CHAPTER 11

Structured Interviewing and Questionnaire Construction

Susan C. Weller

Most of what we know about what people think and do comes from interviews and questionnaires. This chapter focuses on the development of interview materials for collecting direct informant-based information with interviews and questionnaires. It is organized by interview *purpose* and presents different approaches to interviewing and question formats within the context of study goals. A mixed methods approach is recommended, beginning with open-ended questions in an exploratory or pilot phase and then integrating those results into a second phase using structured or systematic interviewing techniques and questionnaires.

This two-step process is widely used across the social sciences in the development of interview materials. The initial stage of any study should include a descriptive exploration of the topic under study. A variety of strategies are available for conducting semi-structured individual or group interviews. In general, the less that is known about an area, the more appropriate unstructured, open-ended interviewing methods are. For new areas of investigation, the goal is to develop questions and materials relevant to the area of inquiry and the people being studied. If an existing questionnaire or scale is to be used, especially if it will be used on a new population, then the initial interviews serve to verify that the questions and content are appropriate for the new population. The initial phase, then, focuses on eliciting relevant themes, questions, and responses for further study. A productive technique for doing this is the *free-listing* interview.

The second stage incorporates those results into the development or modification of structured interview materials for a more systematic and detailed examination of the topic and responses across people. In anthropology, the first phase can be quite lengthy, as the purpose is often to explore topics in a new population, a new setting, and a new language. Descriptive information may then be used to frame a study on cultural beliefs or behaviors. In psychology, an initial phase of interviewing may be used to generate items for a new scale or to modify existing scale items (e.g., questionnaires) for use on a new population. In sociology, large surveys begin with a pilot or preliminary phase of interviewing to test clarity, comprehension, and question content.

The combination of an initial descriptive exploratory phase followed by a systematic, structured phase produces a study much superior to one based on either method alone, but it also involves a greater commitment of time and energy. Projects relying solely on either responses to open-ended questions or on responses to a series of agreement rating scales can be biased and inaccurate. Responses to open-ended questions are limited by memory bias: People can *recall* fewer items (reasons, cases, etc.) than they can *recognize* when presented with a complete listing of relevant items. This means that a spontaneous, unstructured request for information, while retrieving important information, may not retrieve complete information. When someone fails to mention a particular item, the item may not be important or it simply may have been forgotten. Also, there are different response styles that result in different amounts of information per person; some people provide long, detailed answers while others give short ones. Further, the use of different prompts and probes in response to individuals' responses effectively changes the questions and makes it extremely difficult to compare responses to open-ended questions across individuals.

Use of a structured format with the same set of questions and responses for all respondents produces comparable information across people and facilitates detailed comparisons across individuals and groups. If the questions or responses, however, are researcher generated and are not preceded by descriptive interviewing to verify relevance and wording, the interview may focus on items of interest to the researcher and may misrepresent or entirely miss topics of importance to the informants. A preferable approach is to combine both qualitative and systematic interviewing, taking advantage of the strengths of each: using open-ended questions to explore a topic and develop an understanding of relevant themes, questions, and responses and then using a structured interview to collect systematic data with those themes, questions, and responses.

After a descriptive or qualitative phase elicits relevant themes, structured interview materials or questionnaires can be developed to examine knowledge, attitudes, beliefs, and behaviors about those themes. A variety of question formats are available. For example, most interviews contain *general information questions* requesting sociodemographic information from the respondent. These questions can be constructed in a variety of formats (close-ended, multiple choice, or open-ended) and are designed to collect specific information like age, gender, years of education, racial/ethnic identity, religious affiliation, number of children, and the like. Questions may also ask about behaviors ("In the last year, how many times did you visit a doctor?") or relationships ("Name the people with whom you have discussed important personal matters during the past six months."). These types of questions request information about the respondent or about people in his or her social network.

Questions can assess *knowledge*. Knowledge tests evaluate the degree to which an individual or group possesses knowledge about a particular topic. Tests may be constructed with multiple choice, true/false, or open-ended questions. A specific assumption of a knowledge test is that the correct answer to each question is known, so respondents' answers are scored as correct/incorrect in relation to that standard.

Questions can also assess *attitudes*. Attitudinal scales attempt to measure the degree that an individual demonstrates or possesses a specific predefined construct that is usually psychological, such as authoritarianism, acculturation, or depression. The most common format for attitudinal questions is a series of statements, typically with a rating scale for each, where respondents are asked to rate their relative agreement with or the frequency of each statement. Similar to knowledge tests, responses are "scored" with reference to an a priori defined standard or criterion and then combined across statements to create a single score or scale of the construct.

Another type of study explores respondents' *classification* of a set of items and describes the categories or dimensions used by people to discriminate among items in a set. Classification studies try to uncover respondents' own dimensions of discrimination rather than access their adherence to a priori defined dimensions. Informants are asked to compare items in terms of their similarity, without reference to any specific dimensions or criteria. Formats appropriate for the collection of similarity data include: pile-sorting tasks, where respondents are asked to sort items into piles according to their similarity; and paired or triadic comparisons of items, where respondents judge the similarity of pairs of items.

Finally, the purpose of a study may be to describe the *beliefs* of a group of respondents. While a classification study focuses on respondents' beliefs (e.g., how they divide up the world into sets and subsets), beliefs may be studied in greater depth by administering a series of related questions on a single topic. For example, a series of questions might ask about specific attributes or assertions, like possible effects of global warming. Question formats differ from those appropriate for classification studies and include: open-ended, multiple-choice, ordering or ranking, and interval or frequency estimate type questions.

Classification and belief studies depart meaningfully from knowledge and attitudinal studies in the handling of informants' responses. In knowledge and attitudinal studies, responses are recoded or scored against a predetermined standard; in classification and belief studies, responses are not recoded, transformed, or scored. Thus, while many formats are applicable across a variety of study purposes, not all formats lend themselves to every purpose.

PHASE I: EXPLORATORY INTERVIEWS: GETTING INFORMATION TO DEVELOP STRUCTURED INTERVIEW MATERIALS

The first phase of a project should be dedicated to gaining a broad understanding of the area of study. Without general background knowledge, it is impossible to know what *questions* are appropriate. So, depending on how familiar you are with the topic and the population you intend to study, a project begins with unstructured and semi-structured interviews and progresses to more structured interviews. Initial interviews may explore a topic in general to gain a broad understanding of the topic and terminology. The first step in this initial phase of interviewing, however, focuses on learning whether or not the topic of study is relevant to the population and discovering the "right" questions to ask. The second step focuses on eliciting more substantive information. Then, elicited information may be used to develop new interview materials or to check the appropriateness of existing materials.

The elicitation of items, statements, and themes relevant to the topic of study is the focus of the initial stage of interviewing, whether interviews are conducted individually or in groups. The set of items is sometimes called a semantic or cultural domain. A domain is a set of related themes, concepts, or statements on a single topic. For this purpose, items are elicited from informants in their own words. Without such elicitation of items directly from informants, items may reflect ideas of the researcher and not the informants. Convenience or purposive sampling is often used in this phase of a study, where a small group of people are selected for interviewing based on characteristics that are important to the study (Johnson 1990; see also Teddlie and Tashakkori 2009).

ASKING THE RIGHT QUESTIONS

The first step is to find out what questions to ask. What are meaningful and productive questions? If you are new to the topic, the people, and maybe even the language, one of the best sources on getting started is Spradley's (1979) book, *The Ethnographic Interview*. Although more than 30 years old, this book is still one of the best sources on getting started with ethnographic or qualitative interviewing precisely because Spradley begins at the beginning by discussing how to find an informant and what to ask. Informants should be selected according to the purpose of the study and should come from the target population. If the study is about urban gardeners, then initial interviews should be conducted with urban gardeners. In this formative stage, often only a half-dozen people are necessary. Informants should be accessible and have time to sit and talk. An informant should be someone with expertise on the topic, someone with at least a year of full-time experience and who is currently involved in the activity or topic. Initially, the investigator's role is as a "student," to learn enough about a topic to ask reasonable questions about it. Grand tour, mini-tour, and taxonomic questions help you understand what is relevant to your informants and the terminology and organization of the domain.

Grand tour questions (Spradley 1979, 86–88) are very productive in starting interviews and learning about a topic by providing an overview. Grand tour questions are general questions that ask for a description of a place, a process, or a typical day. "Could you describe the inside of the jail for me?" "Can you tell me all the things that happen when you get arrested for being drunk, from the first moment you encounter the police, to going to court and being sentenced, until you finally get out of jail?" (Spradley 1979, 86). What you are asking for is a review of something, allowing informants to talk about whatever they want; you will hear about things of importance to the informants as they tell you their impression of how things are organized.

Mini-tour questions (Spradley 1979, 88) concentrate on unpacking meaning from smaller or more specific activities. Often embedded within a longer, more general description are smaller experiential units and processes. Similar to the grand tour overview questions, mini-tour questions ask for descriptions of these smaller events: "You said that a table of guys gave you a hard time last night, Can you give me an example of someone giving you a hard time?" (Spradley 1979, 88).

Taxonomic questions (Spradley 1979, 132–54) may be used to elicit an entire taxonomy from one or more informants. A taxonomy is a structure of set and subset relations among domain items. General questioning like "What kinds of ____s are there?" with comparative and contrastive questions like "Is ___ a kind of ___?" can be used to construct a taxonomy of domain items. Taxonomic relations can distinguish relevant categories of kinds of things, attributes, functions, causes, and examples (Spradley 1979, 110). This type of interviewing is excellent for mapping out terminology (especially in a new language or with a new population) and gaining an understanding of the interrelations among items. It is a logical process of interviewing, developed from observing courtroom cross-examinations and can be seen in the early work in this area (Conklin 1969; Frake 1964; Meztger and Williams 1966; see also D'Andrade 1995). Ground-breaking work by Berlin et al. (1968, 1973) detailed indigenous knowledge of plants, and Berlin and Kay (1969) described color terms primarily with taxonomic interviewing techniques. Taxonomic interviewing was popular in the 1960s and 1970s and was often relied on as a primary data collection technique to describe structure within a semantic or cultural domain. Few people today consider this a primary data collection technique, but it is a very valuable way to quickly get into and understand terms and relations from an informant's perspective.

Item Generation: Domain Definition

Given that you now have a topic that is relevant to the informants you intend to study and know enough about the topic to ask reasonable questions about it, you are ready to actually begin eliciting items. The goal is to elicit a set of related words, statements, or themes relevant to the domain you wish to study. The emphasis here is to obtain the items directly from informants. While an item pool can be created by simply sitting down and writing a series of statements, it is best to elicit items from informants and write statements from those items. Free listing is a productive technique to elicit terms as the goal is to get an exhaustive list of themes from each person, thereby reducing the number of necessary interviews. Themes may come from individual or group interviews or from other sources, such as from narratives and case reports.

FREE LISTING

In free listing, an open-ended question is used to obtain a list or set of items from each informant. (What kinds of ______s are there? Name all the _____s you know.) For the study of social networks, a question might focus on listing all the network members (Name all your friends.) or listing all the type of relationships or exchanges that you can have with people in your network. The goal is to have a comprehensive sample of items by getting an exhaustive list from each person. Responses should be at the same level of contrast, without any set-subset relationships among items. While taxonomic interviews map out terminology and relationships among sets and subsets of items more broadly, free listing focuses on a single level of contrast. If terms indicate a class of items, then the class should be explored and specific items listed. Free listing can help define the set of domain items and its boundaries. It can be used to discover descriptive terms for Alzheimer's disease (Karlawish et al. 2011) or genitalia (Cain et al. 2011). Some fields rely heavily on free listing to generate inventories of things (e.g., in ethnobotany, free listing is used to generate inventories of plants and to assess plant knowledge) (Canales et al. 2005; Mathez-Stiefel and Vandebroek 2012; Miranda et al. 2007; Schunko and Vogl 2010; Vogl et al. 2004). Listing can also be used to elicit perceptions of other environmental features (Mathevet et al. 2011).

Some areas or topics are so clearly defined that a single question can elicit domain items. These domains are easy to identify because of the ease with which informants can produce a list of items. For example, Henley (1969, 177) asked a sample of 21 students to list (in 10 minutes) all the animals that they knew. Individual lists ranged from 21 to 110 animals, and the median number of animals listed was 55. Weller (1984) asked 20 women in the United States and Guatemala "to name all the illnesses or expressions for being sick that they could think of." These are clear, unambiguous requests that generate many items. When responses were tabulated to see the number of people who named each item, many items were named by a majority of the sample and some were named by only one or two people. A total of 423 animals were listed: Four were named by over 90% of the sample and 175 were named only once. Cancer, the most frequently mentioned illness in the U.S. sample, was mentioned by 75% (15/20) of the sample; 6 items were mentioned by 50% or more of the sample; and 30 items were mentioned by at least 15% (3/20) of the sample. Because salient themes and items tend to be mentioned by more people and mentioned earlier in individual lists (Bousfield and Barclay 1950; Friendly 1977), further study of the animal and illness terms focused on using items easily recognized and omitting items mentioned by only a few people.

FREE LISTING WITH MULTIPLE, RELATED QUESTIONS

A series of related listing questions may be used to elicit exhaustive lists from individuals. The series of questions may be perceived by some informants as being all the same, but others respond differently to each question and provide detailed responses to some questions and not to others. In a study of women's preferences for different infant feeding methods (Weller and Dungy 1986), a series of questions was used to try and tap the set of reasons that might influence a woman's decision to breast- or bottle-feed their infant. Multiple questions were asked of each informant, to capture the positive and negative aspects of each feeding method. In all, women were asked 18 free-listing questions to elicit lists of reasons for choosing either breast- or bottle-feeding, but all 18 questions tapped into the single domain of characteristics of infant feeding methods:

- Please tell me the reasons why you want to breast-feed.
- Why do you think some people breast-feed?
- Why did you decide not to bottle-feed?
- What are the advantages of breast-feeding?
- What are the disadvantages of breast-feeding?
- What are all the things you like about breast-feeding?
- What are all the things you *dislike* about breast-feeding?
- When is breast-feeding appropriate?
- In what situations would you *not* want to breast-feed?

(Each question was repeated substituting bottle-feeding for breast-feeding.)

FREE LISTING WITH CONTRASTING QUESTIONS

A related format, that also uses multiple questions, is the use of contrasting questions. Here, items are compared (in pairs) and informants are asked about the distinguishing features. Young (1980) used this format in studying choices for health care. To elicit reasons for choosing a particular health care source, he asked informants why they might go to a doctor and *not* a pharmacist, why/when they would consult a pharmacist and *not* a doctor. The anchored comparison helps elicit more detailed information than the general question of "why/when would you go to a doctor?" or "why/when would you go to a pharmacist?" Another study elicited descriptive attributes of social success, by free listing positive and negative attributes (Freeman et al. 1981; Romney, Smith et al. 1979). For one subsample, informants were asked to name people who they thought were successful and to describe each one; then they were asked to think of people who were failures and to describe them. For another subsample, they were asked to think of five friends or acquaintances and to describe all the ways that each was successful and then all the ways that each was a failure.

SUCCESSIVE FREE LISTING (LINKED LISTS)

Related lists of items can also be elicited by asking for an exhaustive list of items in one domain; and then using those responses to asking for a new list in a related domain. This linked-listing task has been called "successive free listing" (Ryan et al. 2000). In a study of adolescent behaviors and possible punishments, interviews with Anglo and Hispanic adolescents explored adolescent "misbehaviors" and "adult disciplinary responses" (Weller et al. 1987). Verbatim responses of 29 Anglo and 27 Hispanic adolescents (with approximately equal numbers of males and females) were recorded. Each interview took one to two hours to complete and consisted of open-ended and free-listing type questions, descriptive answers, and probes by interviewers to seek further explanations. The following issues were explored:

1. "What things do you (or other teenagers) do that make your parents/mother/father/adults, etc., angry?"
2. For each response to the previous question:
   - "When you do _____, what do your parents, etc. do?"
   - "What other things might be likely to make adults upset or angry?"
   - "And if _____makes adults/etc. angry, what might they do in response?"

The purpose here was to elicit an exhaustive a list for each informant for each question, so the question was changed slightly and asked again as informants exhausted their list. First, questions focused on eliciting teen misbehaviors. Multiple questions were used as probes: "What things do *you* do that make your parents angry?" "What things do *other teenagers* do that make their parents angry?" Then a second domain of adult behaviors was also elicited, linked to the responses given to the first question(s): "When you do _____, what do your parents do?" "And if _____ makes adults angry, what might they do in response?" Information on the second domain was requested listing possible responses *after* listing all misbehaviors. Thus, two related lists were elicited: the set of things teenagers do and the set of things adults do in response. Responses were tabulated across all 56 adolescents for each of the two domains.

