American Journal of Pharmaceutical Education **2005; 69 (5) Article 97.**

SPECIAL ARTICLES

Construction, Implementation, and Analysis of Summated Rating Attitude Scales

Shane P. Desselle, PhD

Mylan School of Pharmacy, Duquesne University

Submitted December 7, 2004; accepted February 13, 2005; published December 9, 2005.

Surveys are perhaps the most used, and sometimes misused, methodological tool among academic researchers. Self-administered survey questionnaires (and guided interviews) allow us to elicit behaviors, attributes, beliefs, and attitudes among populations. Data gathered from poorly designed instrumentation is suspect and investigators are disappointed when research involving the use of attitude scales is viewed unfavorably by peer reviewers. Summated rating scales may be commonly used to elicit feelings or attitudes from students, faculty members, administrators, and others. Investigators must be conscious of the referent object and the research objectives when constructing items comprising a summated ratings scale. Items should be gathered from careful review of the literature and consultation with experts. Pilot testing a survey instrument and its component attitude scale are critical to success. Adherence to a few basic tenets will help to ensure that responses to the scale are valid and reliable. There are a number of means by which a scale's validity and reliability can be determined. While ordinal in nature, data gathered from summated rating scales may be analyzed with robust, parametric statistics that allow investigators to tackle the multivariate nature of many problems in social, administrative, and clinical research. Novices in questionnaire survey design and analysis are encouraged to consult with a number of sources who may be helpful to them, including survey methodologists and statisticians.

Keywords: attitude scale, Likert-type scale, semantic differential scale, survey research

INTRODUCTION

The Utility of Survey Questionnaires

In the social and medical sciences, and certainly in research on teaching pedagogy and design, important measurements are based on a question-and-answer process. Studies of medical outcomes rely on patients' answers to questions about their health status and quality of life; patients respond to queries estimating their compliance to prescribed therapies; students provide assessments of faculty members' teaching effectiveness, from which promotion and tenure decisions are based; faculty members respond to questionnaire surveys that elicit data pertaining to the types of information covered in various courses in hopes of standardization and quality assurance in the education process.

There is an almost limitless body of desirable and useful information that can be gathered only by asking people questions. Some persons, even within the field of psychology, question the validity and utility of direct questioning for acquiring data. Opponents point to a reluctance among individuals to express their feelings or attitudes about certain issues. They suggest also that some individuals may not be aware of their feelings toward a given psychological object. Those who profess great dislike of something may be reacting against unconscious impulses of the opposite nature (eg, the young male teenager who "hates" girls).[1] An individual's feelings about a psychological object may be mixed, making them difficult to evaluate. Other problems include some individuals' inclinations to provide socially desirable answers or even attempt to strategically bias a survey's results. Sampling bias and poor response rate may also limit the utility of survey research.

No research method, though, is without its limitations. Even direct observation of behaviors can provide results of limited value. In some cases, that is because we want to know facts that are difficult to observe systematically. For example, a student who grows dissatisfied with his/her education at a particular school of pharmacy by the third-professional year of the program is not likely to quit

Corresponding Author: Shane P. Desselle, PhD. Address: Director, Office of Assessment & Educational Strategies, Mylan School of Pharmacy of Duquesne University, 600 Forbes Avenue, Pittsburgh, PA 15282. Tel: 412-396-6363. Fax: 412-396-5130. E-mail: desselle@duq.edu.

or change majors or universities. Thus, relying on dropout rates to measure students' attitudes about a program would produce specious results. When constructed, analyzed, and interpreted well, survey instruments in general, and attitude scales in particular, can provide invaluable information to researchers, teachers, administrators, and students.

Data Gathered from Surveys

Surveys may be used to elicit information about individuals' attributes, behaviors, beliefs, and attitudes. Personal and demographic characteristics (eg, educational degrees, current grade level or year in a PharmD program, academic rank, age, and sex) are respondent attributes elicited in most survey research. The usual purpose for collecting this information is to explore how persons may differ in behaviors, beliefs, or attitudes along these characteristics. Other attributes that may be measured include certain predispositions or traits. Examples of traits used in education research include, but are not limited to, intelligence, aptitude, performance anxiety, learning style, self-efficacy, moral reasoning, critical thinking, and leadership ability. These types of variables may be the primary focus of research or may be factors that must be accounted or controlled for in experiments seeking to determine the effect or impact of an educational intervention.

Beliefs are what a person typically holds as truths. There is not implied goodness or badness in beliefs, but only an assessment of what one thinks exists or does not exist.[2] Beliefs may be elicited in the form of summated ratings scales discussed in this article, such as with measures of "favorability," "opposition," "agreement," or "perceived importance." They also may be elicited with questions using dichotomous response sets, such as "yes/no" or "agree/disagree." In essence, tests of knowledge constitute measuring beliefs, in this case, beliefs about what is "correct." As such, objective tests may take on a variety of forms, including essay, fill-in-the-blank, true/false, and multiple-choice, to name a few. Researchers may elicit beliefs about a host of other matters, such as the need for curricular reform, alternative strategies for making tenure and promotion decisions, or types of conduct that are deemed in violation of standards of academic integrity.

Reports of behaviors are commonly elicited through surveys; strictly speaking, these questions acquire individuals' beliefs about their behavior. Thus, behavior should not be considered a distinct type of information. However, there is a substantial difference between asking people to describe their own behavior and asking for their view about something they have experienced only in a cognitive (as opposed to a physical) sense.[3] Behavioral questions may concern what people have done in the past (eg, students' study habits or cheating behaviors, faculty members' track records of publications and grants), what they are currently doing (eg, student participation in extracurricular activities, faculty members' research interests or teaching styles), or what they plan to do in the future (eg, faculty members' intent to stay/remain with the current employer, students' intent to enter a particular type of setting to begin a pharmacy career, or administrators' intent to change leadership styles).

