed. We will review some of these briefly. Pioneering work was done by L. L. Thurstone in the 1920s and 1930s. Thurstone is most closely associated with *differential scales* (Kidder, 1981, p. 301). These are attitude scales whose items are statements that represent the entire range of possible opinions (i.e., items are included that oppose, support, and are neutral toward the concept). The items are ranked by human judges, and subjects are located on the scale according to the items with which they

agree. If the model fits, subjects should agree only with a subset of the items located adjacent to each other on the scale and should not agree with items to the left and right of these. Scale scores can be computed for those who agree with items spread out along the scale, but the meaningfulness of such scores is questionable. Thurstone applied the scale model characteristics discussed earlier with varying degrees of success. He was aware of the need to free the scale from dependence on a particular subject group or set of questionnaire items. His adjustments for these sample effects were based on group level descriptions of the ability distributions, such as the mean and standard deviation, and on the assumption that the ability distributions were normal (Engelhard, 1984).

Cumulative scales differ from differential scales in the way that individuals respond to their items. Here the items are statements that either support or oppose (in varying degrees) the concept. "The items are related to one another in such a way that, ideally, an individual who replies favorably to item two also replies favorably to item one; one who replies favorably to item three also replies favorably to items one and two; and so on" (Kidder, 1981, pp. 217-18). (The numbers one, two, and three above indicate the items' ranked position on the scale, not their position in the questionnaire.)

Louis Guttman (1944), created the notion of a cumulative scale with his scalogram method. This method graphically and statistically tested for cumulative scale-type patterns of responses. Guttman (1950) saw this method essentially as a test of unidimensionality and described it as follows:

> The basic condition to be satisfied is that persons who answer a question "favorably" all have higher scale scores than persons who answer the question "unfavorably". This constitutes a rigorous definition of a scale. It provides a simple, objective technique for testing the existence of a single variable, that is, for determining whether the questions have the same meaning for all respondents. (pp. 76-77)

Guttman's model is based on the principle that a subject will respond consistently (either favorably or unfavorably) to each item on the scale up to the level of his ability/attitude. Beyond this point, he will continue to respond consistently but in the opposite manner. Once a person's scale score is known, his response to any particular item is predicted by the model to be either definitely favorable or unfavorable (depending on where the item is located on the scale, relative to the subject's location). Guttman realized that perfect scales do not exist and allowed for some deviation from the model's required response pattern. His "coefficient of reproducibility" is a statistic indicating whether the data's deviation from the model is significant or not.

Specific examples of these types of scales can be found in standard textbooks on research methods, such as Selltiz, Wrightsman, and Cook's *Research Methods in Social Relations* (Kidder, 1981).

Rasch Model

Overview. The scaling model which we employ for this study was developed in the 1950s by the Danish mathematician Georg Rasch while developing educational accomplishment scales. This model incorporated much of Guttman's idea of a cumulative scale but improves on it by the addition of a probabilistic response function. Rasch argued that the deterministic models of classical physics did not suit descriptions of human behavior. Here, he believed, it is better to apply the nondeterministic models of modern physics—i.e., "(to employ probability) where chance plays a decisive role: The possible behaviour of a pupil is described by means of a probability that he solved the task" (Rasch, 1980, p. 11). By defining a probabilistic model that relates performance to person and item parameters, Rasch's model makes it possible both to estimate person and item scores and also assess whether the model fits, or reasonably describes, the data.

The Rasch Scaling Model stems from the idea that a person's response to an item in an ability test is governed by two factors: the difficulty of the item and the ability of the person. (In the case of attitude measurement, item "difficulty" corresponds to the difficulty in agreeing with the item; person "ability" corresponds to the amount of the attitude exhibited by a person or the ease with which a person agrees to items expressing the attitude.) Given item difficulties and subject abilities, the model describes the probability of a certain response (contrasting with other models that describe with certainty what that response will be).

The governing factors—person ability and item difficulty—constitute the two parameters of Rasch's model. A distinguishing feature of the model is that these parameters are separable, making it possible to derive estimators of each parameter independently of the other. Thus, the separability makes possible objective measurement in the sense that measurement is not dependent on a particular set of items or persons. Rasch (1980) refers to this quality as "specific objectivity" (p. 11). Previous researchers (e.g., Thurstone) recognized the necessity of such objectivity and attempted to achieve it by procedures accompanying their models. The Rasch Model stands out in that this objectivity is achieved by the model itself. Rasch's model takes the form:

$$P\{success \mid \beta_v; \delta_i\} = e^{(\beta_v - \delta_i)} / [1+e^{(\beta_v - \delta_i)}]$$

which determines the probability of a successful response by person v to item i, where β_v is the ability measurement of person v and δ_i is the difficulty measure of item i.