OBTAINING ITEMS FROM OTHER SOURCES

Lists of items generated by informants also can be supplemented with items from other sources. In a study of possible cultural differences in the definition of punishment and child abuse, punishment items listed by Anglo and Hispanic adolescents were supplemented with physical abuse descriptions from a hospital emergency department (Weller et al. 1987). Because extreme forms of punishment and abuse are infrequent and would not be expected to be reported in a small sample, a list of the most frequently reported forms of physical abuse was obtained from hospital emergency room records and was incorporated into the final list of items. Similarly, items can come from published sources, from existing questionnaires and scales, or from participant observation. To create a taxonomy of factors relating to disclosure of medical errors, Kaldjian et al. (2006) obtained examples primarily from the medical literature and supplemented that set with examples from interviews.

LIST LENGTH

Informants should be able to generate lists of about a dozen items. List length may be affected by context (Miranda et al. 2007) and expertise (Hutchinson 1983). If lists are short, try probing more. Prompts and probes can elicit more items. Avoid asking questions that can be answered with a "yes" or "no." Thus, rather than asking "Are there any more ___s?" Ask "What other kinds of ____s are there?" This nonspecific prompt can help elicit more items and also helps your informant understand that you want an exhaustive list. Another prompt is to repeat what has already been listed: "You said that ____ and ____ are kinds of ____s. What other kinds of ____s are there?" Here, you remind the informant what he or she was thinking and convey the message that there are more items. The main question can be repeated in a slightly different way, as with the multiple questions about infant feeding methods, adolescent behaviors, and health care sources. Brewer (2002) compared the effectiveness of (1) a nonspecific probe; (2) repeating listed items; and (3) reading back each item and asking the informant to think about the item and other items that are similar to it. All three methods produced longer lists, but the third technique increased the list length by almost 50%. If such probes fail to generate richer lists, you might try a different format for the focus of the question, by using multiple or contrasting questions, or by using an altogether different focus. It is possible that the set may exist in your mind (the researcher), but not necessarily in the minds of the informants.

RECORDING RESPONSES

Responses should be recorded verbatim. The point of generating items from the informants is to discover their definition of items in their language (verbatim). All ambiguous phrases and thoughts, however, should be clarified. The interviewer should probe and seek to determine explicitly what is meant: "What do you mean by _____?" Or, "Can you tell me more about that?" The goal is to elicit statements or themes that are clear so that only one meaning is conveyed (e.g., if a statement is repeated to someone not present at the interview, they should understand the exact meaning intended by the informant). An example of this in the infant-feeding study was that some women stated that they had chosen *breast-feeding* because it was *convenient*. Others stated that they had chosen *bottle-feeding* because it was *convenient*.

Further probing in each of these situations revealed that the breast-feeders meant that they could feed their infant without having to prepare or clean bottles and the bottle-feeders meant that they could feed their baby anywhere without the embarrassment of exposing their breasts. Thus, the latter full statements more clearly express the reasons for choosing a particular feeding method. It is not sufficient, then, to use a general theme, such as convenience, when that theme has more than one meaning. A goal in recording responses is to be sure that you have captured the essence or the underlying meaning *in the informants' own words*, as much as possible, so that you may use specific statements, phrases, and idioms in subsequent interviews with the caveat that the exact meaning is understood.

SUMMARIZING RESPONSES

Unique, verbatim answers or themes are tabulated across respondents. Domains may be defined with the use of single questions, multiple questions, contrasting questions, linked listing questions, and sometimes with supplementary items from other sources. Answers are then tabulated by informant and not by question. This is especially important when using multiple questions to elicit items, so that when someone mentions something more than once, it is counted only once, for that informant. The final tabulated list should reflect the number of unique people who mentioned each item. Final statements should be in clear language with consistent syntax. Statements should convey the same meaning to each and every reader.

In the infant-feeding study, the 18 most frequently mentioned themes from the English-speaking Anglo and the Spanish-speaking Hispanic lists were chosen for study and changed to a neutral form "A way to feed your baby that. . . ." The items were also balanced so that half of the items referred to breast-feeding and half to bottle-feeding and half contained positive attributes and half were negative. Although the list contained a culled and modified set of the multitude of statements collected, language and ideas remained concordant with those in the original interviews. Tabulation of responses helps provide a sense of the relative salience for the themes across people.

SAMPLE SIZE

The necessary sample size for free-listing interviews is a function of variation. This is true for both qualitative and quantitative research. The less variation (e.g., the more consistent the responses are) across people, the smaller the necessary sample size. For some domains, a sample size of 10 may be sufficient and for other domains, or for increased accuracy, sample sizes of 50 or more may be needed. Typically, a sample of about 20 informants is adequate, especially with a good list length per person. As the number of interviewed informants increases, say in increments of five; from 5 to 10, 10 to 15, and so forth, there will reach a point where little new information is added to the content and order of tabulated items. This is sometimes referred to as the point of saturation. Thus, the sample size is adequate when the addition of new people or groups does not alter the frequency distribution of items and few new items are added.

By getting a *list* of items from each informant, more information is obtained per informant and fewer informants are needed, and saturation is reached more quickly. With a meaningful question and probing, each informant should be able to generate a list of at least 6 things, usually around 10 to 12, and sometimes many more. Agreement on items, statements, or themes is estimated simply by counting the number of informants that mentioned each. The set or domain is defined by the items mentioned by multiple informants. The most frequently mentioned items are the most salient items. Psychologists have shown that the most salient items will be named by more people and those items will appear higher up in individual lists (Bousfield and Barclay 1950; Friendly 1977). Salience of items is estimated most simply with the frequency distribution (e.g., the percent of the sample that named each item) and can be used to make comparisons between samples (Ross and Medin 2005; Thompson and Juan 2006). Sometimes salience is estimated with a consensus analysis to identify items mentioned by a majority of the sample (Mathevet et al. 2011). While the set of items obtained with free-recall listing is not necessarily definitive, it should nevertheless capture most well-recognized items.

GROUP INTERVIEWS

Free-listing interviews may be conducted with individuals or groups (focus groups). An important thing to remember, however, is that the sample size for group interviews is not the number of participants, but is closer to the number of groups. Lists generated from group interviews do not reflect the thoughts of each individual in the group, rather, the interviews reflect the group's thoughts and thus only one list is generated per group. Individual interviews are much more productive than group interviews in terms of generating ideas. Group interviews generate only about 60% as many topics as do individual interviews (Fern 1982; Morgan 1996). Larger groups are more productive than smaller groups, so one group of eight people is preferable to a group of four; but more groups are better, so two groups of four each are better than one group of eight (Fern 1982). Saturation for group interviews often occurs with four to six groups of eight people each (Morgan 1996).

It is important to note the total amount of time invested in interviewing: 20 individual free-list interviews that average 45 minutes each results in 15 total contact hours of interviewing, four groups of eight people typically result in 4 to 6 total contact hours of interviewing, and eight groups of four would have 8 to 12 hours of interviewing. Production of ideas and differences between methods may also be a function of the time invested in interviewing.

NARRATIVES, CASE HISTORIES, AND TEXTUAL MATERIAL

Another approach to gaining an understanding of a topic or domain is to collect descriptive accounts, like narratives or case histories. Themes can be identified in textual materials in phrases and ideas that are discussed, repeated, labeled as categories of things, or used as metaphors (Ryan and Bernard 2003). Quinn (1987) culled themes relevant to American beliefs about marriage based on informants' descriptions of marriage. Chavez et al. (1995) recorded women's descriptions of possible causes of cancer and used recurring themes for further study. Kempton et al. (1995) also began their systematic study of U.S. environmental beliefs by collecting descriptive narratives and identifying themes from the descriptions. Johnson and Griffith (1996) conducted detailed interviews about pollution, its causes, sea life that is affected, and seafood; and then selected themes from the transcripts for further study.

Analysis of textual materials can only *suggest* possible interconnections and relationships among themes. Unstructured methods of interviewing and response narratives are excellent for suggesting hypotheses, but more systematic data are needed to test the validity of observations and to make comparisons across groups. Personal case histories sometimes yield more detail on a single case, but typically require a larger sample size (more people and more cases) to cover the breadth of cases. A detailed history of the last illness case that occurred in the household collects information on only one case of one illness, and it is difficult to get case information on rare events. In contrast, interviews with individuals about "all the illnesses they know" can uncover information on a variety of illnesses.

PHASE II: STRUCTURED INTERVIEWING TECHNIQUES AND QUESTIONNAIRE CONSTRUCTION

After establishing the items and content for study, a more structured interview format can be pursued. Open-ended, semi-structured formats facilitate the collection of new information with the flexibility to explore topics in-depth with informants. Meaningful comparisons across people may not be possible, however, because informants have been encouraged to discuss different items and thus have not really been asked the "same" questions. Structured formats allow the investigator to make more detailed comparisons across people and groups and can verify impressions from less-structured interview methods. This section describes a variety of question formats. The focus is on designing interview materials (questions, tests, and tasks) appropriate for the goal of the study. Thus, the section is organized by study purpose: general information questions, knowledge tests, attitude scales, classification studies, and assessment of cultural beliefs.

General Information Questions

Most studies include questions to collect information on respondents' socio-demographic characteristics. Questions are straightforward requests for information: age, gender, ethnicity, household composition, length of residence, and sometimes behaviors. These questions parallel those found in surveys.

The term "survey," however, is often used to refer to a combination of methodologies: the selection of respondents, method of interviewing, and questionnaire design (Fowler 2009). The selection of respondents usually focuses on procedures for selecting a *random* or *representative* sample. When a representative sample of respondents is used, results may be generalized from the sample to a larger population. Nonrandom or convenience samples can provide useful information, but generalization of findings should be done with great caution.

The method of interviewing concerns whether interviews are conducted in person, on the phone, or by mail. In-person or face-to-face interviews may be administered by an interviewer or be self-administered and tend to have the highest participation rates. Phone interviews can only be administered by an interviewer, but may be computer assisted by having the questionnaire on a computer. With computer-assisted telephone interviews (known as CATI in the sociological literature), the interviewer enters responses directly into a computer. Mail, email, and web-based questionnaires must be self-administered. More complex responses can be obtained in face-to-face interviews with the use of visual aids, if necessary. Questions and responses must be simplified for oral/phone presentation. Self-administered, open-ended questions usually do not produce useful information due to the lack of probing for clarification.

Participation rates for the three different approaches parallel their costs. In general, face-to-face interviews have the highest participation rates and are the most expensive. Phone and mail methods tend to be less expensive but also have lower rates of participation. As follow-up procedures (call backs and re-mailings) are intensified, phone and mail participation rates (and costs) increase. Self-administered questionnaires in mail, email, and web-based sources tend to have the lowest participation rates. Participation rates below 75% should be examined critically as the sample may no longer be representative and may be biased. It may be preferable to interview a small representative sample on the phone or in face-to-face interviews than to send out a great number of self-administered questionnaires in regular mail or email and get a low participation rate.

Here, the focus is on questionnaire construction and it is assumed that most interviews will be conducted in person. The biggest weakness in questionnaire design is often the result of an investigator that simply drafts a set of questions, assuming that anyone can write a questionnaire. The result is often a set of poorly worded questions with unclear response categories. Unclear questions lead to uninterpretable responses. Sociologists and psychologists have spent an enormous amount of time designing questionnaires, studying the effects of different wording and ordering of questions on responses as well as the interaction between interviewer and respondent. It is a waste of research effort not to take advantage of their experience and knowledge. Recommendations on wording and ordering of items can be found in the sociology literature. See, for example, Fowler's (1995) *Improving Survey Questions* for a very good short, focused description; Bradburn et al.'s (2004) *Asking Questions* is a more complete overall reference; and Fink's (2003) *The Survey Kit* also a handy overview. It is certainly worth investing some time, even if only a few days, to review some of these materials.

Question formats include: open-ended, close-ended multiple choice, and rating scales. Open-ended questions should be simple and seek clear, short answers. Questions should be written as complete questions, so they are asked in the same way for each person. For example, instead of just having "Age ___" on the questionnaire, it is preferable to have "How old are you?" or "What is your date of birth?" Close-ended questions should be concise with a complete listing of mutually exclusive response categories. Rating scales are usually appropriate only for literate informants with a moderate degree of education, although they may be simplified and asked in an oral interview (Weller and Romney 1988).

In general, questions should proceed from broad, general requests for information to questions requesting specific or more detailed information. This is done so that questions requesting detailed information do not bias the responses for more general information. Similarly, less personal questions should precede those perceived to be more private or threatening. Questions requesting sociodemographic information may be asked initially, especially if they help establish whether or not the informant fits the study inclusion criteria. Other sociodemographic questions may be asked at the very end of the interview, especially those adding extra information and those thought to be more personal or threatening, as with questions in the United States regarding income.

Inclusion and exclusion criteria for participants are established as part of the study design or protocol. They are the explicit conditions for including or excluding someone in the study. If you want to study "Latina" women, then before interviewing anyone, you should define who is and who is not a Latina woman. Thus, the initial questions may seek to establish the informant's gender, ethnicity (by self-report and also possibly by birthplace and language preference), and age (in years or parental status). The advantage of having all inclusion and exclusion criteria-related questions first is that an interview may be terminated quickly for people who do not meet study criteria. It is advantageous to collect some information on everyone, even the excluded individuals, to see if there are differences between those who do and do not choose to participate.

Only questions relevant to the study should be included in the interview. Each question should link directly or indirectly to the purpose of the study. Questions should concern factors implicated by theory, factors mentioned in the literature, and factors that might potentially affect results. Too often, extraneous questions are included without considering how responses will be handled. For example, a question on marital status might be included, but if the real interest is whether a woman is living with the father of her child, then a direct question to that effect would provide more useful information. Still, it's best to ask too many rather than too few questions: A question or answer can always be ignored after it is collected, but it's usually difficult or impossible to go back and ask a question that was omitted inadvertently.

If you want to know how your sample results compare with those from a larger population, use questions from large or national surveys. Not only can you compare responses with those of the larger survey, but you can take advantage of the time and effort that went into the development and wording of the questions. Even simple questions can be borrowed directly from such surveys. Also, you can compare different sets of questions purported to measure the same thing. For example, questions about ethnicity can come from multiple sources: You can ask about ethnicity using the questions and categories used in a national census and also from questions you have developed that you believe are more appropriate indicators. Using census categories allows you to discuss your results in terms of national categories and to compare your findings with other reports. Using a new series of questions in conjunction with census questions allows for a direct comparison of the two ways to define ethnicity.

When beginning to design a questionnaire, take advantage of previous scholarly work and look around for published questions (and responses) and do not hesitate to use them. For example, in the United States, check the U.S. Census, the American Community Survey, the General Social Survey (also done in many countries around the world), the National Health Interview Survey, the National Crime Survey, and the Consumer Expenditure Survey. Also see the World Fertility Survey and the World Values Survey.

When writing questions, keep the study's purpose in mind. Translate the purpose into specific questions that will directly or indirectly provide information relevant to the purpose. Also, have a plan about how responses will be used to meet the study's purpose. Good questions are ones that respondents understand, that all respondents interpret in the same way, and that respondents' understanding is the same as the intended meaning. All questions should be administered in the same way to all respondents.

The wording of questions should be clear and simple. Avoid ambiguity in meaning and define terms if necessary. Avoid compound questions with more than one concept embedded in a question. It is preferable to use multiple, related, and simple questions than ask complicated or long questions. It is important to ask things that informants know about and can answer meaningfully. Questions like "What kind of health insurance do you have?" may reveal that people know whether they have health insurance and maybe the company, but they simply may not know much about the actual coverage. And people cannot answer complex questions, like "What proportion of your time is spent doing ___?" To answer involves an estimate of time spent on different activities and then divided by the total time. Instead ask: "Have you done ___ in the last month?" Minimize the difficulty in answering.

Requests for information for a shorter, more recent time period rather than a longer period of time tend to get more accurate answers. The U.S. National Crime Survey asks about experiences over the past six months, and interviews about illness episodes typically ask about experiences in the past two weeks. Bias in responses tends to be toward what people usually do and not what they did on a specific occasion or time period. Respondents also tend to "telescope": When asked about behaviors during a specific time period, they report actions from a longer period of time. If asked if someone went to the dentist in the past year, people tend to say yes if they visited the dentist within the past two years.

To improve recall on behaviors during a specified time period, it is important to mark the time period with an important event and to ask several questions about the behavior. For example, rather than asking, "Has anyone in the household been ill in the past three months?" ask instead, "Since Easter/Holy week, has anyone been ill in the household?" or, "Since our last visit, has anyone been ill in the household?" (Weller et al. 1997) Alternatively, ask this as a series of questions. If you are interested in illnesses during the past week, begin by asking about illnesses over the past year, then in the past three months, and then in the past two weeks. Multiple questions signal that the question is important and can improve accuracy (Fowler 1995).