Most frequently in education research, we are interested in measuring the attitudes of students, faculty members, or administrators. An attitude is an organized predisposition to think, feel, perceive, and behave toward a referent or cognitive object.[4] They are evaluative in nature and reflect respondents' views about the desirability of something. Attitude questions require respondents to show whether they have positive or negative feelings about the "referent object."

A referent object can be any symbol, phrase, slogan, person, institution, ideal, idea, phenomenon, or construct toward which persons have an objective reference, differing from a trait, which has a subjective reference.[1] For example, one who has a hostile attitude toward foreigners may be hostile only to foreigners, but one who bears the trait "hostility" is prospectively hostile to a much broader array of persons.[4]

MEASUREMENT

Issues of Measurement

Questions and answers are part of everyday conversation; however, the researcher must turn an everyday process into a means of rigorous measurement. The fact that the answers to questions comprising a survey are used to measure the phenomenon of interest has at least 2 important implications. First, the researcher is not interested in the answers for his own sake. As a result, a critical standard for a good measure is that it produces meaningful information about what he is trying to describe (validity). Second, the purpose of measurements usually is to produce comparable information about many people or events. Hence, it is important that the measurement process, when applied repeatedly, produces consistent results (reliability).

Although much has been written on the nature of measurement, it simply comes down to rules for assigning numbers to objects in such a way as to represent quantities of attributes.[4] The term "rules" indicates that the procedures for assigning values or numbers must be explicitly stated. In some instances the rules are so obvious that detailed formulations are not required. Certainly, the rules for measuring most psychological phenomena are not

intuitively obvious, eg, students' motivations for learning, communication apprehension, and civic engagement; or professors' lecture styles, preferences for curriculum design, and attitudes toward mentoring.

The creation of scales used in psychological measurement involves the assignment of numbers to objects to represent quantities of attributes. It is necessary to have some internally consistent plan for the development of a measure. The plan is spoken of as a scaling model, and the measure that results from exercising the plan is referred to as a scale (scale being another word for measurement method).[5] The simplest example is that of the ruler as a scale of length. The methods for constructing and applying rulers constitute the scaling model in that case. The purpose of any scaling model is to generate one or more continua on which persons or objects are located.[5]

Some methods of scaling assume that persons are replicates of one another. For example, the percentage of persons in a group that says one teaching method is more effective than another is assumed to be the same as the percentage of times an ideal model individual would say that one teaching method is more effective than another on different occasions. The assumption that individuals are replicates is frequently made in scaling stimuli. In scaling persons, it is frequently assumed that responses are replicates of one another. Thus, an attitude or degree of "favorableness" can be obtained by adding responses over the separate rating scales.[5]

Measurement Issues/Use of Summated Ratings Scales

There are 3 principal types of attitude scales: interval scales, cumulative (or Guttman), and summated rating scales. Most commonly employed in education research are summated rating scales. A summated rating scale is comprised of a set of attitude items, all of which are considered of approximately equal "attitude value," and to each of which participants respond with varying degrees of intensity (eg, 5 to 7 points) on ordinal measures.[4] The scores of the items on such a scale are summed, or summed and averaged, to yield an individual's attitude score. The purpose of the summated rating scale is to place an individual somewhere on a continuum of the attitude in question. Allowing the individual to express intensity (on a multipoint scale) allows for greater variance and precision among responses and, in many cases, the ability to employ more robust statistical testing procedures. One problem, however, is that the tendency among respondents to use certain types of responses differently from one another (eg, "slightly," "very," "strongly," "completely") produces *response set variance*. That said, response set variance is considered only a mild threat to valid measurement, and its importance is considered overrated.[5]

Interval scales can also be used to assign attitude scores to individuals, and these scales accomplish an additional task of scaling the items comprising the scale, itself. Each item, through a series of judgment processes in equal-appearing intervals or paired comparisons procedures, is assigned a scale value that indicates the strength of attitude of an agreement response to the item.

A cumulative or Guttman scale consists of a relatively small set of homogenous items that are measured with dichotomous response sets. It is used more frequently to assess beliefs or knowledge. The scale is created from items in some natural progression, such as level of difficulty for an examination, wherein a difficult item or problem would be placed first, followed by a less difficult problem, followed by an even easier one. One would suspect that few students who missed the first problem would miss either of the latter 2. The responses to a Guttman scale are used to rank individuals. The Guttman scale has been highly criticized by psychologists, particularly in the measurement of attitudes.[4] The consensus, then, is that summated rating scales provide the researcher with the most varied and effective toolbox from which to elicit attitudinal responses. There are 2 types of summated ratings scales: the Likert-type scale and the semantic differential scale. The Likert-type scale is more commonly used, yet somewhat more difficult to create, analyze, and interpret.

Likert-type Scales

Likert-type scales (henceforth, referred to as "attitude scales") are so called because they are a derivation of a scaling procedure developed by Rensis Likert,[6] whose original procedure was designed to collect interval-level data. Attitude scales of this sort typically are comprised of a set of statements or "items" that scale a respondent's level of agreement, favorability, or other similar perception. The class of all possible items that could be made about a given referent object can be called a "universe of content," describing possible stimuli from which attitudes toward that object may arise.[1] While it is highly possible that someone may have a favorable overall impression of a given object, yet be unfavorable about a particular aspect or dimension of it, and vice versa, any item statement used in a scale should be useful in differentiating between persons with favorable and unfavorable attitudes. Following are things to consider when creating an attitude scale.