Scale Characteristics

The scale defined by this model is measured in terms of units called

logits. The measurements β and δ are logit values. A person's ability β represents the natural log of the odds in favor of his succeeding on an item whose difficulty is at the origin of the scale (i.e., whose δ value equals zero). "A person with ability 0.0 (i.e., ability equal to the difficulty of an item at the origin) has an even chance (odds 1 to 1) of succeeding on the item since $\log(1) = 0$..." (Wright & Mead, 1977, p. 20). Similarly, the item difficulty δ represents the natural log of the odds in favor of a person with ability β equal to zero failing on an item. When person ability β is equal to item difficulty δ, the person has a .5 probability of agreeing with that item.

These results are as expected—e.g., one would expect a person to be likely to agree with an item that he surpasses in librariness (i.e., an item that is easy for him). Conversely, one would expect a person to be likely not to agree with an item that exceeds him in librariness. When the person and item coincide, the person is as likely to respond one way as the other.

As mentioned previously, raw scores are sufficient statistics for creating scales. The task of Rasch analysis is to transform the raw scores into a unidimensional interval scale. Item scores are converted to logit difficulty estimates in a process called item calibration. Similarly, person scores can be converted to logit ability estimates. Various estimation techniques can be used (see Wright and Mead [1977] and Wright and Douglas [1977] for examples). In our study we used the unconditional maximum likelihood procedure as developed by Wright and Panchapakesan (1969)—maximum likelihood estimates a probability distribution's parameters by setting them equal to values that make them as consistent as possible with the observed data. For our problem, it is an iterative procedure that takes raw item and person scores and converges them to the best fitting β and δ values. The estimation procedure requires the elimination of zero and perfect item and person scores. The model cannot scale such scores. Consider, for example, a person with a perfect total score. We cannot ascertain if this person's ability should be placed slightly higher than all other ability estimates or much higher. We cannot determine this unless we add an item whose difficulty is greater than the person's ability (i.e., an item that the person cannot agree with or cannot answer correctly).

Tests of Fit. Scale values have now been established for persons and items. Before continuing, however, we would like to confirm that the items do represent values along a single scale and that the subjects responded consistently to this same single variable—that is, we must ascertain that the model fits the data.

We check item and person fit separately. In testing item fit, we look at the individual responses to each item. Based on the model, we expect persons with sensitivity values less than the difficulty value of the item not to agree with the item and those with greater sensitivity values to

agree with the item. As the model describes, there is, of course, some chance that the responses will not occur this way. We take this into account and allow for some variability in expected responses. An item may be judged as not fitting when it receives a significant number of unexpected responses. An item that does not fit is most likely drawing on knowledge and/or attitudes that do not correspond to the concept being measured. For example, Wright and Masters (1981), in a test of drug knowledge, found that a question about drug legality did not fit with other questions that focused more on the use and effects of drugs. The legal question was estimated to be fairly difficult. However, those who had a good knowledge of the more scientific aspects of drugs (and hence received high ability ratings) turned out to be the ones who most often missed this question. Those with less estimated drug knowledge answered this question correctly with an unexpectedly high frequency. Whatever the legal question is measuring, it is not the same knowledge concept required by other items. Such nonfitting items should be removed and the remaining items should be recalibrated.

Similar lack of fit can occur with persons as well. This can be caused by such things as cultural or educational differences. If, for example, a person obtains a low scale rating but agrees with (or answers correctly) only the least likely to be agreed with items, that person is not responding along the same dimension as described by the scale. In such cases, the scale score should not be used as it is an inaccurate and possibly an unfair measurement.

Misfitting items and persons sometimes stand out due to very unexpected responses. However, we generally need a statistical test to determine whether the response variation is significant. The Rasch Model makes this type of test easy. The model describes the probability of each response for every person-item interaction (by substituting fitted β and δ values in equation 1 noted earlier). We can subtract this expected value from the observed value to obtain a score residual. These residuals can be summed over all item responses for a particular person or over all person responses to a particular item. From these sums a t statistic—a standardized residual—can be calculated (Wright & Masters, 1982, p. 101). When this statistic is significant at the .05 level, the person or item should be removed. Other fit tests are described in Wright and Panchapakesan (1969) and in Wright and Mead (1977).