To minimize problems in reporting accuracy, clarify the goal(s) of the study with respondents. Emphasize that there are no right or wrong answers; that responses are confidential and anonymous; and that providing accurate information is important. Inasmuch as possible, the interviewer should be matched to the respondent by gender (men interview men, women interview women) and background (similar ethnicity and SES). Where possible, borrow questions from national surveys. Avoid ambiguous words and complicated concepts, ask simple, straightforward questions. Make questions easy to answer. Give help with recall over a specific time period by marking the period with specific, memorable events and use multiple questions to improve accuracy. Responses to multiple, related questions can be combined to form an index or scale.

COMBINING RESPONSES TO CREATE SCALES AND INDICES

As requested information becomes more abstract (i.e., as questions move from simple ideas like gender and age to more complex ideas such as social class), more questions are needed to get a reliable estimate of the concept. For concepts that cannot be measured simply or directly, use proxy questions to get information associated with or indicative of the underlying concept. Then combine the responses to obtain a more reliable and accurate estimate. For example, we believe that social class or socioeconomic status exists, even though there is no direct, single question or ruler by which we can assess or categorize an individual or household.

In developed countries, we often use combinations of educational level, income, and occupation to estimate socioeconomic status (see Haug 1977). In less-developed countries and among populations with little variation in occupation, education, and income, such variables may not be helpful in differentiating social strata. In lesser developed and rural areas, it's more helpful to ask a series of questions related to socioeconomic status (e.g., questions about house construction materials, water source, ownership of material goods) and to combine responses to differentiate households.

A summative score of responses to a series of questions creates an index or scale. A summative score should be valid; it should measure the idea or construct that it is intended to measure. *Content validity* focuses on the construction of a scale: Are the items reasonable and do they appear to measure the same thing? *Criterion or predictive validity* is the degree to which a scale predicts the idea or construct it is trying to measure. *Construct validity* goes a step further for scales with reasonable content and internal consistency; construct validity concerns the association between a scale and other measures theoretically related (but not necessarily the same construct) to the construct that the scale is attempting to measure.

First, the choice of reasonable questions and proxy variables helps ensure that a combination of responses to those questions will also be reasonable. Second, items selected for combination in a scale should be "scalable" (i.e., they should be positively correlated) with internal consistency and good reliability. A principal components analysis can indicate how to optimally combine variables that are in different units of measurement. A principal components analysis clusters items into groups according to their inter-correlations; items with the same pattern of responses across people (those that have the same pattern of high values and low values across people) are grouped together. Finally, the scale should correlate positively with similar scales and should correlate with other measures in ways predictable by theory.

In developing a scale of financial resources in rural Guatemala, Weller et al. (1997) asked over two dozen questions about household composition, characteristics of the head of household (gender, age, education, ability to read, ability to write), house construction (walls, roof, and floor), and assets (ownership of land, appliances, vehicles, and animals). Some questions requested yes/no type responses: "Do you own your house?" "Do you have a bicycle?" Others requested the number of people or animals; and responses to multiple choice responses (household construction materials) were coded as present or absent.

Seeking to develop a scale concordant with community perceptions, Weller et al. (1997) also asked three informants in six villages to rank 10 families according to their economic resources and retained only those questionnaire items that correlated with the community judgments (10 of the original 28 questions). A principal components analysis of those questions for the larger sample showed that variables most indicative of financial resources (including monthly income) grouped together on the first factor, and variables representing other dimensions of socioeconomic status (educational level and household size) grouped on successive factors.

Weller et al. (1997) wanted a relatively simple scale that could be used in other studies in the region, so they used the principal components solution to identify which variables should be combined (those on the first factor), but not for a weighted combination of variables. To overcome the problem of different units of measure, variables were dichotomized (so they would be in the same units) and summed. Each household received a cumulative score (+1) for the presence of each indicator: monthly income greater than the median; ownership of any appliance; more than two rooms in the house; non-dirt floor; more than three chickens; adobe, brick, or block walls (as opposed to bamboo, wood, or plastic); land ownership; and ownership of a bicycle. Summing across the eight variables created a nine point (0–8) scale. The final scale was concordant with other scales previously constructed to assess socioeconomic status in rural Guatemala (Freeman et al. 1977; Johnston et al. 1987). In fact, such scales are surprisingly similar across rural regions of the world and use indicators such as floor construction (dirt vs. other), type of cooking fuel, and availability of animals for sale.

Guttman scaling is another way to combine household indicators of socioeconomic status. Guttman scaling reveals whether there is a cumulative and sequential ordering of variables: If someone has an item on the list, they would also tend to have the objects that precede the item. Similarly, if a household lacked an item, it would tend to lack subsequent items. Dewalt (1979, 106–15) described a nine-point "material style of life" Guttman scale from the presence or absence of eight variables: iron, radio, bed, cooking facilities off the floor, sewing machine, wardrobe, stove, and television. This means that responses indicated that if a household has a bed, they also had a radio and an iron. Dewalt checked the validity of the scale by comparing the final scale to informant ratings of wealth and found them highly correlated. Guttman scaling has been used to describe the acquisition of consumer goods in Polynesia (Kay 1964; Weller and Romney 1990, 79–83) and in the United States (Dickson et al. 1983; Kasulis et al. 1979). Guest (2000) presents a detailed example for Ecuadorian fishermen using 12 material goods. A related, alternative model for representing the order of acquisition is the Rasch model (Soutar and Cornish-Ward 1997). Guttman scaling can be used to represent a variety of cumulative activities and skills (e.g., social participation activities among the elderly [Bukov et al. 2002] and men's skill in building and creating objects [Johnson 1995]).

Responses can be combined across *related* questions or variables to create a single scale or index. Such indices are more reliable and accurate than use of a single question, especially when the request is for information more abstract than someone's age, height, or weight. While the responses to simple questions may be combined to estimate the household socioeconomic status, a variety of other variables may be similarly combined to obtain better estimates of behaviors and experiences. For example, Handwerker (1996) used a combination of questions to better estimate household activities and experiences of violence and affection.

SOCIAL NETWORK QUESTIONS

Social network studies focus on interrelationships among people and organizations. Questions on social networks tend to focus on two different approaches (Bernard 2012; Scott 2000; Wasserman and Faust 2009). One approach looks at personal or ego-centered networks, where a respondent is asked about his or her relationships with others and the others may or may not know one another. The second approach looks at complete, whole-group (sociocentric) networks, where each person is asked about his or her relationships or interactions with every person in a group.

With ego-centered networks, questions may seek information on the number and types of friends or shared activities and interests. These questions can measure qualitative attributes of relationships between people and can be used to estimate social support or social capital. A first step is to use open-ended and free-list interviews to explore types of people, relationships, and functions that are important, to form a meaningful context for subsequent questions and to be able to ask about those relationships in a meaningful way. Second, in subsequent questions with a new sample, questions would ask systematically about who might offer help, advice, or support in different scenarios (e.g., Burt 1984, 1986, Freeman and Danching 1997). Questions can also drill down and collect detailed information on the type and quality of relationships in the respondent's personal network: for example, "Name 10 people who ___," and then for each person named, ask about their characteristics and the relationships *between* the people to estimate information on the connectedness and density of the personal network (McCarty 2001).

Studies of whole group networks focus on a defined group of people and ask about the frequency and/or quality of their interactions. Here, initial interviews first must list and define all people in the group. Then, each person can be asked about their relationship (advice seeking, exchanges, social interaction, etc.) with every other person in the group. Each person is presented with a list of all group members and asked to check the names of those they interact with the most, or rate them on rating scales on the frequency with which they interact or rank the entire list in terms of how often they interact. For example, Johnson et al. (2003) studied a work group network at the South Pole Research Station and had them rate one another on an 11-point rating scale indicating the frequency of social interaction. Although much more intensive, each person can also be asked about the relationships between everyone else in the group (Krackhardt 1987, 1990). Johnson and Orbach (2002) studied the network of people judged to be important in passing a particular piece of legislation (North Carolina state senators, cabinet-level secretaries, legislative committee chairs and co-chairs, staff, resources managers, lobbyists, and private citizens) and provide an example of how to collect data on a large group of people and minimize the response burden (length of the task) while doing so.

CHALLENGES TO VALIDITY

Accuracy of responses can be compromised by questions that are interpreted differently by different respondents. Questions should be in complete, grammatically correct language and read the same way to each person. One way to understand how informants interpret a question is to interview a small sample of individuals and ask them to think out loud; ask them to describe their interpretation of the question and to list possible answers (cognitive interviews, Fowler 1995).

Another source of inaccurate responses is the informants' own memory. Informants may report an event that actually happened 12 months ago as occurring 6 months ago. Marking a period with an important or widely recognized event (since ____ occurred) reduces this telescoping effect (Loftus and Marburger 1983). Informants also may "misremember" an event, reporting instead what they think happened or what usually happens. Informants are much better at telling you what they typically do, than what happened at a specific time. Freeman et al. (1987) asked a group of individuals about attendance at a group presentation the previous week. Errors consistently counted those who usually were in attendance but were not there as being there, and counted those who usually were absent, but were there, as absent.

Reporting errors tend to be biased in the direction of typical behaviors. Bernard et al. (1980; see also Bernard et al. 1985) found poor informant accuracy, when people reported with whom they interacted for a specified period of time, but reports may have been biased toward more typical behaviors rather than what actually occurred in the specified time interval. When informants' reports of social interactions were compared to observed interactions for a specific time period, studies with longer observation periods (a better sample of typical behaviors) tended to have better informant accuracy. D'Andrade (1974) found that coding of behaviors *immediately after they occurred* corresponded more to the similarity among adjectives rating the behaviors than to the behaviors that actually occurred. So, if someone was remembered as having smiled, they were more likely to be attributed with actions associated with smiling like having been facilitative, friendly, and so on, whether they were or not. One explanation for this bias is that people who smile are usually helpful and friendly.

Accuracy of responses also may be affected by the interview itself. Contextual effects have long been documented and studied by sociologists and, generally, better responses are obtained when the interviewer and the informant share characteristics such as gender and ethnicity (Schuman and Presser1996). An informant's lack of experience with the interview process may decrease accuracy, and informants may offer socially desirable responses or may deliberately mislead you. Accuracy may be increased as participants understand the purpose of the interview and the degree of confidentiality in responding.

Knowledge Tests

A knowledge test consists of a series of questions designed to test someone's ability or knowledge. The answers—the correct answers—to the questions are known, and responses are scored or *recoded* as correct/incorrect. First, the content domain is established that covers the subject matter or ability to be tested. Then, test questions are drafted. Question format may be multiple choice (with two or more choices) or open-ended (requesting single word or short phrase answers). Performance of respondents is usually described as the percentage of correct responses (of the total number of questions) or as a percentile, comparing an individual's performance with the distribution of scores across respondents. Just as sociologists have much expertise in writing general information questions, psychologists have extensive expertise in developing knowledge tests. Nunnally's (1978) book, *Psychometric Theory*, presents a thorough review of issues involved in developing a test.

It is important after drafting, administering, and scoring a test to also assess its reliability. An assessment of a test's reliability and the resultant modification of the test can greatly improve a test's ability to discriminate between knowledgeable and less knowledgeable informants. *Reliability* is the degree to which a variable or test obtains the same result when administered to the same people, under the same circumstances. A test with low reliability is analogous to a very sloppy measuring device; it may be valid, but it has a lot of measurement error. For example, if you measured the height of a sample of college undergraduates with a weight–height measuring device typically found in a physician's office and also with a 6" pocket-ruler, you might find that the pocket-ruler estimates could conceivably contain measurement error large enough to mask the difference in average height between men and women. The more accurate the measuring device, the greater the ability to detect smaller differences. The same is true for tests. If a test can be streamlined and *limited* to questions that best differentiate degree of knowledge of the subject matter (thus, increasing the reliability), it can be a shorter, more accurate, and hence more powerful test.

RELIABILITY

Reliability of a test can be assessed in a variety of ways. One way to assess reliability is to give the same test twice, after an interval of time, to the same individuals. Reliability, then, is estimated by the correlation between the two sets of scores. Because the Pearson correlation coefficient is used, reliability ranges from zero to one. This type of reliability, *test-retest reliability*, is limited in that scores may improve due to practice or learning effects and change can occur in the time interval. Alternatively, equivalent, but non-identical tests can be administered, but it's difficult to develop "equivalent but non-identical" tests. A third approach is to create two tests by arbitrarily dividing one test in half and calculating separate scores for the odd-numbered and even-numbered items and administer the test once. This type of reliability, *split-half reliability*, is estimated by the correlation between the two sets of scores. The best overall estimate of reliability, because it subsumes the previous estimates, is provided by the *reliability coefficient* (Nunnally 1978). The reliability coefficient, sometimes called coefficient alpha or Chronbach's alpha, is mathematically equivalent to calculating all possible split-half reliabilities and, while it may sound complex, it is widely available as an easily accessible option in most statistical software packages.

For a test to have high reliability, all of the test questions must be on a single topic and be at the same general level of difficulty. This means that items should be positively intercorrelated, and performance on individual items should be concordant with the overall score. A test question would not be a good estimate of ability if the "best" or high scorers got it wrong and those with lower total scores tended to get it right. Such questions reduce the accuracy of the total score. An *item analysis* helps identify items that do and do not correlate positively with the total score. The *item-to-total* correlation for each question tells how well responses for each question parallel the total score. If the correlation is not positive, or is small (less than +.20), the question should be dropped (Nunnally 1978). Items considered for omission can be modified in future applications. Writing good questions with multiple choice answers is very difficult!

The overall reliability of a test, the reliability coefficient alpha, is a function of the intercorrelation among the questions (the degree to which they measure the same concept) and the number of items (the more items on a single topic the more accurate the estimate):

$$\text{Reliability} = k\bar{r} \;/\; (1 + (k - 1)\bar{r})$$

where k is the number of questions and $\bar{r}$ is the average Pearson correlation coefficient between all pairs of questions. Thus, a reliable test can be created with a few highly correlated items or with a lengthy test of weakly related items. When dichotomous responses are analyzed, this formula is called Kuder-Richardson 20 (KR-20). The reliability coefficient and the performance of each item (in the item analysis) can readily be obtained in most major statistical packages.

EXAMPLE

In a study in rural Guatemala, Ruebush et al. (1992) developed a test to assess local knowledge about the causes, symptoms, and treatment of malaria. Experience both with residents of the region and the National Malaria Service led to a draft questionnaire or test with 65 true/false items. Since the correct answers to the questions comprised the scientific or biomedical model of malaria transmission and treatment, an initial pilot test was a very simple one to see if National Malaria Service workers (those with more biomedical experience) scored higher than the rural residents. This involved a day's worth of interviewing, in a single rural village, interviewing a half a dozen respondents and a few National Malaria Service workers. A quick tabulation of responses and scores, in the field, helped identify ambiguous questions with unclear answers.

A revised version with 65 true/false questions was administered to a larger sample of residents and National Malaria Service workers. Responses, where 0 = no/false and 1 = yes/true, were compared to the correct answers and recoded to 1 if the answer was correct and to 0 if the answer was incorrect. A reliability analysis, in particular the item analysis, helped identify questions that did not perform well because they did not contribute to the total score. The 65-item test had a reliability coefficient of .82. The reliability analysis indicated that reliability could be improved by *omitting* items with low item-to-total score correlations. The omission of 25 items created a 40-item test with a reliability coefficient of .91. Thus, the shorter version of the test had better discriminatory ability, and comparisons between groups could be made with greater precision. This procedure is also used in identifying poor test questions on multiple choice exams for large college classes.

Scores from knowledge tests indicate how well people know the correct answers. In the above example, the answers constituted the scientific or biomedical model of malaria, but the scores did not indicate whether wrong answers were due to a lack of knowledge or to different beliefs. In the malaria study, Ruebush et al. (1992) also analyzed responses in their original form without coding them as correct/incorrect and used the modal response for each question as an estimate for local *beliefs* regarding the answers. Cultural beliefs can then be compared to the scientific answers used to score the knowledge test to identify areas where errors might be due to differences in beliefs. Trotter et al. (1999) conducted a detailed study of Latino AIDS beliefs and compared the results to performance on the national AIDS knowledge test. They found that Latino errors on knowledge tests were most likely not due to different cultural beliefs (see section below on Exploration of Specific Beliefs).