Defining the Study's Objective

All too often, novices in questionnaire survey construction and design initiate projects involving the use of attitude scales without following basic tenets of scientific inquiry that they might never violate within the scope

of their expertise. Even if a project is undertaken to answer a specific question for "applied" purposes, and not necessarily to test a theory or model, adherence to the scientific process will yield the best results. If the prevailing attitude about a project is anything like, "Oh, we just need to piece together a survey instrument as quickly as we can," the investigators fall prey to the same mistakes that laypersons make when observing phenomena, such as overgeneralization, selective observation, inaccurate observation, illogical reasoning, and ex post facto hypothesizing (aka, "See, I told you so").[7] The scientific process involves stating and shaping the problem, generating relevant and testable hypotheses, reasoning and deduction, observation or testing, and analysis of the results.[4]

As elementary as it may appear, some investigators fall short of clearly shaping and stating the research problem. For example, school administrators may state simply that they want to know whether students are satisfied with their education. One danger here is that "satisfaction with education" is a complicated issue and rooted in perceptions about the curriculum, the faculty members, administration, campus life, housing, meals, financial aid, facilities, and available technology, to name but a few factors that shape this attitude. A lack of clearly stated research objectives will manifest in poorly written survey items that will in all likelihood fail to answer the questions that the administrator had in mind. Clearly defined research objectives will also provide investigators with the types of information (behavioral and attribute questions) that they must solicit from respondents in addition to responses to the attitude scale. Investigators often take the stance, "Well, if we're going through the trouble of administering the survey, we might as well ask them this question as well. Keeping questionnaire surveys brief is difficult enough when only the information needed to fulfill the research objectives is solicited. Moreover, prospective respondents try, at least on some level, to discern the rationale for the questions comprising the survey questionnaire and may become disinterested or offended by an instrument that is too lengthy or contains questions they view as unnecessary or otherwise inappropriate.

STUDY DESIGN

Research Design and Sampling Issues

Prior to constructing the questionnaire survey, the investigators must consider aspects of the study's design and sampling procedures. The choice of design will confer varying degrees of confidence that the observed results are not due to confounders and that they may be extrapolated beyond the study sample. The design will also determine whether the results can be used to explain, correlate, or predict the occurrence of one or more phenomena. It is extraordinarily important to consider the population of interest, ie, prospective respondents, prior to constructing questionnaire survey items. Care must be taken to ensure that the research is meaningful to them, that the items are constructed to elicit views from their perspective, and that the survey questionnaire uses appropriate language and is written at a level they can understand.

Scale Values

An attitude scale measures intensity of a feeling or perception. The measurement of intensity is useful because strength of feeling predict both attitude stability and attitude constraint.[8] Thus, eliciting intensity can help identify respondents whose responses will be more consistent over time.[9] Asking questions about intensity or centrality also may enhance the investigators' understanding of the nature of opinion on an issue. For example, research on abortion has shown Americans to be fairly split on whether a married woman should have access to a legal abortion; however, abortion opponents were 6 times more likely than "pro-choice" advocates to indicate extremely strong sentiments about the issue.[8] Similarly, a preponderance of students may regard an instructor as relatively effective, but a significant minority of students may believe that the instructor is extremely ineffective. These are unique sets of information (plurality, intensity), but may be equally useful to the instructor and administrator. The choice of which information is more relevant and should be presented as the study's results depends upon the research objectives.

Argument has persisted over the number of values or "points" comprising a response scale (ie, that are used with individual items that are combined to form the summated scale), but the majority of researchers agree that 5 to 7 points allows respondents a wide enough range of intensities from which to choose. Having 10-11 points for certain measures is viewed as acceptable, particularly for global assessments and within clinical arenas where respondents are asked to express a level of pain, discomfort, or raw emotion.[4] Certainly, a response scale comprising any more than 10-11 points creates an intrinsic level of precision to which the majority of respondents are not able to adhere.

Investigators must be keenly aware that the choice of response scale anchors, values, and labels could affect the results. Whether to employ an even or odd number of scale values must also be considered. The argument for an odd number of choices is that it permits the use of a middle value such as "neutral," "neither agree nor disagree," or "no opinion." This is thought to make subjects more comfortable in providing ratings, and subjects frequently have neutral reactions that should be measured.[5]

Nunnally argues, however, that the use of a neutral value introduces response styles.[5] Some subjects tend to use the neutral step more frequently than others, and individual differences in that regard might not relate highly to the attitude in question. A respondent who provides a neutral response consistently for much of the items may be sending a message they did not care to participate in the study. Moreover, reliable differentiations can be made among persons who mark the neutral value. One person may have little opinion in regards to a particular item statement, while another may hold a middle of the road position with considerable passion.[8] Having a ''no opinion'' or even a ''not applicable'' option is left to the discretion of the investigators and is highly dependent upon the research objectives and the nature of the referent object. ''Not applicable'' and ''no opinion'' are entirely unique from having a ''neutral'' opinion. Having these options may be necessary when eliciting beliefs or behaviors, or when an attitude requires exposure to some aspect of the stimulus to which many of the respondents may not have had, and as such, should not be scored as a neutral opinion (which is typically the midpoint of a scale).

Another issue to consider is whether to use numerical or adjectival labels for response scale values. The principal argument in favor of adjectival scales is that all of the points are more consistently evaluated by the use of words. Depending on the nature of the items and the referent object, the use of numerical labels may have some respondents treating the response scale like a thermometer or ''feelings ladder,'' rating varying degrees of ''positiveness'' rather than going through a midpoint toward a negative rating.[3] When utilizing numeric labels, one also should consider the response scale's anchors. Responses to a scale ranging from -5 to +5 will be more positive than responses to identical items on a scale ranging from 0 to 10.[10] The choice of anchor may depend partly on whether the investigators are attempting to elicit potential negative affective responses or simply determine the strength of agreement with item statements.