Rasch Model Applications. Rasch scaling is a very general technique and many examples of its application are available. Among them are pistol marksmanship of Military Police (MP) candidates (Wright & Mead, 1977), knowledge about drugs, attitude toward drugs, fear of crime, and knowledge of physics (Wright & Masters, 1982). Some examples of how such a model might illuminate problems in library and information science follow.

1. *Indexing Example.* Consider the task of assigning index terms to documents. Viewed abstractly, this process has much in common with taking a test or completing an attitude-measuring questionnaire. Here, each index term is associated with a scale: with respect to a particular term, a document will be about that term to some estimable degree (from not at all to very much), in the same way that a person might possess a particular ability or attitude in some measurable amount. At the same time, each indexer has a different threshold for assigning a particular index term in the same way that each item on a test or questionnaire has a threshold ability or attitude requirement in order for it to be answered correctly or agreed with. According to this model, the difference between an indexer's location on the scale and a document's position determines the probability that the indexer will assign that term to the document. Therefore, just as test items and people can be measured and located along a common scale with regard to an ability or attitude, indexers and documents can be scaled with regard to a particular indexing term or concept, and the likelihood of an indexer's assigning a term to a document can be modeled by equation 1. Indexing is a complex and poorly understood process. There has been much controversy regarding the degree to which indexers are inconsistent. The possibility of an underlying scaling phenomenon sheds light on one aspect of this problem. This model should be tested; if valid, scaling documents and indexers in this way would be helpful in understanding the problem of interindexer inconsistency in that it would show the differences between indexers in a concrete way. Dealing with those differences in order to improve indexing consistency could then be better addressed.
2. *Collection Development.* The model has been used to illuminate aspects of library collection development (Bookstein, 1988). Libraries differ in the strengths of their collections, and books differ in their desirability to libraries because of the subjects they represent and the strength of those subjects in the libraries' collections. Whether or not a library gets a book can be stochastically associated with these factors. In terms of the Rasch Model, a book is like a test or questionnaire item in that it represents a particular amount of a subject. Its difficulty level might be described as the amount of difficulty an acquisitions librarian would have in selecting the book, taking into consideration the library's strength in the subject area represented by the book. Libraries are like people being tested in that they have varying strengths in a subject area just as people have varying abilities or attitudes regarding a particular test variable.

   If this model is valid, library collections and individual books can then be scaled with regard to a particular subject area such as calculus or botany. Those libraries ranking high on the scale would be those with the strongest collections in this subject and, hence,

those with the greatest ease in selecting a book on the topic. Books ranking high would be those with a high difficulty of being selected. Only libraries with the strongest collection rankings would purchase such books.

A useful concept that follows this model is that of a "Peer Group" of libraries; with respect to a class of books, a peer group of libraries is a group in which the likelihood that a library will acquire a book is determined by the strength parameters of the book and the library alone. Since any other considerations separate a library from the group, another less formal way of expressing this idea is that a peer group of libraries (with respect to a class of books) is a group differing in collection strength but sharing a collection personality. The model proposed here serves as a formal definition of a peer group; it shows us how to evaluate the pertinent parameters, and it provides a mechanism for alerting us to instances when a library has not acquired a book that seems appropriate for it to acquire given its membership in a peer group. A third example, which constitutes the body of this article, is based on an application of the Rasch analysis to questionnaires.

Applying the Rasch Model to Question Ambiguity

In our study, we use a questionnaire of twenty-two items to define the library sensitivity scale (for the entire questionnaire, refer to Bookstein, 1980). The dimension we develop deals with the concept of use as it applies to libraries and to research tools such as journals. We ask subjects to indicate whether they consider certain activities to be uses of research material or of the library itself. Our questionnaire is based on those used by Bookstein (1985) and Kidston (1985), with minor changes. We chose this approach to the concept of library sensitivity since the previous research has already shown the scaling tendencies.

Our questionnaire consists of four sections, each with several items. Section 1 describes various interactions one might have with a journal in a library, and asks if one considers that interaction to be a use of the library. For example, item 3—"Would you say you 'used' the library today if you obtained an issue of a journal or magazine in that library and looked at the advertisements while waiting for a friend." Section 2 asks whether journals are considered to be "used" or "read" under certain conditions; for example, item 4—"Would you describe a journal as being among those you 'read' (on a continuing basis) if you subscribe to it as one of several journals in a field close to your main area of interest? However, you only have time to scan carefully three or four of the twelve issues published each year." Section 3 focuses on use as it relates to books and libraries. For example, item 8—"While looking for a book in the stacks you notice a book you weren't aware of with a title suggesting it is on the same topic. You glance through it. Would you say you 'used' the book if, after seeing the publisher and copyright date, you

return it to the shelf?" Section 4 describes a variety of activities carried out in a library (many of which can also be done elsewhere) and asks whether each constitutes a use of the library; an example is item 13—"If you were asked how often you 'used' the library, would you count the time when you checked the spelling of an author's name by referring to the card catalog?" Each item requested a yes or no answer from the subject.