Attitude Scales and Tests

Similar to knowledge tests, attitudinal scales or tests measure the degree to which individuals and groups possess specific constructs. (A construct is an a priori defined concept.) Development of attitudinal scales begins by defining the domain of items relevant to the particular attitude being studied. Statements are generated that describe or are indicative of the attitude. The statements are then administered to respondents, usually with a checklist or rating scales. Informants indicate whether the statement describes their feelings and thoughts. Responses are scored by summing together responses after reversing or reflecting some responses (e.g., reversing scale values by subtracting them from the largest anchor point value plus one), so that the meaning of the values is consistent and small (or large) scores all indicate the absence (or presence) of the attitude. This recoding of responses parallels the handling of responses with knowledge tests in that responses are scored in accordance with a previously determined standard. Attitude scales have been developed for a variety of topics, like depression, acculturation, and quality of life. Question formats can be dichotomous or checklist questions, but are usually rating scales indicating degree of agreement or frequency.

ADAPTING EXISTING MATERIALS AND SCALES

There are many advantages to using existing questionnaires and standardized scales. Most importantly, it allows you to take advantage of the considerable amount of work that went into the development of the scale and facilitates communication with a larger group of scholars. The main disadvantage in using existing materials, especially standardized attitudinal scales, is the questionable validity of the results when applied to a new population. A scale developed on one population may not be directly transferable to another population as scale meaning and performance may not readily generalize to the new population. The application of a scale in a new setting can miss concepts that are important to the new group; ideas or elaborations of ideas in the new population may not be tapped or fully articulated in the original scale.

Nevertheless, the advantages of adopting existing interview materials, when and where they exist, usually outweigh the disadvantages. One approach is to borrow and adapt materials as necessary. A thorough discussion of how to translate and modify materials (especially, tests) is presented by Brislin (1986; see also Jowell et al. 2007; Schrauf and Navarro 2005). Cross-cultural psychologists have extensive expertise in developing tests and scales that are comparable across cultural boundaries.

The first step in adapting a test for another culture or another setting is to translate statements and rating scales. Materials should be translated from the source language to the target language by one person and then translated back into the source language by another person. Brislin recommends two full translation loops (four people). This is especially important for psychological concepts. Taking statements through such translation loops allows the investigator to see which concepts translate. Statements that retain their meaning through translation and retranslation are easily and directly usable. Statements that change meaning or are not consistent across translations need to be modified.

The next step involves assuring that test questions are appropriate. One way to validate items on a test or statements in an attitude scale is to generate the item pool *de novo*. When applying the scale to a new group, even within the same language, it's advisable to generate new items. Open-ended, free-listing questions with a small sample can sometimes reveal quickly and directly the content validity of the items. If newly generated items match or overlap with statements and concepts already included in the scale or test, the scale probably needs little or no modification. If, on the other hand, open-ended interviews elicit many ideas and themes not well developed or measured on the test, then the test probably needs revision. One solution is to add new questions at the end of the set of standard questions. Adding new questions at the end allows you to score the scale in the accepted way and build on the body of literature relevant to the scales as well as to base an analysis on a new set of items.

In a study of pre-term deliveries among inner-city African American women, a standardized measure of stressful life events was modified for that population. Stress, a severe strain or reaction that can be brought on by events and experiences, was measured with a checklist of 43 stressful life events (Holmes and Rahe 1967) that may have occurred in the past year, such as death of a spouse or change in residence; a greater number of positive answers indicates a greater number of life-changing events and possible higher stress. Before using the scale in a study of inner-city women, the investigators conducted open-ended, descriptive interviews with pregnant African American women about the stressful experiences in their lives.

Interviews began with a discussion of stress to discover how it was defined and understood. Then discussions covered the kinds of things that caused stress. The results showed that although the women shared a general definition of stress and had experienced similar stress-causing situations, their stressful life events didn't correspond completely with those in the Holmes and Rahe scale. For example, the women experienced stressful events that were not captured in the scale, such as loss of heat or electricity, being beaten or hit by a husband or boyfriend, and being evicted from their homes (being homeless). To be able to communicate with a larger group of researchers who might use the same scale, the investigators added new items to the end of the scale, rather than modify the scale itself. This gave them the flexibility to analyze stress in terms of either the standardized approach or as a modified test. Stress scales also have been adapted for use in other cultures (Ice and Yogo 2005).

A limitation with attitudinal scales is their questionable validity when used on populations different from that on which the scale was developed. In general, this does not indicate a problem with the test, but instead is a problem with the application and conclusions. Validity, in its most general sense, is the degree to which something does what it is supposed to do. A valid question, scale, or test is one that measures what it is intended to measure. Content validity refers to the appropriateness of the items: Does the content of the items appear to be relevant to the topic that is being assessed? If responses from open-ended interviews with members of the target population overlap with the ideas contained in the existing set of questions, the questionnaire is appropriate for the new application. If the two sets of items overlap on many ideas but not all, the existing materials can be modified by adding new questions. If there is little overlap between the ideas and themes captured in new interviews and the existing materials, an alternative or new test is needed.

CREATING A NEW INDEX OR SCALE

Nunnally (1978, 604–9) describes the process of creating an attitude scale. His discussion is summarized here as five steps:

1. An item pool is created by writing about 40 items on a single topic. (Themes may be taken from free-listing interviews or other sources.) Half of the statements should be moderately positive and half moderately negative. Statements where all or many respondents answer similarly do not help to differentiate people. Thus, neutral statements are not helpful nor are strong statements.
2. Statements are composed into a draft questionnaire and administered to individuals similar to whom the scale will eventually be administered (the target population). Questions may have dichotomous or rating scale responses. The number of respondents should be approximately ten times the number of items. (The sample size recommendation is because principal components analysis is used to ensure that statements are inter-correlated and cluster together as a single conceptual group.)
3. Responses are scored so that high scores all indicate the presence of the concept or trait and low scores indicate an absence of the trait. This means that some responses must be reversed or reflected prior to summation. If items were rated on 7-point rating scales where 1 = agree and 7 = disagree for positive statements, then responses for negative items need to be subtracted from 8 so that 1 = disagree and 7 = agree. Similarly, when responses are dichotomous and 0 = no and 1 = yes, then coding for negative statements should be reversed prior to summation and analysis.
4. Fourth, an individual's score is the sum of his or her responses across items (after appropriate reversal of some items). Reliability of the total score is calculated from the average correlation among items and the number of items (alpha or KR-20). Reliability of individual items is determined by each item's correlation to the total score (item-to-total correlation). All items should have a positive item-to-total correlation. (Items with a negative item-to-total correlation need to be reflected or omitted; see step 3).
5. The final items are selected with high item-to-total correlations, say 10 positive and 10 negative statements from the original 40. A 20 item summative scale should have a reliability coefficient greater than .80.

Development of reliable and valid attitudinal scales is usually an iterative process involving data collection from several samples. For example, Lewis et al. (1984) were interested in measuring stress in pre-adolescent children. Previous studies of stress contained items relevant to adults or items *thought to be relevant* for children. The investigators began with individual and small group interviews with 50–60 fifth and sixth graders and asked, "What happens that makes you feel bad, nervous, or worry?" From the responses to this question (three questions), the researchers compiled a list of 22 items agreed on by the group.

The degree to which the themes were well captured and expressed in existing scales of stress for children provides evidence for the validity of those scales. The degree to which the themes were mutually exclusive with existing scales challenges the valid use of such scales with children. The researchers determined that the themes were sufficiently unique to this population that they proceeded to create a new scale. Their next step was to pretest the 22 items as a questionnaire, rated on 5-point scales as to "How bad each would make you feel" and "How often each occurs." The results of the pretest indicated that two items were almost always rated as "not bad," and so were eliminated. The final 20-item test was then administered to 2,400 fifth graders.

RATING SCALES

Modifying existing materials or developing new materials involves making sure the content of the questions is appropriate for the population you are studying. Equally important is the format for responses to those questions. Cliff (1959) studied the effect of different descriptors that can be used to anchor rating scales and to help respondents interpret the points on a rating scale. He found that certain adverbs increase or decrease the value of an adjective by a predictable and measurable amount: “slightly” decreases an adjective by about half and “extremely” increases an adjective by about 50%. Thus, a rating scale constructed with “slightly pleasant,” “pleasant,” and “extremely pleasant” would have three ordinal categories with fairly equal intervals. Rating scales can be collected orally, if the task is simplified. For example, a 4-point rating scale can be presented orally (for phone administration or for someone who cannot read) by using two questions: First, “Is your health good or poor?” Then, “Is your health poor [#2] or very poor [#1]?” Or, “Is your health good [#3] or very good [#4]?”) (See Weller and Romney 1988, chapter on Rating Scales.)

Classification Studies

In a departure from knowledge tests and attitudinal scales where the answers are known, classification studies attempt to understand and describe the ways in which individuals classify items into categories. This technique helps us understand categories of things *according to informants*. For a set of items, similarity data are collected from respondents without directing them to the criteria for making comparisons; judgments are made only in terms of the similarity or difference between items. Similarity distinctions are very basic distinctions in all cultures. Formats appropriate for similarity data collection are: pile sorting of items and paired or triadic comparisons of items. Typically, responses are aggregated across informants and the similarity information is represented with a spatial plot or tree structure to summarize the relationships among items. Results reveal relevant categories and sub-groupings of items that are salient to informants.

A classification study has at least three parts. First, the set of items for study must be defined. Second, similarity between each pair of items is estimated. Third, the similarity data are represented with a spatial or tree model. Similarity information can be collected *directly* with judged similarity or *indirectly* with a measure of similarity between pairs of items across a series of questions (their similarity in profiles). Direct, judged similarity may be collected with the names of items written on cards and sorted into piles according to their similarity (*pile sorting*); with items presented in pairs and each pair is rated on the degree of similarity (*paired comparisons*), or items can be presented in sets of three and the most different item is selected (*triadic comparisons*).

PILE SORTING

After the set of items for study has been defined, the name of each item can be written on a card or visual stimuli (pictures or objects) can be used. Informants are asked to read or review all of the items and to put them into piles, so that similar items are together in the same pile. Instructions are deliberately kept at a general level: Group the items according to their similarity without providing any specific criteria or examples.

Individuals may make as many or as few piles as they wish. Pile sorting was originally described by Miller (1969) and is reviewed in detail in Weller and Romney (1988) (also see Bernard and Ryan 2010).

Judged-similarity data help us understand *informants*' categories or perceptions of items. Sometimes in research you may be faced with a list of informant-generated items and want to know if some are redundant and whether there are categories of items that can be used to summarize main themes, or how to reduce the number of items but retain some items from each important category, or simply to describe perceptions and subcategories of things. Pile sorting is a way to find the categories *as perceived by informants* and not coded by the investigator, although pile sorting by the investigators can be used to develop coding categories (Hsaio et al. 2006; Sayles et al. 2007).

Instructions are given to ask informants to sort the cards (or photos or things) into piles so that things that are similar are together in a pile. Things that belong together or are alike go together in a pile. Informants can make as many piles as they wish: "These are things that people have said. Please read all the cards and then sort them into piles, so that similar things are together in a pile, and different things are in different piles." Additionally, instructions can be added about the number of piles: "Please make two or more piles" or "Please make seven to nine piles." This is generally an easy task, and the fewer words on the cards, the easier it is.

Pile sorting has been used to describe illnesses (Breiger 1994; Ross et al. 2002; Weller 1983, 1984), HIV risky behaviors (Macauda et al. 2011; Stanton et al. 1993), drugs (Carlson et al. 2004), addictions (Penka et al. 2008), problems among homeless youth (Ensign and Gittelsohn 1998), and types of dental pain (Moore et al. 1986). The method can be also used to study perception of plants (Benz et al. 2007; Berlin 1992; Berlin et al. 1974; Calvet-Mir et al. 2008; Nolan 2002) and animals (Boster and Johnson 1989; Lopez et al. 1997). Some applications include the study of social networks (Freeman et al. 1988, 1989; Johnson and Miller 1983; Miller and Johnson 1981), recreational activities (Miller and Hutchins 1989; Parr 1996; Roberts and Chick 1979; Roberts and Natrass 1980; Roberts et al. 1981); concepts of success and failure (Freeman et al. 1981; Romney et al. 1979), pilot errors (Roberts et al. 1980), activities of the elderly (Harman 2001), and emotions (Alvarado 1998; Lutz 1982). Pile sorting has also been used to develop salient and reliable categories for coding of qualitative data (Hsaio et al. 2006; Sayles et al. 2007).

Kirk and Miller (1978) were interested in the perception of coca in South America and used pile sorting to discover if it was considered as a food product, a beverage, or a drug. They collected pile sort similarity data on 16 words, including foods, condiments, beverages, cigarettes, and drugs. They selected samples of 12 informants from each of 12 different sites: 2 cities in Colombia, 1 in Ecuador, and 6 locales in Peru (with 4 separate samples in Lima). Because Kirk and Miller used small, convenience samples, they used multiple samples to check the consistency or reliability of their results. Although a single, large representative sample would provide information on perceptions of coca; multiple, diverse, convenience samples can sometimes provide similar information—*if* the results are consistent across the diverse groups. If the results differ, then further work is necessary to discover what factors are associated with the difference. In this case, results were similar across samples, so the samples were combined.

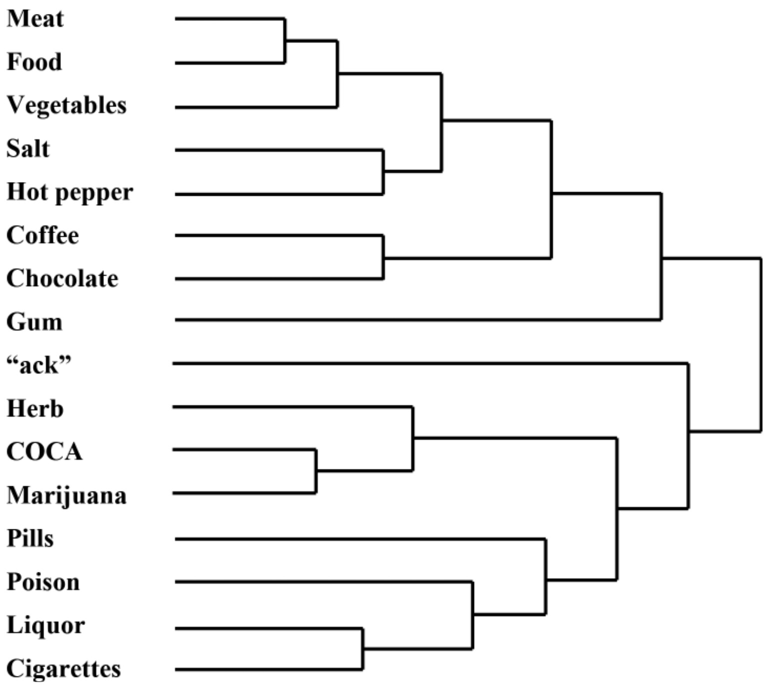

Figure 11.1. Perceived similarity among foods and drugs (adapted from Figure 2 in Kirk and Miller 1978, 144; reprinted with permission).

Kirk and Miller's classification or grouping of items appears here as Figure 11.1 in a dendrogram or tree-like representation (an adaptation of their "Figure 2 Diameter Method," p. 144); it is a taxonomy of the similarity between items from a hierarchical clustering analysis. Here, "meat" and "food" are the most similar pair and are linked together at the lowest level of the tree, indicating the highest level of similarity. A cluster of edible things is then formed with other foods and condiments: meat, food, and vegetables joins with salt, and hot pepper. The beverages, coffee and chocolate, also belong to this cluster. Another cluster contains the drugs: herb, COCA, and marijuana are in one subgroup; and liquor, cigarettes, poison and pills are in another. Thus coca, although chewed and often drunk as tea, is perceived to be a drug, similar to marijuana.

The pile sort is a widely used and quick way to estimate similarity among items for a group of people. Informants are asked to sort the items into piles so that things that are similar are in the same pile together. The task is easily understood and facilitates conversation. After an individual has finished sorting items, she or he can describe the groupings. The data are best used to describe a group of individuals, rather than a single individual, because the data are sparse. Information from each individual only indicates if an item is paired with another or not (without information on the degree of similarity). Thus, only dichotomous (yes/no or one/zero) data are collected for each pair of items from each individual. Because of the sparseness of information at the individual level, the method is recommended for larger samples of people (at least 30 people) and for larger sets of items (two dozen or more items, where other methods of data collection become prohibitive). Note also that informants must know how to read to sort words, although pictures or things can be sorted.