It is important that the response scale be balanced when utilizing adjectival labels. Consider the following response scale: very satisfied, somewhat satisfied, satisfied, not satisfied. Many observers would contend that ''somewhat satisfied'' is a less positive category than ''satisfied.'' If some respondents concur, the assumption that the responses are ordinal would be violated, because each term or label is assigned a value, which is used for statistical evaluations. Thus, the order of adjectival labels must be unambiguous. Additionally, the intervals between each value must be as equal as possible. The following represents a significant improvement: completely disagree, mostly disagree, slightly disagree, slightly agree, mostly agree, completely agree.

Consider also a scale that measures frequency: always, usually, often, sometimes, never. The scale's midpoint is ''often,'' which would not appear indicative of moderation in frequency. Moreover, the interval between sometimes and never may be larger than other intervals in the scale. The scale would be more balanced if a label such as ''seldom'' were placed in between ''sometimes'' and ''never.'' Care should be taken in semantics issues over the adjectival labels regardless of what is being scaled (eg, agreement, favorability, frequency). Words like ''fairly,'' ''very,'' ''most,'' and ''somewhat,'' to name a few, must be carefully scrutinized for unintended meanings, ambiguity, and improper scale balance.

There is at least some support for the use of forced-choice questions as opposed to the use of an agree-disagree format. Lenski and Leggett[11] proffered 2 items that would appear quite contradictory:

1. ''It is hardly fair to bring children into the world, the way things look for the future.''
2. ''Children born today have a wonderful future to look forward to.''

Approximately 1 in 10 respondents of their sample agreed to both statements. This has been attributed to a tendency among some respondents to agree irrespective of item content (that being said, this author believes that the leading and compound nature of the first item invites the occasional unreliable response, which I will elaborate upon later). This tendency has been demonstrated to be more prevalent among those with less formal education.[3] A student may plausibly agree with both of the following statements:

1. ''The university should see to it that every student receives a proper education.''
2. ''Students should be responsible for their own educational outcomes.''

A forced choice format in the above situation would have the respondent indicate on a scale which of the 2 parties was more responsible for educational outcomes and might look something like this:

__ Entirely the student
__ Mostly the student and partly the university
__ Equally shared among the student and university
__ Mostly the university and partly the student
__ Entirely the university

However, this may not alleviate the problem considerably because many respondents may be reluctant to place the responsibility entirely on either party. Additionally, the previous 2-statement items are not antithetical to one another (ie, it may be possible to find some agreement with both), and they too violate basic tenets of good item construction.

Nunnally argues, then, that decisions on whether to include neutral values and on the number of values (points) to include in the response scale are not the most important issues, particularly if scores are summated over a number of items. The use of a forced-choice or other formats may not be necessary. What is unequivocally most important is that the items comprising the attitude scale are properly constructed.[5]

Developing Items to Comprise a Likert-type Scale

The response stimuli (captured by item statements) comprising an attitude scale should be representative of the universe of existing stimuli to that referent object. In other words, researchers need to consider all aspects of a phenomenon that may constitute a person's attitude towards it. For example, investigators gathering students' attitudes about use of a new technology in a course need to consider everything about that technology that could evoke an affective response. This would include, but not necessarily be limited to the technology's access, convenience, navigability or ease of use; enjoyment of use; cost; compatibility with the course format; assistance in helping students to meet course objectives; use as a study aid; appeal to various learning styles; and time-savings capabilities. More often than not, certain stimuli comprising attitudes toward the referent objective may not be readily apparent. Failure to capture at least the majority of possible stimuli will compromise the scale's validity. Developing a comprehensive and appropriate set of items requires considerable commitment from the investigators.

Sources for the Development of Items

There are at least 3 sources to consider for identifying potential items. One is consultation of the literature, particularly primary literature. A thorough literature review will provide considerable insight into the referent object and invariably bring to light unique aspects and approaches to the topic not previously considered by the investigators. The review also might help determine whether the investigators' goals are too ambitious or whether the research problem is not one that lends itself well to investigation as currently structured. On the other hand, it may assist the investigators in identifying a previously validated scale or other tool that can be used to measure the referent object more reliably and to avoid the burden of creating an entirely new scale.

The literature search should be comprehensive. For example, if an investigator is interested in identifying innovative curriculum designs, the literature review cannot be confined to merely plugging in the terms ''curriculum'' and ''design'' into a search engine to see what this yields. Just like a thorough literature review in a researcher's own area of expertise, a thorough literature review used to guide good survey research should take days to weeks and involve initial queries with a comprehensive set of search terms in a variety of databases, and refining subsequent searches based upon findings in the initial search. Depending upon the nature of the research project and the referent object, it may be necessary to conduct searches on International Pharmaceutical Abstracts, Medline, ERIC, PsycInfo, Social Sciences Index, Social Sciences Citations, ProQuest, and CINAHL, to name a few.

A second source for identifying items is that of fellow colleagues and experts in the field. Much can be gained from conversations with colleagues, administrators, and consultants. Aside from acquiring their unique perspectives, the process of discussing the project in and of itself is stimulating. While the expert may have published extensively on the subject already, a discussion with him or her may reveal additional information not yet in print or provide the same sort of stimulation derived from talking with colleagues. It also may result in an additional collaborator.