The motivation behind this investigation was to understand better what people mean when they respond in a library use survey that they used a library. The variation in response that we found suggests that it is quite possible for several people to engage in the same or similar activities, yet some would, on the basis of those activities, say yes, they did use the library, while others would say no, they did not. Thus, when we learn from a survey that a certain percent of the population used the library last month, we should recognize that this figure reflects differences in interpretation of the word *use* as well as differences in the behavior of interest.

The questionnaire was given to forty-two individuals—thirty-three students in a research methods class at the Graduate Library School of the University of Chicago, four librarians at a Chicago special library, and five nonlibrarians. Random selection of subjects was not necessary since item calibration can be done with any set of subjects as long as the model fits for most of the subjects.

The responses were analyzed with the MSCALE computer program developed in the Department of Education at the University of Chicago (Wright et al., 1987). This program estimates person abilities, item difficulties, and tests the fit of the model. It can be used with multiresponse category items or with dichotomous data.

Results

Our data matrix (see Table 2) shows the responses of each tested subject to each item. The number of items listed has been reduced from the original twenty-two to seventeen as five items had perfect scores (all subjects responded "yes" to these) and were therefore removed. One subject responded "yes" to every item and was removed leaving the final count of scalable subjects at forty-one. The subjects are ordered from those most likely to interpret an activity as a library use at the top, to those least likely at the bottom of the chart. Similarly, items are ordered from those easy to interpret as a library use (on the left) to those most difficult. This arrangement emphasizes any inconsistency in the pattern of responses.

Item Analysis

A chart of item calibrations as calculated by the MSCALE program is shown in Table 3. The score column indicates the number of subjects agreeing with the item out of a total sample size of forty-one. The

TABLE 2
PATTERN OF RESPONSE OF PERSONS (ROWS) TO ITEMS (COLUMNS). A ONE INDICATES THE PERSON AGREES WITH ITEM; A ZERO INDICATES DISAGREEMENT. BOTH ROWS AND COLUMNS HAVE BEEN ORDERED ACCORDING TO SCALE VALUES

		ITEM NUMBER (easy items — — → difficult items)																	
		11	9	16	12	18	4	10	6	5	7	15	22	3	14	20	17	8	TOTAL
	11	1	1	1	1	1	1	1	1	1	1	1	1	0	0	1	1	1	15
PERSON	6	1	1	1	1	1	1	1	1	1	1	0	1	1	0	1	1	0	14
NUMBER	39	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	14
	30	1	1	1	1	1	1	1	1	1	1	1	0	1	1	0	0	0	13
	34	0	1	1	1	1	1	0	1	1	1	1	1	0	1	1	0	1	13
(high	8	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	13
scores	2	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	13
	25	1	1	1	1	1	0	0	0	1	1	1	1	1	0	1	1	0	12
!	36	1	1	1	1	0	1	1	1	1	1	0	1	1	0	0	0	1	12
!	37	1	1	1	1	0	1	1	1	1	0	1	1	1	1	0	0	0	12
↓	24	1	1	1	1	1	0	1	1	1	1	0	1	1	1	0	0	0	12
	38	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	12
low	12	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	12
scores)	21	1	1	1	1	1	1	1	1	0	1	0	1	0	1	0	0	0	11
	19	1	1	1	1	1	0	1	0	1	1	1	0	1	1	0	0	0	11
	35	1	1	1	1	0	1	1	1	1	1	1	1	0	0	0	0	0	11
	17	1	1	1	1	1	1	1	1	0	1	1	1	0	0	0	0	0	11
	23	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	11
	33	1	1	1	1	1	1	1	0	1	1	0	0	1	0	0	0	0	10
	32	1	0	1	0	1	1	1	1	1	1	1	1	0	0	0	0	0	10
	28	1	1	1	1	1	1	1	0	1	0	0	1	0	0	0	1	0	10
	16	1	1	1	1	1	1	0	1	1	1	0	0	1	0	0	0	0	10
	3	1	1	1	1	1	1	1	1	0	1	0	1	0	0	0	0	0	10
	27	1	1	1	1	1	1	1	0	1	1	1	0	0	0	0	0	0	10
	10	1	1	1	1	0	1	0	1	1	0	0	1	1	0	0	0	0	9
	1	1	1	1	1	0	1	1	1	1	0	1	0	0	0	0	0	0	9
	14	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	1	9
	9	1	1	1	1	0	0	1	0	1	1	1	1	0	0	0	0	0	9
	4	1	1	1	1	0	0	1	1	0	1	0	0	0	1	0	0	0	8
	13	1	1	1	0	1	0	1	0	1	1	1	0	0	0	0	0	0	8
	20	1	1	1	1	1	1	0	0	0	1	0	1	0	0	0	0	0	8
	31	1	0	0	0	1	1	1	1	1	1	0	0	1	0	0	0	0	8
	29	1	1	1	1	1	1	0	1	0	0	1	0	0	0	0	0	0	8
	7	1	1	1	1	1	0	1	0	1	1	0	0	0	0	0	0	0	8
	5	1	1	1	1	1	0		0	1		1	0	0	0	0	0	0	7
	18	1	1	0	0	1	1	1	1	1	0	0	0	0	0	0	0	0	7
	26	1	1	1	0	1	1	0	1	0	0	1	0	0	0	0	0	0	7
	15	1	1	0	0	1	1	0	1	0	1	0	1	0	0	0	0	0	7
	40	1	1	1	1	0	0	1	0	0	0	1	0	0	1	0	0	0	7
	22	1	1	0	0	1	1	0	1	1	0	0	0	0	0	0	0	0	6
	41	1	0	0	1	0	0		1		0	0	0	0	0	0	0	0	3
TOTAL		40	38	36	34	32	31	30	30	30	29	23	22	14	9	4	4	4	