To collect pile sort data, write or type the names of items on cards (and number the backs of each card). Then shuffle (and randomize) the cards and present them to an informant. Ask the informant to sort the cards according to their similarity so that similar things are in a pile together. Responses can be recorded immediately or later, if the piles are preserved by putting colored cards between the piles and putting a rubber band around the total set. Responses are recorded by piles. For example, if someone sorts seven things into four piles:

Pile 1: 1, 2, 3
Pile 2: 4, 5
Pile 3: 6
Pile 4: 7

Here, seven items have been sorted into four piles: items #1, 2, and 3 are together; and items # 4 and 5 are together. Items # 6 and 7 were not put into piles with any other items. Similarity between *each pair* of the seven items can then be recorded into a square, symmetric table or matrix. Since items 1, 2, and 3 are together, each pair in the group (1 and 2, 2 and 3, 1 and 3) are tabulated as similar. Items 4 and 5 also occur together and are tabulated as similar. All other pairs are not perceived to be similar and are coded with zeros. (See Weller and Romney 1988; for more detail, also see Bernard and Ryan 2010.) Responses are tabulated into a matrix for each individual and then summed together into an aggregate matrix for the entire sample of informants. The tabulation of responses can be done by hand or with the aid of computer software. ANTHROPAC (Borgatti 1996) translates the pile sort information for each respondent into individual and group matrices. The matrices can then be analyzed in ANTHROPAC or in other statistical software to represent and see the clustering of items into groups and subgroups.

Variations on pile sorting include: allowing informants to "split" items, so that an item may go into more than one pile; constraining the number of piles an informant may make; or collecting successive pile sorts from each individual. Stefflre (1972) asked informants, when they were finished sorting items, if any items should go into more than one pile. Items or cards were then split and put into multiple piles. In the unconstrained version of the pile sort, informants may make as many or as few piles as they wish. In the constrained version; informants are instructed to make a specific number of piles, say between seven and nine piles (Romney et al. 1979). The constrained version of the pile sort attempts to control for individual differences in style; some individuals make finer discriminations between items (splitters) than others (lumpers). Burton (1975) proposed a method for assigning greater weight to the responses of splitters in an unconstrained sorting task.

Because of the strong effect of such style differences, sorting tasks are usually not appropriate for comparisons between informants (Arabie and Boorman 1973; Boorman and Arabie 1972; Boorman and Oliver 1973). Comparisons between informants, rather than items, can be made only with an equal number of piles per informant or with successive pile sorts (Boster 1986a; Truex 1977; and see Weller and Romney [1988] and Boster [1994] for more information on successive sorts). Successive pile sorting allows for more detailed information to be collected on each person (Lynch and Holmes 2011; Ross et al. 2011).

PAIRED COMPARISON AND TRIADS SIMILARITY DATA

Since similarity data technically concerns pairs of items, sets of items can be created and informants can be asked directly about each pair. The advantage of such a design is that more detailed information is collected per informant and these designs can be used orally with people who cannot read. With k items there are $k(k–1)/2$ pairs or relationships to be estimated. Pile sort similarity data provide only dichotomous information (two values; co-occur = 1, do not co-occur = 0) on the $k(k–1)/2$ pairs for each informant. A direct rating of pairs, say on a 9-point rating scale, provides a 9-point range of information for each pair for each informant. A triad design offers a measurement range equal to the number of times each pair occurs. Thus, a paired comparison (two at a time) or a triadic (three at a time) design collects the same type of information as the pile sort, but collects more detail (finer discriminations of similarity) from each informant. The tradeoff is that more information is collected per person, allowing for a smaller sample size and more reliable representation, but the tasks may be less interesting to informants than doing a pile sort.

In triad designs, items are systematically arranged into sets of three (see Weller and Romney 1988; also Coombs 1954). Usually informants are instructed to pick the "most different" item in each set, which, in turn, identifies the most similar pair (the two remaining items). Pairwise similarity is thus estimated from responses. Picking the most different item is a simple task and can be done orally. Triads are really the only practical way to collect similarity data orally. Because of that, it is the method preferred for interviewing people with low literacy levels. Psychologists, working in more controlled conditions like classroom data collection, sometimes collect much more detailed information. For example, because a triad of items actually contains three pairs, some ask informants to identify the *most* similar pair in each triad and the *least* similar pair (Coombs 1954). In that way, all three pairs within each triad are ranked (1 = least, 2, and 3 = most similar). This latter method is much more intensive than the simple "pick the most different one," and provides much more information per informant but is not practical for most field applications.

Tasks collecting judged similarity data with systematic comparisons of items can collect more detailed information per informant, but the task can be lengthy. With k items there are:

$k\ (k–1)\ /\ 2$ pairs in any set of items and
$k!\ /\ [3!\ (k–3)!]$ triads.

Thus, with 10 items there are 45 pairs and 120 triads; with 21 items there are 210 pairs and 1330 triads.

Because the paired comparison and triad designs quickly become cumbersome, there are special designs to limit the number of necessary subsets and still collect similarity judgments on pairs of items. These designs provide a systematic subset of possible comparisons. An incomplete cyclic design for paired comparisons may include only 30 to 40% of the possible pairs and still accurately represent all pairs (Burton 2003). For triads, *balanced-incomplete-block designs* can be found in Burton and Nerlove (1976) or in Weller and Romney (1988). Balanced-incomplete-block designs are identified by the number of items to be compared (*k*), the size of the subsets (2 = pairs, 3 = triads, etc.), and the number of times each pair appears (lambda). A complete triads design for 21 items contains 1,330 unique sets of three items, but only 70 triads are necessary if a design is created where each pair occurs only once. A lambda-one design for 21 items has a large enough number of items to provide interesting results and yet is simple enough to be administered orally in the field.

The first step in creating a triad design for a set of items is to select a balanced-incomplete-block design for the number of items that you have (see Weller and Romney 1988). Second, enumerate the sets as specified in the design. Third, randomize the order of the sets and the order of items within each set (see Weller and Romney 1988, 33–34). Failure to randomize items can lead to biased selections by informants and might confound results (Romney et al. 1979). ANTHROPAC (Borgatti 1996) has an option to develop and print the data collection forms for many of the triad designs as well as tabulate the responses into a similarity matrix. Informants are asked to pick the most different item in each set. The task may be preceded by an example or two, but the examples should have obvious answers, they should come from a different domain, and the correct answer in each example should be in a different position within the set (second, first, third item). The similarity matrix containing the aggregate responses across all informants (whether from pile sorting, triads, or paired comparisons) can be analyzed to determine the perception or categorizations for the group.

If pairs are rated, the first step is to list all possible pairs of items or to use a systematic subset of pairs. Remember that although there are $k(k{-}1)/2$ pairs in k items, there are systematic designs that can cut the number of necessary pairs in half (see Burton 2003). Second, the ordering of the pairs and the order of items within each pair is randomized. Third, a rating scale is created, where the smallest number indicates the least similar and the largest number indicates the most similar. Informants then judge the similarity of pairs of items on the rating scales. The rating scale value selected for each pair is tallied into a matrix.

Applications using triads to collect similarity data include the study of kinships terms (Romney and D'Andrade 1964), animals (Henley 1969), occupations (Burton and Romney 1975; Magaña, et al. 1995), illnesses (Lieberman and Dressler 1977; Nyamongo 2002; Weller 1983; Young and Garro 1982), personality descriptors (Burton and Kirk 1979; Kirk and Burton 1977), and emotions (Alvarado and Jameson 2011; Romney et al. 1997). Triadic comparisons have also been used to study ethnobotanical classifications (Reyes-Garcia et al. 2004; Ross et al. 2005).

In a study of emotion terms, Romney et al. (1997) compared monolingual English speakers' and monolingual and bilingual Japanese-speakers' similarity judgments of 15 emotion terms using triads (lambda 3) and paired comparisons (5-point rating scale) to judge the similarity of pairs of items. Figure 11.2 displays the similarity between terms and across the two monolingual samples in a spatial representation (adapted from Romney et al., 1997, Figure 2, p. 5491). Correspondence analysis was used to

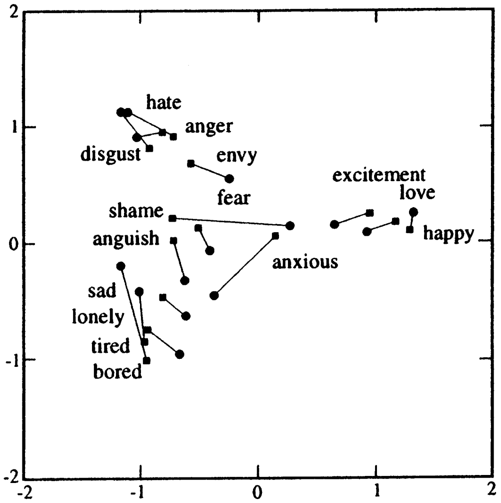

Figure 11.2. Spatial representation of similarity among emotion terms.

represent the similarity data in two dimensions. The figure may be interpreted as a "map"; where closeness in the picture indicates similarity. Thus, "disgust," "anger," and "hate" are perceived as similar to one another and different from "sad" and "happy." Differences between the two samples are negligible for four terms, small for eight terms (e.g., "disgust/*mukatsuku*," "hate/*kirai*," and "anger/*haragatatsu*"), and large for three terms ("shame/*hazukashii*," "anxious/*fuan*," and "bored/*tsumaranai*"). Romney et al. conclude that there is a substantial amount of shared meaning in emotions between the English and Japanese samples.

In a study of societal problems, Wish and Carrol (presented in Kruskal and Wish 1990, 36–41) asked 14 individuals to rate 22 societal problems in terms of their similarity. Rating scales were used to collect judged similarity on all 231 pairs. Additional rating scales were used to rate the 22 problems on other, specific dimensions to aid in interpretation of the similarity dimensions (e.g., the degree to which each problem affects most people). The similarity between the 22 items (aggregated across informants) was represented spatially in three dimensions using multidimensional scaling (MDS). MDS is another multivariate analysis appropriate for the analysis of inter-item similarity data. Similarity relations are translated into distances creating a spatial representation like a map. Thus, closeness in the representation indicates similarity.

The three dimensions that best explained informants' perception of the societal problems were the degree to which the problem affected most people, the degree to which the problem was the responsibility of local government, and the degree to which the problem was technological. Figure 11.3 shows the latter two dimensions (adapted from Kruskal and Wish 1990:40, Figure 12b). In the lower-left quadrant of the figure are problems ("Failures in welfare") thought to be the responsibility of local government; in the upper-right quadrant are those that are not the responsibility of local government ("Inflation"). Technological problems are in the lower-right quadrant and nontechnological problems are in the upper-left.

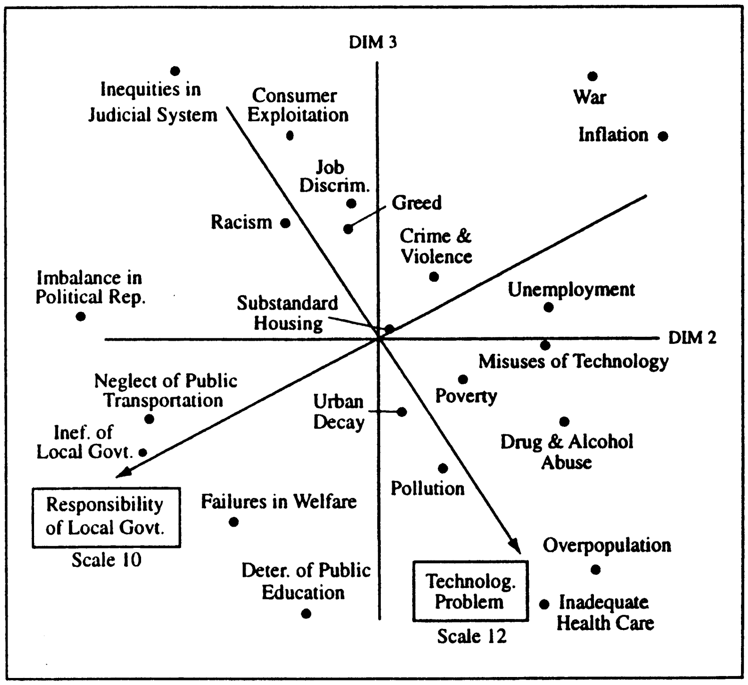

Figure 11.3. Spatial representation of similarity among societal problem.

SENTENCE-SUBSTITUTION OR PROFILE DATA

Similarity between items can be collected directly with judgments of similarity (pile sorting, triads, or paired comparisons) or similarity can be estimated indirectly, between the "profiles" of pairs of items across a series of questions. For example, D'Andrade et al. (1972) asked about the attributes of 50 illness terms by repeating the set of 50 attribute questions for each illness (2,500 questions); then they estimated the similarity between the illnesses from their proportion of shared attributes. This interviewing procedure—the systematic comparison of a set of items with a set of attributes or features—is sometimes called sentence-substitution data collection because the items are systematically substituted into sentence frames containing the attributes for the interview. Similarly, information can be collected in this way for social relationships within a group of people, where people are asked to rate the social relationship, the frequency of interaction, or the similarity between themselves and each person in the group (Wasserman and Faust 2009, 45–55).

Sentence-substitution interviews begin with two related lists. The first list is the set of domain items and the second is a set of statements about the domain items. The latter list may include descriptive statements, attributes, features, or uses (behaviors) relevant to the domain items. In the interview, each item is paired with every attribute and informants are asked to judge the acceptability or veracity of the newly formed statement. The task is easy to understand and may be administered orally. For oral administration, a matrix can be used to indicate the intersection of the two lists (rows as attributes and columns as domain items) and each question regarding each attribute can be read by systematically substituting each of the domain items. For written administration, all statements should be completely written out with correct syntax. Responses can be dichotomous (yes/true or no/false), or a rating scale can be used for each question. Usually, the responses of informants are aggregated into a single item-by-attribute table, where responses are represented with the modal response (for categorical data), averaged (for rating scales), or cultural consensus is used to estimate the best answers (see the section below on Exploring Beliefs).

Similarity between *items* may be calculated from their shared attributes (or similarity between attributes can be calculated from their co-occurrence in items). From either, a square symmetric matrix of similarities is obtained. In D'Andrade et al.'s (1972) study of illnesses and illness attributes, the similarity between each pair of illnesses (across attributes) was calculated with a Pearson correlation coefficient. The item-by-item correlation matrix was represented with MDS and hierarchical clustering. Clustering results can be used to interpret the similarity between items and to reorder the rows (attributes) and columns (items) in the aggregate item-by-attribute response table so that the joint item-attribute clusters can be seen. This multi-step process (correlation matrices, multivariate analysis, and reordering of the rows and columns to see patterns) can now be accomplished in a single step with correspondence analysis (Weller and Romney 1990).

Several applications have used sentence-substitution questions to explore illnesses and their symptoms and treatments (D'Andrade et al. 1972; Garro 1986; Maupin et al. 2011, Ross et al. 2012; Stefflre 1972; Young 1978). Other examples include looking at how Peace Core workers are perceived (types of people by behaviors, in Stefflre 1972) and beliefs about adolescent punishments (adolescent behaviors by punishments, in Weller et al. 1987). D'Andrade (1976) and Young (1978) also attempted to identify attributes that best differentiated illness categories. A strength of this method is the linking of two related sets of items. Sentence-substitution data provide rich and valuable information, but the interview can be lengthy. Interviews like Stefflre's (1972) and D'Andrade et al.'s (1972) comparison of 50 items and 50 attributes (2,500 questions) were carried out over a few days and informants were reimbursed for their time.

A more general form of this type of interviewing is the systematic collection of information on any two related lists of items to create a profile of information for one set of items based on the second set of items. For example, interviews with members of a small face-to-face social group (whole network) may ask that each group member "Name the individuals with whom you interact the most," "Name the three people with whom you interact the most," or "Rate each person in terms of how much you interacted with them in the past month." Although these three questions vary from unconstrained and constrained dichotomous responses (named = 1, not named = 0) to rated (or ranked) responses for each person in the group, the information collected refers to the set of all group members.

The two related lists each contain the names of all members: The first list indicates the informant or person interviewed and the second list indicates the informant's responses or choices for everyone else in the group. Similarity is then calculated between informants, based on their profile of responses/choices. Similarity in their pattern of choices may be calculated with a Pearson correlation coefficient (phi) or other measure and represented spatially with MDS, correspondence analysis, or graph theoretic techniques (Wasserman and Faust 2009).

RELIABILITY AND VALIDITY OF SIMILARITY DATA REPRESENTATIONS

Data collection and analysis for the study of classifications include three steps: (1) getting the similarity data; (2) tabulating the data into a single table or matrix for each group; and (3) getting a descriptive model or representation of the similarity relationships. Similarity data may be collected directly with pile sorting, triads, or paired comparisons or measured indirectly from the shared attributes across items. With direct judged similarity, a similarity matrix is created for each individual and then the matrices are summed together into a single matrix. Tabulation of similarity can be done by hand or by computer (Borgatti 1996). With indirect measures of similarity, a matrix of similarity coefficients (e.g., Pearson correlation coefficients) is generated by a computer program. Finally, the aggregate similarity between items in the form of a square, symmetric matrix of similarities is usually represented with a descriptive, visual multivariate technique.