A third source comprises persons meeting the sampling criteria. Referring to the previous example concerning students' attitudes about technology, gathering affective responses from a small group of students about the technology will prove invaluable. A research project concerned with measuring faculty members' opinions about workload, for example, should involve consultation with colleagues within and outside the researcher's discipline to gain a sense of the types of activities in which other faculty members are involved, which may be unique from activities familiar to the investigators. Gathering this information simply may involve interviewing a small number of persons chosen on the basis of convenience. A more formal and potentially rich data source for soliciting domains of the referent object involves the use of focus groups or nominal group techniques.[11-13]

Identifying Relevant Domains

The aforementioned procedures will assist the investigators in identifying specific items, some of which may be entire domains of the referent object. This is especially true for complex, theoretical research of referent objects that are multidimensional in nature. Specific items would then have to be developed for each domain. Each domain should consist of several items. Generating a greater number of item stimuli will increase the scale's reliability.

Writing Good Item Stimuli

Table 1 summarizes some guidelines for constructing items.[14] The most fundamental principle for constructing good items is to bear in mind the intended respondents.

Table 1. Suggestions for Writing Items Comprising an Attitudinal Scale

- Use appropriate language. Avoid technical jargon and abbreviations.
- Design the items at the appropriate level of reading comprehension.
- Avoid the use of vague or ambiguous terms.
- Avoid the use of terms that elicit bias on the part of respondents.
- Avoid questions that will engender socially desirable answers.
- Provide contextual material judiciously and only when necessary.
- Avoid double-barreled items or items with more than one meaning.
- Avoid the use of superlatives and unnecessary descriptors.
- Carefully intersperse negatively worded statements or those that require reverse coding.
- Construct the items in a form that make them as personally relevant to the respondent as possible.
- Do not elicit content knowledge in an item that purportedly measures attitudes or feelings.
- Choose relevant time periods to achieve accuracy in responses.

The items should use language and terminology familiar to the respondents while maintaining proper grammar and punctuation. For example, in a pretest to acquire students' knowledge prior to implementing an educational intervention, a pharmacology instructor should not utilize the abbreviation "GPCR" (G-protein coupled receptor), or in an attempt to measure students' attitudes about managed care, an instructor in social and administrative sciences should not use the abbreviation "PBM" (pharmacy benefits manager). In the former case, it is knowledge of the abbreviation and not knowledge of the receptor that is tested; in the latter, students would not know how to respond and would probably either leave the question blank or provide a response at the scale's midpoint. Similarly, while many of us in academia would prefer that our students' have better vocabularies, using words that a number of respondents might fail to understand (eg, "ostensibly" or "perfunctory") is not prudent. Less common but perhaps somewhat problematic would be to construct items in a manner that would insult the intelligence of the respondent.

Examples of vague item statements are "How are things coming along with your academic career?" or "How has your academic progress been?" "Coming along" is ambiguous. The respondent does not know whether this refers to a general level of satisfaction, to research/teaching productivity, progress toward tenure and promotion, or something else. "Academic progress" is a nebulous term that could be inferred as meaning that the respondent is on pace to graduate on time, has improved from previously poor academic standing, or is learning substantially during his or her studies.

Biasing words and phrases are those that elicit emotional responses that have little to do with the referent object. The investigator has to be careful when using a term like "academic freedom" when designing survey items. Additionally, words such as "fair" can take on any of a variety of meanings, such as just, equitable, impartial, or not very good, and should be used judiciously and with caution, if not entirely avoided. The investigator must be objective and resist the temptation to agitate respondents who may end up responding more to the survey itself than to the referent object. On the other hand, investigators must avoid eliciting socially desirable responses. For example, if inquiring about an instructor's self-reported teaching effectiveness, an item phrased, "I put as much effort as do other faculty in my teaching," or "I believe that I am an effective teacher" places the respondent in a compromising position, almost forcing him or her to provide an affirmative response. In this case, more specific and reliable answers would be obtained if the respondent were asked to report the frequency in which they engage in certain behaviors, employ specific strategies, and perhaps how they might respond in certain teaching situations.

It may be appropriate on rare occasions to accompany 1 or more items with some contextual information, particularly if there is concern that respondents might provide socially desirable answers. For example, one might consider informing potential faculty respondents that "some instructors may utilize technology effectively in the classroom, while other instructors may effectively teach without substantial use of technology," before eliciting attitudes about the use of technology. This technique should be used sparingly, however, as too much information increases response burden and may result in inappropriately leading the respondent or complicating the instructions.

A common mistake even among more experienced researchers is to construct compound or double-barreled item statements. For example, consider the item, "I agree with the general direction that my department and my school are heading." The respondent might disagree with the direction that the department is taking, but agree with that of the school. An item worded in this fashion places respondents in a quandary, resulting in them leaving the question blank or perhaps "averaging" their feelings to the 2 items when selecting a response. The remedy for this is simply to divide the 1 item into 2 unique items.

Investigators may be under the false impression that the use of superlatives somehow enhances the clarity of an item, but more often than not, the opposite is true. An item phrased, "This is the best elective course I have ever had," creates problems on several fronts. For one, a word like "best" is ambiguous and could be inferred to mean easiest, the one wherein the instructor is liked most, the one in which the greatest amount of learning took place, an so on. Secondly, the respondent may have had little experience with other courses that could be used as a basis for comparison. Finally, the respondent could be relatively satisfied with the course, but recall another course that was preferred. As currently stated, the respondent would have little choice but to disagree or strongly disagree with the item, in spite of their satisfaction with the course. Similar problems occur when adding descriptors like "very" and "quite" to items.