(Blanks indicate missing responses)

measure column is the item difficulty figure (in logit units), with high values indicating items more difficult to interpret as a library use. Error is the standard error of the measurement. The fit statistic is calculated from the item residuals. The expected value of this statistic is zero with values of absolute size greater than two indicating lack of fit.

Some of the most difficult items are those that ask about library use in connection with activities that might occur in a different setting. (Examples are items referring to using the restroom, meeting a friend, and looking at journal advertisements.) Attending a lecture in the library, while not a library specific activity, was much more frequently considered as a library use. Perhaps this is because it has a more intellec-

TABLE 3
ITEMS APPEARING IN TEST, WITH RAW SCORE, RASCH SCALE VALUE, ESTIMATED STANDARD ERROR AND FIT MEASURE. MOST ITEMS HAD ONE OF THREE STANDARD INITIAL SEGMENTS FOLLOWED BY A SPECIFIC ACTION. THESE INITIAL SEGMENTS ARE ABBREVIATED BY A, B, AND C

ITEM STATISTICS					
NUM	QUESTION CONTENT	SCORE	MEASURE	ERROR	FIT
8	B + Would you say you "used" book if, after seeing publisher and copyright date, you return it to shelf.	4	3.14	.57	.27
17	C + Went only to use restroom.	4	3.14	.60	-.10
20	C + Use library as place to meet friend.	4	3.14	.64	-.72
14	C + Went in to return a book.	9	2.06	.46	.93
3	A + Looked at advertisements while waiting for a friend.	14	1.33	.37	.01
22	C + Attended a lecture held in library's meeting room.	22	.35	.38	-.93
15	C + Went to check out a book, but it wasn't available. You left, planning to try again next week.	23	.23	.37	.78
7	You scan journal in library, but don't subscribe. You see about 25% of the issues this way. Would you say you regularly "use" the journal.	29	-.52	.41	-.69
5	You subscribe to a journal, but have only time to scan carefully 3 or 4 of the 12 yearly issues. Would you say you regularly "used" the journal.	30	-.68	.40	-.40
6	You scan journal in library, but don't subscribe. You see about 25% of the issues this way. Would you say you "read" this journal.	30	-.75	.45	.55
10	B + Would you say you "used" book if you glance through only to see how conventional diagrams are now being represented.	30	-.78	.41	.58
4	You subscribe to a journal, but have only time to scan carefully 3 or 4 of the 12 yearly issues. Would you say you regularly "read" the journal.	31	-.81	.41	.39
18	C + Brought own materials into library to study.	32	-.96	.44	.64
12	C + Went to duplicate an article in a journal you already knew the library had.	34	-1.31	.46	-.59
16	C + Went to check out a book, but it wasn't available so you had it recalled.	36	-1.74	.59	-.87
9	B + Would you say you used *library* if, after seeing publisher and copyright date, you return it to shelf.	38	-2.34	.67	.02
11	B + Would you say you used *library* if you glance through book only to see how conventional diagrams are now being represented.	40	-3.52	1.10	1.18
2	A + Looked up an article you were referred to. After examining it, you decide not to use it.	41	n/a	n/a	n/a
19	C + Checked out a book, but soon found it not useful and returned it.	41	n/a	n/a	n/a
13	C + Checked spelling of a name by referring to card catalog.	41	n/a	n/a	n/a
21	C + Had online search performed.	41	n/a	n/a	n/a
1	A + Read titles/abstracts of several articles before deciding none was useful.	41	n/a	n/a	n/a