Descriptive statistical analyses used for the representation of similarity data include clustering (Mezzich and Solomon 1980), nonmetric MDS (Kruskal and Wish 1990; Mezzich and Solomon 1980), factor analysis or principal components analysis (Weller and Romney 1990), and correspondence analysis (Weller and Romney 1990). These analyses are available in most major statistical packages. Hierarchical clustering represents the relationships between items in a tree-like structure or dendogram, like a taxonomy.

Both MDS and correspondence analysis provide spatial representations of data so that similar items are closer together on a map or plot of items, as can factor analysis or principal components. Correspondence analysis is a sister of principal components, appropriate for scaling qualitative/categorical data. Correspondence analysis allows for the simultaneous scaling of items and attributes in the same spatial configuration, facilitating a sentence-substitution data analysis.

A variety of studies have been undertaken to test the validity and reliability of using one of these multivariate models to represent similarity data. Validity concerns the degree that these multivariate models actually represent how people perceive and think about the items. Simple exercises include submitting a set of interpoint distances (where similarity is the degree of propinquity) for analysis and checking to see if the same information can be retrieved. As mentioned, although there are many types of clustering methods, the average-link method (Sokal and Sneath 1963) tends to outperform others in being able to retrieve known structures (Milligan 1980). Green and Carmone (1970) illustrate MDS's ability to translate such information into an accurate "map" with a configuration of points representing the letters "A" and "M"; Kruskal and Wish do so with a map of the United States. Weller and Romney (1990) repeat Kruskal and Wish's example and show that correspondence analysis also can translate inter-city mileages into a map. Magaña et al. (1981) studied the perception of a college campus and compared estimates of distances, triad judgments, and distances from hand-drawn maps and found the MDS representations to accurately reflect true distances.

A more complicated form of validation concerns the degree to which such models are accurate representations of what people think. Judged similarity data, when represented with multivariate techniques, predict memory performance, judgments, and reasoning task performance. Friendly (1977) used hierarchical clustering and MDS models of free-recall listing and similarity data to successfully predict memory performance tasks. Similarly, Romney et al. (1993) used a MDS model of similarity data to predict list length in a free-recall listing task. Hutchinson and Lockhead (1977) used MDS model inter-item distances to predict reaction time judgments concerning similarity. Rumelhart and Abramson (1973) used a MDS model to predict informants' responses on analogical reasoning tasks concerning animals.

Most studies have found similarity judgments to be highly reliable. This means that there often tends to be little intracultural variation in these judgments. Romney, Smith et al. (1979) in a study of concepts of success and failure, compared results across several samples and found them to be highly concordant. A check on the internal consistency (agreement) in similarity judgments is an important step in justifying an aggregate representation. Similarity between items, using different methods of estimating similarity, is usually concordant (compare D'Andrade et al. 1972 and Weller 1983; and see Young and Garro 1982; Romney et al. 1997).

Exploration of Specific Beliefs

A series of questions on a single topic may be used to evaluate knowledge, attitudes, or beliefs. In studies of beliefs, however, the purpose is to discover the answers and not to measure deviance from a standard. Thus, only the original responses are used and they are not scored, transformed, or recoded as for attitudinal scales. Studies focusing on beliefs are similar to classification studies, except that classification studies rely on similarity data without reference to specific criteria and studies of beliefs often focus on specific criteria. *Questions for studies of beliefs are written following the same process as for studies concerning knowledge tests and attitude scales.* Question formats include: open-ended questions requesting short answers or phrases; questions with predetermined multiple choice response categories (including dichotomous yes/no or true/false); requests to rank order items on a specific topic; and open-ended questions requesting numeric estimates (like frequencies, distances, or ages). Typically, beliefs are estimated by aggregating responses across informants.

To assess beliefs, interviews are conducted with a series of statements or questions all on the same topic, all in the same format, and all at the same level of difficulty. As with all interview materials, the items should be relevant to the informants and should be developed from content obtained in open-ended interviews. Clear and simple wording should be used, so that each question is understood in the same way by each person. The actual format of questions is guided by the purpose of the study. If the purpose is to discover detailed beliefs (e.g., a cultural model of the causes, symptoms, and treatments for an illness), then an appropriate format may be a series of yes/no or true/false questions covering attributes of the illness (e.g., Weller et al. 2012). With yes/no or true/false questions, care must be taken to balance the list with approximately half positive (true or yes) and half negative (false or no) statements.

Alternatively, a project might focus on a single question, "What causes breast cancer?" (Chavez et al. 1995), and a set of possible causes can be rank ordered from most to least likely causes. Or possible sources of support can be judged as appropriate for scenarios where help might be needed, "To whom would you go for advice or support?" (Berges et al. 2006; Dressler et al. 1997). Or a researcher may ask simple open-ended questions such as asking informants to identify plants (Boster 1986b).

Questions may look like those for a knowledge test or an attitudinal scale; the difference is how responses are handled and analyzed. A description of beliefs involves a summarization or aggregation of responses for each question. Intuitively, the best estimate of an answer is provided by the majority response or an average of responses across informants (D'Andrade 1987). Such measures, called *central tendency* measures in statistics, are the best single description of responses to a question. Thus, open-ended or categorical responses are best described by the majority or modal response, and ranked or interval data are best described by the median (midpoint) or mean (average) response.

Aggregate measures, however, are accurate only to the degree that there is little to moderate variability in responses. The notion of homogeneity in responses for a single question can be generalized to a set of questions. Homogeneity across informants' responses for a series of questions can be assessed with a measure of agreement. Field data indicate that agreement is related to accuracy (Young and Young 1962); if you ask three people where the post office is and they all give identical answers, then it is more likely that the information is true and correct, than when their answers conflict. The relation between consistency and validity can be expressed as a general principle of aggregation. *The accuracy of aggregated responses is a function of the agreement among informants and the number of informants* (the Spearman-Brown Prophesy Formula, described in Weller and Romney 1988). In other words, the agreement between *each pair* of informants is measured with a Pearson correlation coefficient and averaged across all pairs of informants; the higher the average agreement among informants, the fewer informants are necessary to achieve an accurate estimate of the "true" answers from an aggregation of their responses (see also Weller 1987). Thus, shared beliefs can be estimated by combining the responses of informants, if there is sufficient agreement among informants.

The cultural consensus model estimates culturally appropriate answers to a set of questions and the degree to which each informant shares those answers (for an overview, see Weller 2007). It assumes that the ethnographer does not know the answers to the questions or how much each informant knows about the domain under consideration. The analysis determines *if* there are highly shared beliefs and, if so, estimates the answer for each question and how much each informant knows those answers. The model also includes a method for estimating the number of informants needed to provide given levels of confidence in the answers for different levels of shared cultural knowledge. With highly shared beliefs, accurate results can be obtained with few informants.

Within cultural consensus theory, there are two models or approaches: formal and informal. The formal model is a psychological process model of how questions are answered with varying degrees of knowledge and bias that estimates the knowledge levels of respondents and likelihood that specific answers are correct (Romney et al. 1986). The model can only accommodate categorical responses, such as multiple choice data, including dichotomous data (yes/no or true/false) or open-ended responses (a word or short phrase). The analysis is similar to a factor analysis of people, but requires special software, as Bayesian methods are used to solve for estimates.

Categorical responses can be accommodated by the formal cultural consensus model. The formal model is appropriate for open-ended responses (a series of questions requesting a single word or short phrase), lists of statements requiring dichotomous choices (true/false or yes/no), dichotomous judgments of statements formed by linking two related lists (sentence-frame substitutions), and multiple choice responses. Open-ended questions were used by Boster (1986b), who walked Jivaroan women through a garden and asked them to name plants. Extensive work has been done with dichotomous responses (true/false and yes/no), especially on illnesses: AIDS (Baer et al. 1999b; Baer et al. 2004; Trotter et al. 1999), asthma (Pachter et al. 2002), the common cold (Baer et al. 1999a), diabetes (Smith 2012; Weller et al. 2012), folk illnesses (*nervios*, Baer et al. 2003; *empacho*, Weller et al. 1993; *susto*, Weller et al. 2002), and maternal health knowledge and infant health (Miller 2011). A true/false format also was used to explore beliefs about pollution and safety of seafood (Johnson and Griffith 1996). Sentence substitutions can be used to find normative answers to the joint assertions formed by combining two related lists (Garro 1988; Maupin et al. 2011; Ross et al. 2012).

Multiple choice responses have been used to study gender concepts (de Munk et al. 2002) and shared knowledge about fish habitats and behaviors (Garcia-Quijano 2009). Garcia-Quijano (2009) asked five types of questions for 16 different species of fish: For example, fishermen were asked about each fish's habitat (bays, mangrove channels, sand bottoms, seagrasses, reefs, open water, mud bottoms, and deep water); and their seasons (early winter, late winter, spring, early summer, late summer, or fall).

The informal cultural consensus model is the most widely accessible model with the least assumptions about the data (Romney et al. 1987). The informal model is a collection of analytical procedures that approximate the results of the formal model, estimating answers or the ordering of answers on a specific construct and estimating the degree to which each person's responses correspond with that ordering. For this model, items are typically ordered from most to least on a specific concept. Conceptually, this model averages responses across people to estimate answers and then correlate each person's answers with the aggregated answers of the group to estimate their correspondence to the group consensus or their "cultural knowledge." This approach includes a reliability or factor analysis of people rather than items and can be run in most major statistical software packages (see Weller 2007). For example, Caulkins (2001; Trosset and Caulkins 2001) studied ethnic identity by having people in different regions of the United Kingdom rate scenarios (e.g., a "child performing a song for family guests while standing on the kitchen stool") on how "Welsh" each was. Consensus can also be used to identify shared values and norms within organizations (Caulkins and Hyatt 1999, Jaskyte and Dressler 2004, 2005; Smith et al. 2010).

Ranked responses are accommodated in the informal cultural consensus model. Applications using the informal model include studies of illnesses, social support, and occupational prestige to examine shared beliefs within and across subgroups. Chavez et al. (1995) compared the beliefs of four different groups of Latinas and one group of physicians by having each group of informants rank order 30 potential causes of breast cancer. Magaña et al. (1995) compared U.S., Mexican, and Guatemalan perceptions of socioeconomic status and prestige by comparing informants' rank-orderings of occupations. Koster and Tankersley (2012) examined perceived hunting ability in dogs by having hunters rank the dogs and Koster et al. (2010) examined hunters' desire for particular meat flavors by ranking meats in terms of their desirability and taste. In a high AIDS mortality area of Africa, Kiš (2007) had residents rank order reasons for attending a funeral to understand changing values about why people would and would not attend a funeral.

Fully ranked data may be collected with paper and pencil, interactively with cards, and orally (for detail on ranking methods, see Weller and Romney 1988). Respondents may be asked to rank *k* items by putting a number next to each item using paper-and-pencil data collection. Or names of items can be written on cards and spread out on a table and the informant is asked to pick up the most important item, then the next important item, and so on until all the cards are ordered. For both cases, data recording would have the list of items and the researcher would record the *rank* assigned to each. Smith et al. (2004; Smith et al. 2010) used cards to collect information on patient and physician priorities in primary care. Balanced incomplete designs also allow for the collection of information orally by presentation of subsets (triads or pairs) of items, and combining responses to create a full rank order of items for each person. Cain et al. (2011) studied cultural norms on the appropriateness of genitalia terms by having respondents rank terms in sets of three. Reyes-Garcia et al. (2004) had informants rank plants on their usefulness by orally presenting the plants in pairs and then combining responses.

Rating scales can sometimes be used, but care must be taken to ensure that positive and negative items are used and that the range of scale values is used by each person. One means for doing this is to use a constrained rating scale task called a Q-sort (Weller and Romney 1988). The rating scale is typically arrayed on a table and informants are instructed to put each item on a rating scale value, with the caveat that the researcher constrains the task by limiting how many items can go on each value. For example, a researcher can request that 16 items be placed so that one item is rated as "1," two as "2," three as "3," four as "4," three as "5," two as "6," and one as "7." This can be accomplished by having the desired number of pockets under each rating scale value so it is clear how many items "go with" each value. Rocha (2005, 363) used a Q-sort to collect ordered data: For example, photographs of 34 crops were ordered on their difficulty to be tended and were put into five piles from most to least difficult, and soil types and fertilizers were rated on a 3-point scale with three piles, ensuring that all respondents used all three ordinal categories.

Similarity data may also be used, *if* similarity is collected with a systematic comparison method (triads or paired-comparisons) or with successive pile sorts. Reyes-Garcia et al. (2004) examined intra-cultural variation in shared knowledge of plants by collecting judged similarity data with triadic comparisons, calculating the similarity between pairs of items, then representing shared knowledge with the cultural consensus model. Romney et al. (1997) and Alvarado and Jameson (2011) used triad similarity data to study normative meaning of emotion terms cross-culturally. Lynch and Holmes (2011) used successive pile sorts to study lay perceptions of food group categories and Ross et al. (2011) used successive sorting to study illnesses.

An important application has been the comparison of what people have (social support, material goods, etc) with what local norms indicate they should have. Dressler et al. (1997) studied cultural preferences for different sources of social support by having informants rank order possible sources of support in terms of their appropriateness in different scenarios. The agreement between individual circumstances and group norms has been called cultural consonance (Dressler 1996; Dressler et al. 1997; Dressler et al. 2005; Dressler et al. 2012; Reyes-Garcia et al. 2010).

The validity and accuracy of estimates provided by the informal consensus model are illustrated by Romney et al.'s (1987) study on causes of death, where rankings of the frequency of perceived causes of death in the population were compared with actual mortality rates. Dawes's (1977) study on estimated heights as compared to actual heights illustrates validity, although the study preceded the formalization of the cultural consensus model. Webster et al. (2002) correlated answers obtained from consensus rankings to other performance measures and personality characteristics. Also, Romney and Weller (1984) used individual correspondence to group answers to predict individual accuracy in reporting social interaction patterns.

While simple aggregation of responses with moderate to high agreement is a sound procedure, there are some limitations in the application of consensus theory to response data. Clarity of questions is always an issue. Questions must be clear and understandable to all participants and interpreted in the same way. Missing data can be an issue, and care must be taken to get answers to as many questions as possible. The formal model assumes that there is no response bias, although newer formulations of the model can estimate the amount of bias as well as item difficulty (Karabatsos and Batchelder 2003; Oravecz et al. 2014). Response bias can have many forms; with field data, it may be the simple pattern of respondents to tend to say "yes" to all questions about which they have doubt or conversely to say "no" when in doubt. It is also important to note that "I don't know" currently cannot be handled as a response choice, but is instead accounted for with guessing. Another issue is to be sure that positive and negative statements are both represented; a very skewed distribution (very few positive answers or very few negative answers) can affect the model's estimates.

SUMMARY

Sociological and psychological literature offer many lessons about writing questions. The authors in those areas have extensive experience in writing questions for surveys and tests, important for all types of questionnaires and interview materials. Questions should be simple, clear, and interpreted in the same way by everyone (the respondent and the researcher).

First, preliminary or ethnographic interviewing and free listing provide valuable information for the development of content for questions. Second, when questionnaires are drafted, pilot testing with responses and/or interpretation of questions provides valuable feedback on the clarity of questions. Great insight can be gained with as few as three interviews. Finally, development of questions for surveys, knowledge tests, attitude scales, or belief studies all involve the same processes. Time invested during development can save grief in the interpretation stage. There are no short cuts.

REFERENCES

Alvarado, N. 1998. *A reconsideration of the structure of the emotion lexicon. Motivation and Emotion* 22: 329–44.

Alvarado, N., and K. A. Jameson. 2011. Shared knowledge about emotion among Vietnamese and English bilingual and monolingual speakers. *Journal of Cross-Cultural Psychology* 42: 963–82.

Arabie, P., and S. A. Boorman. 1973. Multidimensional scaling of measures of distances between partitions. *Journal of Mathematical Psychology* 10: 148–203.

Baer, R. D., S. C. Weller, J. E. Garcia de Alba Garcia et al. 2003. A cross-cultural approach to the study of the folk illness "nervios." *Culture, Medicine, and Psychiatry* 27: 315–37.

Baer, R. D., S. C. Weller, J. E. Garcia de Alba et al. 2004. A comparison of community and physician explanatory models of AIDS in Mexico and the U.S. *Medical Anthropology Quarterly* 18: 3–22.

Baer, R. D., S. C. Weller, L. M. Pachter et al. 1999a. Cross-cultural perspectives on the common cold: Data from five populations. *Human Organization* 58: 251–60.

Baer, R. D., S. C. Weller, L. M. Pachter et al. 1999b. Beliefs about AIDS in five Latin and Anglo-American populations: The role of the biomedical model. *Anthropology and Medicine* 6: 13–29.

Benz, B., H. Perales, and S. Brush. 2007. Tzeltal and Tzoltil farmer knowledge and maize diversity in Chiapas, Mexico. *Current Anthropology* 48: 289–300.