Investigators should consider reversing the effect of some of the item statements. For example, if most of the items represent something positive about the referent object, inject a few items that point to something negative about it. If a respondent is agreeing or disagreeing with all of the statements whether good or bad, this tends to indicate that he or she is really not paying attention to the stimuli, perhaps warranting his or her responses to be discarded. If this occurs frequently, the investigator must question the validity of the scale, the directions provided to respondents for completing the survey instrument, and/or the sampling procedures. That being said, there are arguments against this procedure, as disagreement with a negatively worded item is not necessarily the same as agreement with a positively worded item, and vice versa. Therefore, the use of this procedure should be determined on a case-by-case basis and depends upon the nature of the referent object and the intended respondents.

Making the items as personally relevant as possible will improve the reliability and validity of the responses. For example, if investigators are interested in eliciting student attitudes toward an increase in the grade point average required to remain in the PharmD program, an item such as, "Because of the new GPA requirement, students are likely to study even harder," induces speculation by the respondent regarding the habits of other individuals. Rewritten, "I will improve my study habits as a result of the new GPA requirement," requires the respondent only to predict his or her own behavior. Similarly, asking students to rate the "convenience" of a new practice laboratory's hours of operation is inferior to asking them to respond to an item stimulus such as: "The Practice laboartory's hours of operation make it accessible for me to use." The latter is clearer and may provide more specific information to discern preferences among different types of persons (eg, by sex, student classification, age, whether the respondent has an off-campus job).

Including factual statements in a scale designed to elicit attitudes or feelings creates several problems. Favorable responses toward an item may simply mean those respondents were more familiar with the subject matter of a questionnaire item than were respondents who expressed unfavorable attitudes, or vice versa. If a statement is equally likely to be endorsed by those with favorable versus those with unfavorable attitudes, then it is not useful in differentiating respondents. Similarly, if an item requires knowledge that the respondent may not have (eg, asking students about the frequency in which tenured faculty members are evaluated by students), the responses will be unreliable.

Investigators may be interested in eliciting from respondents a frequency or timeframe for a particular behavior, for example, the number of peer-reviewed articles one has published, the number of hours spent utilizing a Web-based tutorial, or a general state of well-being or level of satisfaction. A time period that is overly brief such as "yesterday" is transient and not necessarily indicative of someone's typical behavior or state of mind. On the other hand, asking a respondent to provide information about an entire year, or perhaps even an entire academic semester requires significant recall and may involve attitudes or behaviors that fluctuated during this timeframe. Given that the research objectives differ from one study to the next, there is no universally accepted timeframe; however, periods such as "the past 30 days," "the past 4 weeks," or "the past 2 months" are often appropriate and may yield more reliable responses.

Semantic Differential Scales and Other Question Formats

An alternative to the Likert-type format is the semantic differential scale. Rather than responding to a set of items on 1 scale, the semantic differential acquires responses to a number of scales anchored by a set of bipolar descriptors that putatively describe the reference object. Each scale usually is scored on a 5 to 7 point basis, just as a Likert-type scale is scored. The descriptors are usually one-word adjectives or short phrases. For example, investigators may be interested in obtaining student perceptions of the faculty members, in general. They might ask respondents to rate the faculty members on the type of scales shown in Figure 1.

Semantic differential scales have a number of appealing features. They are easily adapted to a number of concepts and easy to apply to one concept in the same rating form or survey. Moreover, they have intuitive appeal because characteristics of various phenomena (objects)

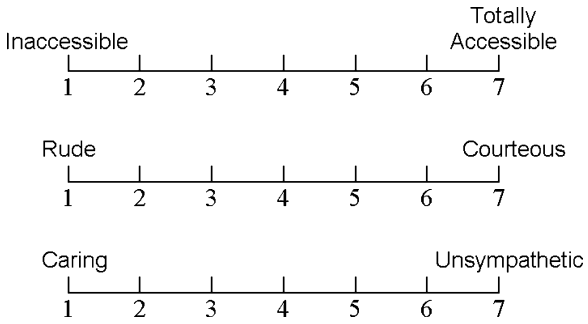

Figure 1. Examples of semantic differential scales.

are communicated largely by adjectives. For most adjectives there are words that are relatively logical opposites, or for which an opposite can be created by adding a prefix such as "in-" or "un-." Of course, care has to be taken in the selection of bipolar words or phrases to ensure that they are indeed as antithetical or as "opposite" as possible, that they are unambiguous, well known to the respondents, and can be appropriately and logically used to describe the object. Investigators also must be aware of how to interpret the results. An indication that a faculty member is "unsympathetic" is perhaps stronger or has a slightly different meaning than a respondent's disagreement on a Likert-type scale that the faculty member is "caring." For more on semantic differential scales, see Nunnally[5] and Snider and Osgood.[15]

Investigators may also consider using questions that elicit rank data. This may be useful when there is concern that respondents might rate all items on a Likert-type scale similarly. For example, if investigators were interested in having faculty members evaluate the importance of various topics that should be covered in particular courses, they might be concerned that faculty respondents would provide a favorable rating to each and every topic. As an alternative, the investigators might ask the respondents to rank the topics in order of importance. There are considerable drawbacks to this method, however, as the cognitive load might become too great if the number of items to be ranked are too numerous.[4] Additionally, the use of rank data limits the use of many of the more powerful statistical procedures available to analyze the data, because rank data elements are not independent and not normally distributed. Other types of scaling methods exist, such as the Q Sort.[5]

Pilot Testing

Investigators should ascertain the face validity of the attitude scale by obtaining expert advice from colleagues familiar with the subject area and from trusted survey methodologists. After the items have been generated, the investigators can construct a draft of the survey questionnaire (or interview) that includes the attitude scale along with other questions needed to meet the research objectives. There are excellent resources to help investigators with issues such as formatting, font size, paper color, use of graphic illustrations, and providing adequate directions to respondents for completing the survey.[2,14,16]

The investigators should pilot test the initial draft of the survey questionnaire on persons similar to those for whom it is intended. While there is no hard and fast rule, it is a good idea to disseminate at least 20 to 30 survey questionnaires during the pilot. Pilot testing can be accomplished through convenience sampling to individuals or through the use of additional focus group procedures. A poor response rate, an abundance of questions left blank, and/or unsolicited comments in the margins of the questionnaire surveys during pilot testing would indicate the need to reword or remove certain questions. Pilot testing also will determine whether the questionnaire survey is too long and burdensome to complete.