CONDITIONS
A: Would you say you "used" library today if you obtained a journal in that library and ...
B: While looking for a book in stacks you notice a book you weren't aware of with a title suggesting it is on same topic. You glance through it ...
C: If asked how often you "used" the library, would you count the time when you ...

tual quality and is therefore more closely associated with traditional library operations. Actions that require interaction with library personnel or procedures are among the least difficult to agree with. Examples are checking out a book—even though it proves not to be useful—recalling a book, and having an online search performed. All of these have either low difficulties or were seen as library uses by all subjects (and so could not be scored).

There appears to be a difference in the perception of the concept of use as it relates to libraries and library materials. An item asking whether a book was used when its publisher and copyright date were

noted and then it was returned to the shelf was the most difficult to agree with. However, a similar item, asking whether the same interaction constituted a library use, was one of the easiest items.

In examining item fit, we first look at the ordering of items by difficulty estimates to see if it makes sense. Our results do seem to be a subjectively reasonable ordering. Similarly, the pattern of responses shown in Table 2 seems compatible with what we would expect if the model were correct. Beyond these observations, we look at the fit statistic (the standardized weighted mean square residual figure). When this is more than 2.00 or less than -2.00, it indicates a significant lack of fit. That is, the item's standardized mean squared residual is more than two standard errors away from its expected value of zero. T distribution tables tell us that such a deviation would occur by chance only 5 percent of the time. Therefore, we consider such a deviation from modeled values to be significant and reject the item as not fitting. In our case, however, all items are acceptable. The largest fit statistic is item 11's 1.18—well within the +/-2.00 limit. In other words, our subjective sense that the pattern of responses is compatible with the existence of an underlying scale is confirmed by formal analysis—discrepancies from perfect scaling are explainable by the probability model.

Examining the measure column of Table 3 shows that the distribution of our items is not uniform over the length of the scale. We would prefer that items were evenly spread out, rather than clustered, as we find around values -1.00 and 3.10. Including several items with similar or equal difficulties does not harm the scaling process; the additional items are just superfluous. There is even an advantage to developing an item pool that includes multiple items at each difficulty location. With this, tests and/or questionnaires can be developed that use different items while representing equivalent difficulty distributions.

We would, however, prefer to eliminate significant gaps between item measurements. As Table 3 illustrates, there are gaps in our scale between values -3.5 and -2.3, -0.5 and 0.2, 0.4 and 1.3, 2.0 and 3.1. Filling in gaps between item difficulties makes more precise person measurement possible. To improve our scale, we would need to design additional items that fall within the gaps observed. The questionnaire could then be readministered and item difficulties recalibrated to locate the new items on the scale.

Finally, we would like our scale to include as many easy items as hard ones. Our scale is slightly weighted toward the easy side, as we have ten items with negative difficulties, but only seven with positive difficulties. The estimation procedure we used centers the item difficulties around a mean value of zero. Ideally, we would also like the median difficulty value to be zero. When items are evenly distributed over both positive and negative halves of the scale, measurement becomes more precise.

Person Analysis

Person ability estimates from the MSCALE program are organized in the same fashion as item measures. High scores indicate a strong

willingness to agree with questionnaire items while low scores indicate agreement with only the easiest items.

Fit statistics indicate that there are two nonfitting persons, 25 and 34. If we examine the data matrix (see Table 2) we see the response inconsistencies that lead to the high fit statistics.

The columns of the data matrix contain the responses to individual items. The bold line cutting through the matrix indicates the point at which the item difficulties overtake the person abilities. In other words, this is the point where the most likely response changes from one (agreement) on the left of the line to zero (nonagreement) on the right. As our model describes, persons with high scale values are expected to agree with more items than those with lower values. The position of this line demonstrates this expectation. In the Guttman model, a perfect scale would show all ones to the left of this line and all zeros to the right.