Berges, I. M., F. Dallo, A. DiNuzzo et al. 2006. Social support: A cultural model. *Human Organization* 65: 420–29.

Berlin, B. 1992. *Ethnobiological classification: Principles of categorization of plants and animals in traditional societies*. Princeton, NJ: Princeton University Press.

Berlin, B. O., D. Breedlove, and P. Raven. 1968. Covert categories and folk taxonomies. *American Anthropologist* 70: 290–99.

Berlin, B. O., D. Breedlove, and P. Raven. 1973. General principles of classification and nomenclature in folk biology. *American Anthropologist* 75: 214–44.

Berlin, B. O., D. Breedlove, and P. Raven. 1974. *Principles of Tzeltal plant classification: An introduction to the botanical ethnography of a Mayan-speaking people of Highland Chiapas*. New York: Academic Press

Berlin, B. O., and P. D. Kay. 1969. *Basic color terms: Their universality and evolution*. Berkeley: University of California Press.

Bernard, H. R. 2013. *Social research methods: Qualitative and quantitative approaches*, 2nd ed. Thousand Oaks, CA: Sage.

Bernard, H. R., P. D. Killworth, D. Kronenfeld, and L. Sailer. 1985. On the validity of retrospective data: The problem of informant accuracy. *Annual Review in Anthropology* 13: 495–517.

Bernard, H. R., P. D. Killworth, and L. Sailer. 1980. Informant accuracy in social network data IV: A comparison of clique-level structure in behavioral and cognitive network data. *Social Networks* 2: 191–218.

Bernard, H. R., and G. W. Ryan. 2010. *Analyzing qualitative data*. Los Angeles: Sage.

Boorman, S. A., and P. Arabie. 1972. Structural measures and the method of sorting. In *Multidimensional scaling: Theory and applications*, Vol. 1, ed. R. Shepard, A. K. Romney, and S. B. Nerlove, 225–49. New York: Seminar Press.

Boorman, S. A., and D. C. Olivier. 1973. Metrics on spaces of finite trees. *Journal of Mathematical Psychology* 10: 26–59.

Borgatti, S. P. 1996. ANTHROPAC. Version 4.98. Natick, MA: Analytic Technologies.

Boster, J. S. 1986a. Can individuals recapitulate the evolutionary development of color lexicons? *Ethnology* 25: 61–74.

Boster, J. S. 1986b. Exchange of varieties and information between Aquaruna manioc cultivators. *American Anthropologist* 88: 428–36.

Boster, J. S. 1994. The successive pile sort. *Field Methods* 6: 11–12.

Boster, J. S., and J. C. Johnson. 1989. Form or function: A comparison of expert and novice judgments of similarity among fish. *American Anthropologist* 91: 866–89.

Bousfield, W. A., and W. D. Barclay. 1950. The relationship between order and frequency of occurrence of restricted associative responses. *Journal of Experimental Psychology* 40: 643–47.

Bradburn, N. M., S. Sudman, and B. Wansink. 2004. *Asking questions: The definitive guide to questionnaire design for market research, political polls, and social and health questionnaires*. New York: John Wiley & Sons.

Breiger, W. R. 1994. Pile sorts as a means of improving the quality of survey data: Malaria illness symptoms. *Health Education Research* 9: 257–60.

Brewer, D. 2002. Supplementary interviewing techniques to maximize output in free-listing tasks. *Field Methods* 14: 108–18.

Brislin, R. W. 1986. The wording and translation of research instruments. In *Field methods in cross-cultural research*, ed. W. J. Lonner and J. W. Berry, 137–64. Thousand Oaks, CA: Sage.

Bukov, A., I. Maas, and T. Lampert. 2002. Social participation in very old age: Cross-sectional and longitudinal findings from BASE. *Journal of Gerontology: Psychological Sciences* 57B: 510–17.

Burt, R. S. 1984. Network items and the General Social Survey. *Social Networks* 6: 293–340.

Burt, R. S. 1986. A note on sociometric order in the General Social Survey. *Social Networks* 8: 149–89.

Burton, M. L. 1975. Dissimilarity measures for unconstrained sorting data. *Multivariate Behavioral Research* 10: 409–24.

Burton, M. L. 2003. Too many questions? The uses of incomplete cyclic designs for paired comparisons. *Field Methods* 15: 115–30.

Burton, M. L., and L. Kirk. 1979. Sex differences in Maasai cognition of personality and social identity. *American Anthropologist* 81: 841–73.

Burton, M. L., and S. B. Nerlove. 1976. Balanced designs for triad tests: Two examples from English. *Social Science Research* 5: 247–67.

Burton, M. L., and A. K. Romney. 1975. A multidimensional representation of role terms. *American Ethnologist* 2: 397–407.

Cain, D., S. Shensul, and R. Mlobeli. 2011. Language choice and sexual communication among Xhosa speakers in Cape Town, South Africa: Implications for HIV prevention message development. *Health Education Research* 26: 476–88.

Calvet-Mir, L., V. Reyes-Garcia, and S. Tanner. 2008. Is there a divide between local medicinal knowledge and Western medicine? A case study among Native Amazonians in Bolivia. *Journal of Ethnobiology and Ethnomedicine* 4: 18–28.

Canales, M., T. Hernandez, J. Caballero et al. 2005. Informant consensus factor and antibacterial activity of the medicinal plants used by the people of San Rafael Coxcatlán, Puebla, Mexico. *Journal of Ethnopharmacology* 97: 429–39.

Carlson, R. G., J. A. McCaughan, R. S. Falck et al. 2004. Perceived adverse consequences associated with MDMA/Ecstasy use among young polydrug users in Ohio: Implications for intervention. *International Journal of Drug Policy* 15: 265–74.

Caulkins, D. D. 2001. Consensus, clines, and edges in Celtic cultures. *Journal of Cross-Cultural Research* 35: 109–26.

Caulkins, D. D., and S. B. Hyatt. 1999. Using consensus analysis to measure cultural diversity in organizations and social movements. *Field Methods* 11: 5–26.

Chavez, L. R., F. A. Hubbell, J. M. McMullin et al. 1995. Structure and meaning in models of breast and cervical cancer risk factors: A comparison of perceptions among Latinas, Anglo women, and physicians. *Medical Anthropology Quarterly* 9: 40–74.

Cliff, N. 1959. Adverbs as multipliers. *Psychology Review* 66: 27–44.

Conklin, H. 1969. Lexicographical treatment of folk taxonomics. In *Cognitive anthropology*, ed. S. Tyler, 41–59. New York: Holt, Rinehart and Winston.

Coombs, C. H. 1954. A method for the study of interstimulus similarity. *Psychometrika* 19: 183–94.

D'Andrade, R. G. 1974. Memory and the assessment of behavior. In *Measurement in the social sciences*, ed. H. M. Blalock, 139–86. Chicago: Aldine.

D'Andrade, R. G. 1976. A Propositional Analysis of U.S. American beliefs about illness. In *Meaning in anthropology*, ed. K. Basso and H. Selby, 155–80. Albuquerque: University of New Mexico Press.

D'Andrade, R .G. 1987. Modal responses and cultural expertise. *American Behavioral Sciences* 31: 194–202.

D'Andrade, R .G. 1995. *The development of cognitive anthropology*. Cambridge: Cambridge University Press.

D'Andrade, R. G., N. Quinn, S. B. Nerlove, and A. K. Romney. 1972. Categories of disease in American-English and Mexican-Spanish. In *Multidimensional scaling: Theory and applications in the behavioral sciences*, Vol. 2, ed. A. K. Romney, R. Shepard, and S. B. Nerlove, 9–54. New York: Seminar Press.

Dawes, R. M 1977. Suppose we measured height with rating scales instead of rulers. *Applied Psychological Measurement* 1: 267–73.

de Munck, V. C., N. Dudley, and J. Cardinale. 2002. Cultural models of gender in Sri Lanka and the United States. *Ethnology* 41: 225–61.

DeWalt, B. R. 1979. *Modernization in a Mexican ejido: A study in economic adaptation*. Cambridge: Cambridge University Press.

Dickson, P. R., R. F. Lusch, and W. L. Wilkie. 1983. Consumer acquisition priorities for home appliances: A replication and re-evaluation. *Journal of Consumer Research* 9: 432–35.

Dressler, W.W. 1996. Culture and blood pressure: Using consensus analysis to create a measurement. *Field Methods* 8: 6–8.

Dressler, W. W., M. C. Balieiro, and J. E. dos Santos. 1997. The cultural construction of social support in Brazil: Associations with health outcomes. *Culture, Medicine, and Psychiatry* 21: 303–35.

Dressler, W. W., M. C. Balieiro, R. P. Ribeiro, and J. E. dos Santos. 2005. Cultural consonance and arterial blood pressure in urban Brazil. *Social Science & Medicine* 61: 527–40.

Dressler, W. W., K. S. Oths, M. C. Balieiro et al. 2012. How culture shapes the body: Cultural consonance and body mass in urban Brazil. *American Journal of Human Biology* 24: 325–31.

Ensign, J., and J. Gittelsohn. 1998. Health and access to care: Perspectives of homeless youth in Baltimore City, USA. *Social Science and Medicine* 47: 2087–99.

Fern, E. F. 1982. The use of focus groups for idea generation: The effects of group size, acquaintanceship, and moderator on response quantity and quality. *Journal of Marketing Research* 19: 1–13.

Fink, A. 2003. *The survey kit*, 2nd ed. Thousand Oaks, CA: Sage.

Fowler, F. J. 1995. *Improving survey questions design and evaluation*. Applied Social Research Methods Series, Vol. 38. Newbury Park, CA: Sage.

Fowler, F. J. 2009. *Survey research methods*, 4th ed. Applied Social Research Methods Series. Newbury Park, CA: Sage.

Frake, C. O. 1964. Notes on queries in ethnography. *Transcultural Studies in Cognition* 66: 132–45.

Freeman, H. E., R. E. Klein, J. Kagan, and C. Yarbrough. 1977. Relations between nutrition and cognition in rural Guatemala. *American Journal of Public Health* 67: 233–39.

Freeman, H. E., A. K. Romney, J. Ferreira-Pinto et al. 1981. Guatemalan and U.S. concepts of success and failure. *Human Organization* 40: 140–45.

Freeman, L. C., and R. Danching. 1997. An international comparative study of interpersonal behavior and role relationships. *L'Année Sociologique* 47: 89–115.

Freeman, L. C., S. C. Freeman, and A. G. Michaelson. 1988. On human social intelligence. *Journal of Social and Biological Structures* 11: 415–25.

Freeman, L. C., S. C. Freeman, and A. G. Michaelson. 1989. How humans see social groups: A test of the Sailer-Gaulin models. *Journal of Quantitative Anthropology* 1: 229–38.

Freeman, L. C., A. K. Romney, and S. C. Freeman. 1987. Cognitive structure and informant accuracy. *American Anthropologist* 89: 310–25.

Friendly, M. L. 1977. In search of the M-gram: The structure of organization in free recall. *Cognitive Psychology* 9: 188–249.

García-Quijano, C. G. 2009. Managing complexity: Ecological knowledge and success in Puerto Rican small-scale fisheries. *Human Organization* 68: 1–17.

Garro, L. C. 1986. Intracultural variation in folk medical knowledge: A comparison between curers and noncurers. *American Anthropologist* 88: 351–70.

Garro, L. C. 1988. Explaining high blood pressure: Variation in knowledge about illness. *American Ethnologist* 15: 98–119.

Green, P. E., and F. J. Carmone. 1970. *Multidimensional scaling and related techniques in marketing analysis*. Boston: Allyn and Bacon.

Guest, G. 2000. Using Guttman scaling to rank wealth: Integrating quantitative and qualitative data. *Field Methods* 12: 346–57.

Handwerker, W. P. 1996. Constructing Likert scales: Testing the validity and reliability of single measures of multidimensional variables. *Cultural Anthropology Methods* 8(1): 1–6.

Harman, R. C. 2001. Activities of contemporary Mayan elders. *Journal of Cross-Cultural Gerontology* 16: 57–77.

Haug, M. R. 1977. Measurement in social stratification. *Annual Review in Sociology* 3: 51–77.

Henley, N. M. 1969. A psychological study of the semantics of animal terms. *Journal of Verbal Learning and Verbal Behavior* 8: 176–84.

Holmes, T. H., and R. H. Rahe. 1967. The social readjustment rating scale. *Journal of Psychosomatic Research* 11: 213–18.

Hsiao, A. F., G. W. Ryan, R. D. Hays et al. 2006. Variations in provider conceptions of integrative medicine. *Social Science and Medicine* 62: 2973–87.

Hutchinson, J. W. 1983. Expertise and the structure of free recall. In *Advances in consumer research*, Vol. 10, ed. R. P. Bagozzi and A. M. Tybout, 585–89. Ann Arbor, MI: Association for Consumer Research.

Hutchinson, J. W., and G. R. Lockhead. 1977. Similarity as distance: A structural principle for semantic memory. *Aluman Learning and Memory* 6: 660–78.

Ice, G. H., and J. Yogo. 2005. Measuring stress among Luo elders. *Field Methods* 17: 394–411.

Jaskyte, K., and W. W. Dressler. 2004. Studying culture as an integral aggregate variable: Organizational culture and innovation in a group of nonprofit organizations. *Field Methods* 16: 265–84.

Jaskyte, K., and W. W. Dressler. 2005. Organizational culture and innovation on non-profit human service organizations. *Administration in Social Work* 29: 23–41.

Johnson, A. 1995. A Guttman scale analysis of Matsigenka men's manufacturing skill. *Field Methods* 7: 1–12.

Johnson, J. C. 1990. *Selecting ethnographic informants*. Qualitative Research Methods Series, Vol. 22. Thousand Oaks, CA: Sage.

Johnson, J. C., J. S. Boster, and L. Palinkas. 2003. Social roles and the evolution of networks in isolated and extreme environments. *Journal of Mathematical Sociology* 27: 89–122.

Johnson, J. C., and D. C. Griffith. 1996. Pollution, food safety, and the distribution of knowledge. *Human Ecology* 24: 87–108.

Johnson, J. C., and M. L. Miller. 1983. Deviant social positions in small groups: The relations between role and individual. *Social Networks* 5: 51–69.

Johnson, J. C., and M. K. Orbach. 2002. Perceiving the political landscape: Ego biases in cognitive political networks. *Social Networks* 24: 291–310.

Johnston, F. E., S. M. Low, Y. deBessa, and R. B. MacVean. 1987. Interaction of nutrition and socioeconomic status as determinants of cognitive development in disadvantaged urban Guatemalan children. *American Journal of Physical Anthropology* 73: 501–6.

Jowell, R., C. Roberts, R. Fitzgerald, and G. Eva. 2007. *Measuring attitudes cross-nationally. Lessons from the European Social Survey*. London: Sage.

Kaldjian, L. C., E. W. Jones, G. E. Rosenthal et al. 2006. An empirically derived taxonomy of factors affecting physicians' willingness to disclose medical errors. *Journal of General Internal Medicine* 21: 942–48.

Karabatsos, G., and W. H. Batchelder. 2003. Markov chain estimation for test theory without an answer key. *Psychometrika* 68: 373–89.

Karlawish, J., F. K. Barg, D. Augsburger et al. 2011. What Latino Puerto Ricans and non-Latinos say when they talk about Alzheimer's disease. *Alzheimer's & Dementia* 7: 161–70.

Kasulis, J. J., R. F. Lusch, and E. F. Stafford, Jr. 1979. Consumer acquisition patterns for durable goods. *Journal of Consumer Research* 6: 47–57.

Kay, P. 1964. A Guttman scaling model of Tahitian consumer behavior. *Southwestern Journal of Anthropology* 20: 160–67.

Kempton, W., J. S. Boster, and J. A. Hartley. 1995. Environmental values in American culture. Cambridge, MA: M.I.T. Press.

Kirk, J., and M. L. Miller. 1978. Cognitions of coca in Columbia, Ecuador, and Peru. In *A multicultural view of drug abuse*, ed. D. E. Smith, S. M. Anderson, M. Buxton, N. Gottlieb, W. Harvey, and T. Chung, 132–46. Cambridge, MA: Schenkman Publishing.

Kirk, L., and M. Burton. 1977. Meaning and context: A study in contextual shifts in meaning of Maasai personality descriptors. *American Ethnologist* 4: 734–61.

Kiš, A. D. 2007. An analyisis of the impact of AIDS on funeral culture in Malawi. *NAPA Bulletin* 27: 129–40.

Koster, J. M., J. J. Hodgen, M. D. Venegas, and T. J. Copeland. 2010. Is meat flavor a factor in hunters' prey choice decisions? *Human Nature—An Interdisciplinary Perspective* 21: 219–42.

Koster, J. M., and K.B. Tankersley. 2012. Heterogeneity of hunting ability and nutritional status among domestic dogs in lowland Nicaragua. *PNAS* 109: E463–E470.