A Few More Words About Social Desirability

One of the main concerns among opponents and proponents of survey research is the tendency among certain individuals to provide socially desirable responses, in other words, responses that would appear to put themselves in a better light or responses that they believe the researchers desire. There are individual differences in style when responding to self-inventories;[5] however, this does not necessarily mean that "frankness" in responses differs among populations of individuals in a manner that would affect the results of inferential statistical tests. One important mechanism to mitigate the problem of socially desirable responses is careful construction of the survey questions. Alternatively, investigators may consider using one or more instruments that actually measure frankness in response, such as the Marlow-Crowne Social Desirability Scale.[17]

ANALYZING DATA

Validity and Reliability

After the survey instruments have been returned (or interviews completed), the investigators should determine the usable rate of return and assess the potential for non-response bias (not discussed, here). At this point, the process of scale development and purification is still not complete. Prior to the use of any inferential statistics, the investigators must take steps to verify the scale's validity and reliability. A statistical procedure called *factor analysis* utilizes the covariance existing between responses to the items to group them together into "factors" or domains. This enables the investigator to determine whether items load onto the domains previously

hypothesized when constructing the scale. It may also reveal that certain items did not load onto any of the domains, thus calling to question whether these items should be retained.

Validity within the context of survey research concerns whether the scale measures what it purports to measure. The aforementioned face validity check is a somewhat crude measure of content validity, which gets at whether the scale's content is representative of the referent object's universe of content that may describe it.[4] Although seldom attainable, it is desirable to determine a measure's *criterion-related validity*. Criterion-related validity can be thought of as comprising concurrent validity and predictive validity. To ascertain a scale's concurrent validity, it would be administered concurrently and compared with some alternative measure of the same construct (perhaps an action or behavior). The scale's predictive validity would be determined through its ability to "predict" an outcome, such as successful performance of a task or job. Criterion-related validity is of greater concern when designing objective tests or tests of knowledge more so than with attitude scales.

Most critical in attitudinal research, particularly research that is theoretical in nature, is determination of a scale's construct validity. Results from a factor analysis procedure assist with building a case for or against a scale's construct validity. The investigators should conduct a "principal components" factor analysis procedure (which infers that it is exploratory in nature and not forced into a particular number of factors or iterations), usually with an orthogonal (Varimax) rotation of the factor loadings that results in a simpler, more interpretable factor structure.[4,5] The output for this procedure yields the number of factors or domains comprising the construct, the amount of variation in responses to the scale explained by these factors, and the loadings of each item. A factor is comprised of the items that loaded highest upon it; however, some items may not load well onto any factor (any item that fails to yield any loading above the absolute value of 0.50 or 0.60 may be considered for removal, pending subsequent item analysis[18]). The resulting factors, if more than one, essentially should "make sense." There should be some discernable pattern of the items that load together; moreover, the factors should help to explain much of the covariance among the items. Factor analysis is known as a "data reduction" procedure. Having taken a scale that comprised for example, 20-30 items and reduced them to 2-4 factors or domains, the investigators can name these factors and express that respondents' attitudes toward the referent object are basically a function of "Factor 1 (name)," "Factor 2 (name)," and so on. Most statistical software packages allow researchers to save "factor scores" for each respondent as variables for subsequent analyses. While this is recommended, the investigators may alternatively create unique subscales by summing the responses only to those items comprising that subscale (factor). Responses to the subscales may then be used in subsequent inferential statistical procedures. Factor analysis is a useful but complex and sometime misused statistical procedure. One often overlooked requirement for factor analysis is that the data should be split into 2 random subsamples. The first subsample is used to find a conceptually plausible structure for the data, while the second subsample serves as a validation sample.[4,5,18]

Investigators may be in a position to compile even greater evidence for a scale's validity by examining it across 2 components of construct validity: convergent validity and discriminant validity. Responses to a scale demonstrating convergent validity should be highly correlated with responses to a scale measuring a similar construct or one that is theoretically proposed to be correlated (somewhat similar to concurrent validity for an objective test). For example, if investigators developed a scale that purportedly measures students' entrepreneurial orientation in an attempt to identify those who may go on to own their own pharmacies or start new businesses, there might be theory to suggest a relatively high correlation between entrepreneurial orientation and general propensity for risk taking. On the other hand, that same scale should exhibit discriminant validity, or be unrelated or weakly related, to other measures. For example, responses to the entrepreneurial orientation scale might be weakly correlated, at best, to measures of self-esteem and grade point average.

The investigators should pursue concurrently an item analysis to determine the scale's overall consistency and the adequacy of each item stimulus. Item analysis (and factor analysis) procedures are available on statistical software packages such as SPSS and SAS. The scale's reliability, or the consistency of respondents in reacting to the items, is determined through calculating a Cronbach's alpha. Although subject to debate, a Cronbach's alpha of 0.70 or higher may be considered acceptable.[18] The item analysis should involve calculation of the Cronbach's alpha and Cronbach's alphas with the deletion of each individual item stimulus. If the overall Cronbach's alpha is improved with the deletion of an item, the investigator may consider deleting the item from the scale. The decision to retain or delete items in the scale should be derived from a compilation of evidence including the factor analysis, item-to-total correlations, and changes in Cronbach's alpha.