The arrangement of the data matrix allows us to see inconsistent responses at a glance. Looking at the row for person 34, we see that this person, who tended to agree with most items, unexpectedly did not agree with three items, including item 11, which for most subjects was the easiest item to agree with. Person 34 also unexpectedly agreed with item 8 which was the hardest one to agree with. While the model allows for an imperfect response pattern, extreme inconsistencies are unacceptable. Subjects who exhibit them cannot be measured on the scale, and if many such subjects existed, the concept of an existing scale would be called into question.

Sometimes examination of the items responded to unexpectedly by a subject reveals a difference between the outlook of this person and the outlook of those who can be measured by the scale. If we examine the responses of person 25, we see a possible pattern. This person, who agreed with most items, unexpectedly did not agree with items 4, 10, and 6, and unexpectedly agreed with items 20 and 17. Items 4, 10, and 6 deal with reading and use of library materials as opposed to the library itself. Items 20 and 17 deal with library use. This person appears to be exceptionally willing to see something as an overall library use, but apparently is much more conservative regarding statements concerning the use of library materials. While for most people these constitute a single scale, for some two scales seem to be required.

Our subjects were selected from three populations. Ranking the subjects by their score shows how the measurement of subjects relates to these three categories (library school students, librarians, and nonlibrarians). The librarians all scored in the top half of this ranking, an event that could have occurred one time in sixteen by chance. Students and nonlibrarians appeared in both halves of the list—at the very top as well as at the bottom.

In this study, we have not included enough nonlibrary school students to see any real group response patterns. If more had been tested, we would anticipate two possible patterns of response. One is that those

with library training would tend to score higher than those without, as this training might be sensitizing students to the information and community services libraries may offer. The other response possibility is that those with library training would score in a midrange, while those without such training would score in either extreme. This would reflect the idea that those with library training are open to all the information potential of a library—including chance encounters with library material whether or not that material is useful at the time of its discovery. These people, however, might not see noninformation seeking tasks, such as meeting friends or restroom use, as being related to the library's function. Nonlibrarians, on the other hand, might have a less sophisticated, more black and white attitude toward the library. They may see nearly every activity carried out in the library as a use, or they may see only successful, traditional activities as uses (resulting in either very high or very low scores). The nonlibrarians we questioned seem to fit this last pattern, though, because of their small numbers, this can readily be dismissed as a chance effect.

Conclusion

Our results indicate that: (1) a personality characteristic (or attitude), herein referred to as library sensitivity, does exist and can be scaled by means of Rasch analysis, and (2) this trait influences how one responds to questions about library activities. There are, of course, several ways in which the current scaling project could be extended.

As mentioned in the last section, our measurements could be improved by the development of additional items to fill in the gaps in the sensitivity scale. It might also be useful to see if new items relating to the sensitivity concept, but not following the "use" line of questioning, could be successfully integrated with the existing scale. These extensions should enhance the robustness of the scale, enabling its measurement to be more precise.

To check further the appropriateness of the model, we might test the consistency of item calibration as follows. Items could be divided into two equal groups—most of the easiest items in one group and most of the more difficult in the other. Each could be given to a different group of subjects. Using the response data, items could be recalibrated for each group. Some item values would be estimated twice—once for each of the two item sets. If the model fits, we would expect the difficulty estimates for each item measured twice in this test to be similar to each other, and, within group translations, to be similar to the estimates described in this article. Our concern here is whether context is influencing the responses.

Although the scaling model arose out of research on questionnaire design, and it was to illuminate the problem of ambiguity in questionnaires that led us to carry out this investigation, the current scale might also be applied in quite different lines of further study. Our scale could

be used to examine correlations between library sensitivity and a variety of demographic and sociometric variables. For example, we could test the relationship between library sensitivity and variables often thought to influence information-seeking behavior, such as education and income. We could also look at the scale values of men versus women, librarians versus nonlibrarians, etc.

As detailed in the Rasch Model Applications section of this article, the Rasch techniques demonstrated here could be extended to other problem solving applications in information studies. We suggest, for example, Rasch analysis of interindexer consistency. Do variations in indexing perhaps result, in part, from response differences between indexers regarding the concepts being indexed (as opposed to variations stemming from differences such as specialized subject training or level of indexing experience)? If so, can such perceptual differences be controlled through training in order to attain greater interindexer consistency?

We also suggest Rasch applications in the area of library acquisitions, using the model to develop peer groups of libraries for the purpose of comparing holdings and discovering subject coverage gaps. Thus, the Rasch methodology used here to analyze our questionnaire data appears to us to be a tool that can prove valuable for a wide range of investigations in the information sciences.