Krackhardt, D. 1987. Cognitive social structures. *Social Networks* 9: 109–34.

Krackhardt, D. 1990. Assessing the political landscape: Structure, cognition, and power in organizations. *Administrative Science Quarterly* 35: 342–69.

Kruskal, J. B., and M. Wish. 1990. *Multidimensional scaling*. Quantitative Applications in the Social Sciences Series, Vol 11. Thousand Oaks, CA: Sage.

Lewis, C. E., J. M. Siegel, and M. A. Lewis. 1984. Feeling bad: Exploring sources of distress among pre-adolescent children. *American Journal of Public Health* 74: 117–22.

Lieberman, D., and W. M. Dressler. 1977. Bilingualism and cognition of St. Lucian disease terms. *Medical Anthropology* 1: 81–110.

Loftus, E., and W. Marburger. 1983. Since the eruption of Mt. St. Helens did anyone beat you up? Improving the accuracy of retrospective reports with landmark events. *Memory and Cognition* 11: 114–20.

Lopez, A., S. Atran, J. D. Coley et al. 1997. The tree of life: Universal and cultural features of folk biological taxonomies and inductions. *Cognitive Psychology* 32: 251–95.

Lutz, C. 1982. The domain of emotion words on Ifaluk. *American Ethnologist* 9: 113–28.

Lynch, E. B., and S. Holmes. 2011. Food group categories of low-income African-American women. *Journal of Nutrition Education and Behavior* 43: 157–64.

Macauda, M. M., P. I. Erickson, M. C. Singer, and C. C. Santelices. 2011. A cultural model of infidelity among African American and Puerto Rican young adults. *Anthropology & Medicine* 18: 351–64.

Magaña, J. R., M. Burton, and J. Ferreira-Pinto. 1995. Occupational names in three nations. *Journal of Quantitative Anthropology* 5: 1149–68.

Magaña, J. R., G. W. Evans, and A. K. Romney. 1981. Scaling techniques in the analysis of environmental cognition data. *Professional Geographer* 33: 294–310.

Mathevet, R. M. Etienne, T. Lynam, and C. Calvet. 2011. Water management in the Carmague Biosphere Reserve: Insights from comparative mental models analysis. *Ecology and Society* 16(1): 43. www.ecologyandsociety.org/vol16/iss1/art43/ES-2011-4007.pdf (accessed March 9, 2013).

Mathez-Stiefel, S., and I. Vandebroek. 2012. Distribution and transmission of medicinal plant knowledge in the Andean highlands: A case study from Peru and Bolivia. *Evidence-Based Complementary and Alternative Medicine*, P1-18. http://www.hindawi.com/journals/ecam/2012/959285/ (accessed March 9, 2013).

Maupin, J. N., N. Ross, and C. A. Timura. 2011. Gendered experiences of migration and conceptual knowledge of illness. *Journal of Immigrant and Minority Health* 13: 600–608.

McCarty, C. 2001. Structure in personal networks. *Journal of Social Structure* 3(1).

http://www.cmu.edu/joss/content/articles/volume3/McCarty.html (accessed March 9, 2013)

Meztger, D., and G. Williams. 1966. Some procedures and results in the study of Native categories: Tzeltal firewood. *American Anthropologist* 68: 389–407.

Mezzich, J. E., and H. Solomon. 1980. *Taxonomy and behavioral science*. London: Academic Press.

Miller, E. M. 2011. Maternal health and knowledge and infant health outcomes in the Ariaal people of northern Kenya. *Social Science & Medicine* 73: 1266–74.

Miller, G. A. 1969. A psychological method to investigate verbal concepts. *Journal of Mathematical Psychology* 6: 169–91.

Miller, M. L., and E. Hutchins. 1989. On the acquisition of boardsailing skill. In *The content of culture: Constants and variants, studies in honor of John M. Roberts*, ed. R. Bolton, 153–70. New Haven, CT: HRAF Press.

Miller, M. L., and J. C. Johnson. 1981. Hard work and competition in an Alaskan fishery. *Human Organization* 40: 131–39.

Milligan, G. W. 1980. An examination of the effect of six types of error perturbation of fifteen clustering algorithms. *Psychometrika* 45: 325–42.

Miranda, T. M., M. C. de Mello Amorozo, J. S. Govone, and D. M. Miranda. 2007. The influence of visual stimuli in ethnobotanical data collection using the listing task method. *Field Methods* 19: 76–86.

Moore, R., M. L. Miller, P. Weinstein et al. 1986. Cultural perceptions of pain and pain coping among patients and dentists. *Community Dental Oral Epidemiology* 14: 327–33.

Morgan, D. L. 1996. Focus groups. *Annual Review of Sociology* 22: 129–52.

Nolan, J. M. 2002. Wild plant classification in Little Dixie: Variation in a regional culture. *Journal of Ecological Anthropology* 6: 69–81.

Nunnally, J. C. 1978. *Psychometric theory*. New York: McGraw-Hill.

Nyamongo, I. K. 2002. Assessing intracultural variability statistically using data on malaria perceptions in Gusii, Kenya. Field Methods 14: 148–60.

Oravecz, Z., J. Vandekerckhove, and W. H. Batchelder. 2014. Bayesian cultural consensus theory. *Field Methods*.

Pachter, L. M., S. C. Weller, R. D. Baer et al. 2002. Asthma beliefs and practices in mainland Puerto Ricans, Mexican-Americans, Mexicans, and Guatemalans: Consistency and variability in health beliefs and practices. *Journal of Asthma* 39: 119–34.

Parr, M. G. 1996. The relationship between leisure theory and recreation practice. *Leisure Sciences* 18: 315–32.

Penka, S., H. Heiman, A. Heinz, and M. Schouler-Ocak. 2008. Explanatory models of addictive behavior among Native German, Russian-German, and Turkish youth. *European Psychiatry* 23: S36–S42.

Quinn, N. 1987. Convergent evidence for a cultural model of American marriage. In *Cultural models in language and thought*, ed. D. Holland and N. Quinn, 173–92. Cambridge: Cambridge University Press.

Reyes-Garcia, V., E. Byron, V. Vadez et al. 2004. Measuring culture as shared knowledge: Do data collection formats matter? Cultural knowledge of plant uses among Tsimane' Amerindians, Bolivia. *Field Methods* 16: 135–56.

Reyes-Garcia, V., C. C. Gravelee, T. W. McDade et al. 2010. Cultural consonance and body morphology: Estimates with longitudinal data from an Amazonian society. *American Journal of Physical Anthropology* 143: 167–74.

Roberts, J. M., and G. E. Chick. 1979. Butler County eight ball: A behavioral space analysis. In *Sports, games, and play: Social and psychological viewpoints*, ed. J. H. Goldstein, 65–99. Hillsdale, NJ: Lawrence Erlbaum.

Roberts, J. M., G. E. Chick, M. Stephanson, and L. L. Hyde. 1981. Inferred categories for tennis play: A limited semantic analysis. In *Play as context*, ed. A. B. Cheska, 181–95. West Point, NY: Leisure Press.

Roberts, J. M., T. V. Golder, and G. E. Chick. 1980. Judgment, oversight and skill: A cultural analysis of P-3 pilot error. *Human Organization* 39: 5–21.

Roberts, J. M., and S. H. Nattrass. 1980. Women and trapshooting: Competence and expression in a game of skill with chance. In *Play and culture*, ed. H. B. Schwartzman, 262–91. West Point, NY: Leisure Press.

Rocha, J. 2005. Measuring traditional agro-ecological knowledge: An example from peasants in the Peruvian Andes. *Field Methods* 17: 356–72.

Romney, A. K., W. H. Batchelder, and S. C. Weller. 1987. Recent applications of consensus theory. *American Behavioral Scientist* 31: 163–77.

Romney, A. K., D. D. Brewer, and W. H. Batchelder. 1993. Predicting clustering from semantic structure. *Psychological Science* 4: 28–34.

Romney, A. K., and R. G. D'Andrade. 1964. Cognitive aspects of English kin terms. *American Anthropologist* 66: 146–70.

Romney, A. K., M. Keiffer, and R. E. Klein. 1979. A normalization procedure for correcting biased response data. *Social Science Research* 2: 307–20.

Romney, A. K., C. C. Moore, and C. D. Rusch. 1997. Cultural universals: Measuring the semantic structure of emotion terms in English and Japanese. *PNAS* 94: 5489–94.

Romney, A. K., T. Smith, H. E. Freeman et al. 1979. Concepts of success and failure. *Social Science Research* 8: 302–26.

Romney, A. K., and S. C. Weller. 1984. Predicting informant accuracy from patterns of recall among individuals. *Social Networks* 4: 59–77.

Romney, A. K., S. C. Weller, and W. H. Batchelder. 1986. Culture as consensus: A theory of cultural and informant accuracy. *American Anthropologist* 88: 313–38.

Ross, J. L., S. L. Laston, P. J. Pelto, and L. Muna. 2002. Exploring explanatory models of women's reproductive health in rural Bangladesh. *Culture, Health, & Sexuality* 4: 173–90.

Ross, N., T. Barrientos, and A. Esquit-Choy. 2005. Triad tasks, a multipurpose tool to elicit similarity judgments: The case of Tzotzil Maya plant taxonomy. *Field Methods* 17: 269–82.

Ross, N., and D. L. Medin. 2005. Ethnography and experiments: Cultural models and expertise effects elicited with experimental research techniques. *Field Methods* 17: 131–49.

Ross, N., J. Maupin, and C.A. Timura. 2011. Knowledge organization, categories, and ad hoc groups: Folk medical models among Mexican migrant in Nashville. *Ethos* 39: 165–88.

Ross, N., C. Timura, and J. Maupin. 2012. The case of curers, noncureres, and biomedical experts in Pichataro, Mexico resiliency in folk-medical beliefs. *Medical Anthropology Quarterly* 26: 159–81.

Ruebush, T. K., II, S. C. Weller, and R. E. Klein. 1992. Knowledge and beliefs about malaria on the Pacific coastal plain of Guatemala. *American Journal of Tropical Medicine and Hygiene* 46: 451–59.

Rumelhart, D. E., and A. A. Abrahamson. 1973. A model for analogical reasoning. *Cognitive Psychology* 5: 1–28.

Ryan, G. W., and H. R. Bernard. 2003. Techniques to identify themes. *Field Methods* 15: 85–109.

Ryan, G. W., J. M. Nolan, and P. S. Yoder. 2000. Successive free listing: Using multiple free lists to generate explanatory models. *Field Methods* 12: 83–107.

Sayles, J. N., G. W. Ryan, J. S. Silver et al. 2007. Experiences of social stigma and implications for health care among a diverse population of HIV positive adults. *Journal of Urban Health* 84: 814–28.

Schrauf, R. W., and E. Navarro. 2005. Using existing tests and scales in the field. *Field Methods* 17: 373–93.

Schuman, H., and S. Presser. 1996. *Questions & answers in attitude surveys. Experiments on question form, wording, and context*. Thousand Oaks, CA: Sage.

Schunko, C., and C. R. Vogl. 2010. Organic farmers use of wild food plants and fungi in a hilly area in Styria (Austria). *Journal of Ethnobiology and Ethnomedicine* 6: 17.

Scott, J. 2000. *Social network analysis: A handbook*, 2nd ed. Newbury Park, CA: Sage.

Smith, C. A. S. 2012. Living with sugar: Influence of cultural beliefs on type 2 diabetes self-management of English-speaking Afro-Caribbean women. *Journal of Immigrant Minority Health* 14: 640–47.

Smith, C. S., M. Morris, W. Hill et al. 2004. Cultural consensus analysis as a tool for clinic improvements. *Journal of General Internal Medicine* 19: 514–18.

Smith, C. S., M. Morris, F. Langois-Winkle et al. 2010. A pilot study using cultural consensus analysis to measure systems-based practice performance. *International Journal of Medical Education* 1: 15–18.

Sokal, R., and P. Sneath. 1963. *Principles of numerical taxonomy*. San Francisco: W. H. Freeman.

Soutar, G.N., and S. P. Cornish-Ward. 1997. Ownership patterns for durable goods and financial assets: A Rasch analysis. *Applied Economics* 29: 903–11.

Spradley, J. P. 1979. *The ethnographic interview*. New York: Holt, Rinehart and Winston.

Stanton, B. F., R. Aronson, S. Borgatti, and J. Galbraith. 1993. Urban adolescent high-risk sexual behavior: Corroboration of focus group discussions through pile-sorting. *AIDS Education and Prevention* 5: 162–74.

Stefflre, V. J. 1972. Some applications of multidimensional scaling to social science problems. In *Multidimensional scaling: Theory and applications in the behavioral sciences*, Vol. 2, ed. A. K. Romney, R. Shepard, and S. B. Nerlove, 211–43. New York: Academic Press.

Teddlie, C., and A. Tashakkori. 2009. *Foundations of mixed methods research: Integrating quantitative and qualitative approaches in the social and behavioral sciences*. Thousand Oaks, CA: Sage.

Thompson, E. C., and Z. Juan. 2006. Comparative cultural salience: Measures using free-list data. *Field Methods* 18: 398–412.

Trosset, C, and D. D. Caulkins. 2001. Triangulation and confirmation in the study of Welsh concepts of personhood. *Journal of Anthropological Research* 57: 61–81.

Trotter, R. T., II, S. C. Weller, R. D. Baer et al. 1999. Consensus theory model of AIDS/SIDA beliefs in four Latino populations. *AIDS Education and Prevention* 11: 414–26.

Truex, G. F. 1977. Measurement of intersubject variations in categorizations. *Journal of Cross-Cultural Psychology* 8: 71–82.

Vogl, C. R., B. Vogl-Lukasser, and R. K. Puri. 2004. Tools and methods for data collection in ethnobotanical studies of homegardens. *Field Methods* 16: 285–306.

Wasserman, S., and K. Faust. 2009. *Social network analysis*. Cambridge: Cambridge University Press.

Webster, C. M., A. L. Iannucci, and A. K. Romney. 2002. Consensus analysis for the measurement and validation of personality traits. *Field Methods* 14: 46–64.

Weller, S. C. 1983. New data on intra-cultural variation: The hot-cold concept. *Human Organization* 42: 249–57.

Weller, S. C. 1984. Cross-cultural concepts of illness: Variation and validation. *American Anthropologist* 86: 341–51.

Weller, S. C. 1987. Shared knowledge, intercultural variation and knowledge aggregation. *American Behavioral Scientist* 31: 178–93.

Weller, S. C. 2007. Cultural consensus theory: Applications and frequently asked questions. *Field Methods* 19: 339–68.

Weller, S. C., R. D. Baer, J. E. Garcia de Alba et al. 2002. Regional variation in Latino beliefs about *susto*. *Culture, Medicine, and Psychiatry* 26: 449–72.

Weller, S. C., R. D. Baer, J. E. Garcia de Alba, and A. L. Salcedo Rocha. 2012. Mexican and Mexican-American explanatory models of diabetes: A comparison of community members, patients, and physicians. *Social Science & Medicine* 75: 1088–96.

Weller, S. C., and C. I. Dungy. 1986. Personal preferences and ethnic variations among Anglo and Hispanic breast and bottle feeders. *Social Science and Medicine* 23: 539–48.

Weller, S. C., L. M. Pachter, R. T. Trotter, II, and R. D. Baer. 1993. Empacho in four Latino groups: A study of intra- and inter-cultural variation in beliefs. *Medical Anthropology* 15: 109–36.

Weller, S. C., and A. K. Romney. 1988. *Systematic data collection*. Qualitative Research Methods Series, Vol. 10. Thousand Oaks, CA: Sage.

Weller, S. C. and A. K. Romney. 1990. *Metric scaling: Correspondence analysis*. Quantitative Applications in the Social Sciences Series, Vol. 75. Thousand Oaks, CA: Sage.

Weller, S. C., A. K. Romney, and D. P. Orr. 1987. The myth of a sub-culture of corporal punishment. *Human Organization* 46: 39–47.

Weller, S. C., T. R. Ruebush, II, and R. E. Klein. 1997. Predicting treatment-seeking behavior in Guatemala: A comparison of the health services research and decision-theoretic approaches. *Medical Anthropology Quarterly* 11: 224–45.

Young, F. W., and R. C. Young. 1962. Key Informant reliability in rural Mexican villages. *Human Organization* 20: 141–48.

Young, J. C. 1978. Illness categories and action strategies in a Tarascan town. *American Ethnologist* 5: 81–97.

Young, J. C. 1980. A model of illness treatment decisions in a Tarascan town. *American Ethnologist* 7: 106–31.

Young, J. C., and L. Y. Garro. 1982. Variation in the choice of treatment in two Mexican communities. *Social Science and Medicine* 16: 1453–65.