REPORTING RESULTS

Results from summated scales typically are reported as means for each item along with means for each domain, if applicable, and perhaps a mean response to the entire scale. Depending upon the nature of the referent object and the research objectives, it may be appropriate to list the percentage of responses for each item with the appropriate response category in table form (ie, percent of respondents who strongly agree, agree, etc). This is particularly useful when the research is exploratory in nature and the main objective is simply to identify opinions about various aspects of a referent object. Presenting the results this way also may be appropriate when attempting to categorize persons more likely to agree or disagree with a statement in a subsequent Chi squared or logistic regression analysis procedure.

It has been argued that, because scale data are ordinal, only non-parametric statistics should be used in their analysis;[19] however, Kerlinger contends that it is safe to assume equality of intervals in the scale.[4] Moreover, the results obtained from using robust, parametric statistics are quite satisfactory. In fact, the strictest application of rules about the use of parametric statistics for scale data would leave many researchers ill-equipped to handle the multivariate nature of most problems existing within the social, administrative, and clinical sciences. Thus, scale scores can be subjected to tests such as Student *t* tests, one-way analyses of variance (ANOVAs), multivariate ANOVAs, and multiple regression analysis procedures, as appropriate, to test the research hypotheses. Researchers must be careful about violating critical assumptions about the distribution of data and the collinearity among independent variables when conducting such statistical tests. Additionally, when multidimensional scales are the focus of study, it is important not to inflate type I (alpha) error by conducting repeated tests (eg, *t* tests and one-way ANOVAs) on each item or domain of items. Investigators may employ the use of multivariate analysis of variance (MANOVA) with subscale totals or factor scores acting as the dependent variables. Significant findings may be followed up with post hoc *t* tests and one-way ANOVAs, as appropriate. Investigators not familiar with the use of such procedures should refer to statistics texts or, preferably, consult a statistician.

SUMMARY

Survey methodology may be considered a nearly indispensable tool for the academic researcher. Self-administered survey questionnaires (and guided interviews) allow us to elicit behaviors, attributes, beliefs, and attitudes among populations of persons; however, basing decisions or making inferences on data gathered from poorly constructed measures may be more problematic than doing so in the absence of any data at all. Summated rating scales are commonly used to elicit feelings or attitudes from students, faculty members, administrators, and others. Items comprising a summated ratings scale should be gathered from a variety of sources and in accordance with a few basic tenets to ensure that responses to the scale are valid and reliable. There are a number of means by which a scale's validity and reliability can be determined. Data gathered from summated rating scales may be analyzed with the use of robust and powerful parametric statistics. Consultation and advice from survey methodologists and statisticians may help investigators achieve publication of their scholarly work.

REFERENCES

1. Edwards AL. *Techniques of Attitude Scale Construction.* New York, NY: Appleton-Century-Crofts, Inc; 1957.
2. Dillman DA. *Mail and Telephone Surveys: The Total Design Method.* New York, NY: John Wiley & Sons; 1978.
3. Fowler FJ Jr. *Improving Survey Questions: Design and Evaluation.* Thousand Oaks, Calif: Sage Publications, Inc.; 1995.
4. Kerlinger FN, Lee HB. *Foundations of Behavioral Research,* 4th ed. Orlando, Fla: Harcourt College Publishers; 2000.
5. Nunnally JC. *Psychometric Theory.* New York, NY: McGraw-Hill; 1978.
6. Likert R. A technique for the measurement of attitudes. *Arch Psychol.* 1932;No. 140.
7. Babbie E. *The Practice of Social Research,* 7th ed. Belmont, Calif: Wadsworth; 1995.
8. Converse JM, Presser S. *Survey Questions: Handcrafting the Standardized Questionnaire.* Thousand Oaks: Calif: Sage Publications, Inc; 1986.
9. Schuman H, Presser S. *Questions and Answers in Attitude Surveys: Experiments on Question Form, Wording, and Context.* New York, NY: Academic Press; 1981.
10. Schwarz N, Knauper B, Hippler HJ, Noelle-Neumann E, Clark L. Rating scales: Numberic values may change the meaning of scale labels. *Public Opin Quart.* 1991;55:570-82.
11. Lenski G, Leggett J. Caste, class, and deference in the research interview. *Am J Sociol.* 1960;65:463-7.
12. Hassell K, Hibbert D. The use of focus groups in pharmacy research: processes and practicalities. *J Soc Adm Pharm.* 1996;13:169-77.
13. Tully MP, Cantrill JA. Use of the nominal group technique in pharmacy practice research: processes and practicalities. *J Soc Adm Pharm.* 1997;14:93-104.
14. Fink A. How to Ask Survey Questions. Thousand Oaks, Calif: Sage Publications, Inc; 1995.
15. Snider JG, Osgood CE, eds. Semantic Differential Technique. Chicago, IL: Aldine; 1969.
16. Fink A. *How to Design Surveys.* Thousand Oaks, Calif: Sage Publications, Inc.; 1995.
17. Crowne DP, Marlow D. A new scale of social desirability independent of psychopathology. *J Consult Psychol.* 1960;24:349-54.
18. Hair JF, Anderson RE, Tatham RL, Black WC. *Multivariate Data Analysis,* 5th ed. Upper Saddle River, NJ: Prentice Hall; 1998.
19. Cliff N. Ordinal Methods for Behavioral Data Analysis. Mahwah: NJ: Lawrence Erlbaum; 1996.