References

Bookstein, A. (1980). On the complexities of asking questions: Difficulties in an interpretation of library surveys. In N. K. Kaske, & W. G. Jones (Eds.), *Library effectiveness: A state of the art. Proceedings of the 1980 ALA preconference, 27-28 June.* Chicago, IL: ALA.

Bookstein, A. (1985). Questionnaire research in a library setting. *The Journal of Academic Librarianship, 11*(March), 24-28.

Bookstein, A. (1988). *Loglinear analysis of library data.* Dublin, OH: OCLC Report (Office of Research).

Bookstein, A., & Biggs, M. (1987). Rating higher education programs: The case of the 1986 White Survey. *Library Quarterly, 57*(October), 351-399.

Bradburn, N. M.; Sudman, S. and Associates. (1979). *Improving interview method and questionnaire design.* San Francisco, CA: Jossey-Bass.

Bradburn, N. M.; Rips, L.; & Shevell, S. (1987). Answering autobiographical questions: The impact of memory and inference on surveys. *Science, 236*(April 10), 157-161.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*(2), 81-105.

Engelhard, G., Jr. (1984). Thorndike, Thurstone, and Rasch: A comparison of their methods of scaling psychological and educational tests. *Applied Psychological Measurement, 8*(1), 21-38.

Goldhor, H. (1972). *An introduction to scientific research in librarianship.* Urbana-Champaign: University of Illinois, Graduate School of Library Science.

Guttman, L. (1944). A basis for scaling quantitative data. *American Sociological Review, 9*(2), 139-150.

Guttman, L. (1950). The basis for scalogram analysis. In S. A. Stouffer, L. Guttman, E. A. Suchman, P. F. Lazarsfeld, S. A. Star, & J. A. Clousen (Eds.). *Measurement and prediction* (pp. 60-90). New York: Wiley.

Hoyt, J. S., Jr. (1972). *Do quantifying adjectives mean the same thing to all people?* Minneapolis, MN: University of Minnesota, Agricultural Extension Service.

Hyman, H. H. et al. (1954). *Interviewing in social research.* Chicago: University of Chicago Press.

Kidder, L. H. (1981). *Selltiz, Wrightsman and Cook's research methods in social relations* (4th ed.). New York: Holt, Rinehart and Winston.

Kidston, J. S. (1985). The validity of questionnaire responses. *Library Quarterly, 55*(April), 133-150.

Kolata, G. (1987). How to ask about sex and get honest answers. *Science, 236*(April 24), 382.

Neter, J., & Waksberg, J. (1964). A study of response errors in expenditure data from household interviews. *Journal of the American Statistical Association, 59*(305), 18-55.

Payne, S. L. (1951). *The art of asking questions,* Princeton, NJ: Princeton University Press.

Pepper, S., & Prytulak, L. S. (1974). Sometimes frequently means seldom: Context effect in interpretation of quantitative expressions. *Journal of Research in Personality, 8*(June), 95-104.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests.* Chicago, IL: University of Chicago Press.

Schuman, H., & Scott, J. (1987). Problems in the use of survey questions to measure public opinion. *Science, 236*(May 22), 457-459.

Simpson, R. H. (1944). The specific meaning of certain terms indicating differing degrees of frequency. *Quarterly Journal of Speech, 30*(October), 328-330.

Sudman, S., & Bradburn, N. M. (1973). Effects of time and money factors on response in surveys. *Journal of the American Statistical Association, 68*(344), 805-815.

Thurstone, L. L. (1928). Attitudes can be measured. *American Journal of Sociology, 33*(4), 529-554.

White, H. S. (1987). Response to Bookstein and Biggs article. *Library Quarterly, 57*(4), 396-398.

Wright, B. D., & Panchapakesan, N. (1969). A procedure for sample-free item analysis. *Educational and Psychological Measurement, 29*(1), 23-48.

Wright, B. D., & Douglas, G. A. (1977). Best procedures for sample-free item analysis. *Applied Psychological Measurement, 1*(Spring), 281-295.

Wright, B. D., & Masters, G. N. (1981). The measurement of knowledge and attitude. *Research Memorandum,* MESA Psychometric Laboratory, University of Chicago, Department of Education.

Wright, B. D.; Rossner, M.; & Congdon, R. T. (1987). *MSCALE-A Rasch program for ordered categories* (Computer program). Chicago, IL: University of Chicago, Department of Education (March).

Wright, B. D., & Mead, R. J. (1977). BICAL: Calibrating items and scales with the Rasch Model. *Research Memorandum,* University of Chicago, Department of Education, *23*(January), 6.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis.* Chicago, IL: MESA Press.