Steps in Constructing a Written-Response Survey for Applied Linguistics[1]

H. P. L. Molloy, Asia University

A note on terminology

In this and the accompanying piece, "written-response survey" (hereafter "survey") means a written document given to participants and containing questions, statements, descriptions of situations, or other stimuli (hereafter "prompts") and to which participants respond by directly recording responses themselves. "Survey" is sometimes used distinctly from "questionnaire" (see, *e.g.,* Cohen, Manion, & Morrison, 2000), "questionnaire" often referring to a series of prompts delivered orally by another person and for which responses are recorded by the other person. The use of questionnaires in this sense involves complexities that will not be addressed here. In applied linguistics, "questionnaire" and "survey" are often used interchangeably, however.

INTRODUCTION

The written response survey is a popular information gathering technique in applied linguistics and in any other field concerned with measuring a few things common to large groups of people.

The results of surveys are familiar to virtually anyone exposed to news media in post-industrial societies: as I write, news media daily report the results of surveys and questionnaires on the U.S. presidential campaign; the convenience shop around the corner carries a couple of popular magazines devoted entirely to reporting results of popularity surveys.

Despite the familiarity of surveys, however, constructing and interpreting them for applied linguistics research is a time-consuming process, one with potential pitfalls at nearly every step. This article and the companion piece on analysis of survey results will, I hope, serve as a bit of familiarization with planning and constructing a useful survey.

My research background concerns mainly constructing language-elicitation and reaction surveys in L2 pragmalinguistics and sociopragmatics; my foci here are influenced by that work. For those who wish to pursue survey construction seriously, it will be necessary to consult longer works. In the applied linguistics area, Brown (1997, or, especially, 2001) has written well and clearly on survey construction and gives an especially good introduction to analytic procedures. The newer volume by Dörnyei (2003) is less detailed, but has a wider range of examples from the literature. Books from the consumer research area (*e.g.,* Backstrom & Hursh-César, 1981) and, among other fields, psychology, sociology, or political science treat of survey-construction problems more like those of applied linguists than do those written for other frequent survey users, such as geographers or epidemiologists. For analyses, I find many useful ideas in texts or articles in descriptive studies in population studies or parasitology. I have made a particular effort to include a wide range of referenced examples to allow the reader varied routes in the literature, but have limited them to resources that can be accessed with relative ease in Japan or by contacting the authors directly (as I have often done).

[1] Miyake Hiroko, Bruce E. Carrick, and Chieko Mimura gave valuable comments on a near-complete version of the manuscript; Mika Shimura advice on the general direction of the piece. My debt to these colleagues is greater than my ability to express my thanks. Opinions and mistakes in this paper are entirely mine.

Surveys are an old and popular research technique. In this short introduction, I concentrated on recent, easily available studies in L2 studies for illustrations. Most of the illustrations are of potential pitfalls in survey construction, but my choices should not be taken to imply that the studies I discuss are unusually bad. Because surveys are such crude yet complex research tools, virtually any extant survey can be criticized on one or (usually) more points.

Because surveys are so familiar, they may seem a relatively easy research technique; it is my impression that a survey is often a first choice for beginning researchers. In my own experience, surveys are not extraordinarily difficult to construct or analyze, but they do necessitate care. Certain steps I discuss below have to be taken, and each step takes time.

How much time do the steps take? Given the time constraints many applied linguistics researchers work under (classes, meetings, private lives, and so on), I recommend planning on at least a year of work in getting a survey ready from scratch. I am presently working on several surveys: the oldest one not yet finished I started working on three and a half years ago.

IS A SURVEY THE BEST CHOICE?

A survey is a useful research instrument if your goal is to characterize a small number of behaviors or a much smaller number of attitudes, beliefs, or similar constructs of a large group of people. Data selection (that is, including some data or types of data and excluding others) is inevitable in any kind of research, but it tends to be particularly great in survey research. Likewise, measurement error exists in any research, but it tends to be particularly important in survey research.

If the researcher is interested in painting a complete portrait of the important factors in a group's language learning, for example, a survey is probably not a good choice, as to reliably measure everything necessary would take an extraordinarily long survey.

If the researcher is interested in collecting data from a relatively small number of people, it is probably better to use open ended interviews or direct observation of participants than to use a survey. This is for two reasons: first, conducting interviews with, say, each member of a language class of 40 students is probably less time-consuming, all told, than constructing a survey. Second, most of the popular statistical techniques for analyzing the results of surveys are only usefully used with relatively large groups of people: with highly reliable and distributionally normal sets of numbers, the popular *t*-test is hypothetically useful with only 60 participants (30 in each of two groups), but the seemingly unavoidable error that accompanies most real-world surveys makes me hesitate to use surveys with fewer than 100 people in each group (as detailed in Molloy & Shimura, 2003, where we demonstrated that survey results can differ radically when the number of participants is smaller, even if the participants are simply a randomly selected subset of a larger set).

How many (and which) participants?

Generally, if you can get 400 randomly selected participants from the group you are interested in characterizing, you should be able to get a reasonably accurate characterization of the group of interest, provided you have a good survey. For example, if you are interested in characterizing Japanese university students, 400 randomly selected students would be an adequate sample size. Generally, the larger the number of participants, the more accurate the measurement. Also, the smaller the number of people in the target group, the larger the percentage of that group you will have to survey.

The excellent textbook by Cohen, Manion, and Morrison (2000) has useful tables for determining adequate sample sizes (pp. 94-95). The tables show that, for example, to make a reasonable portrait of a year of students involved in the Asia University Freshman English program (about 1500 per year), it would be necessary to give a survey to 306 randomly

selected students. With smaller groups, much larger percentages would be required: for example, in the first of two surveys reported by Chiba and Matsuura (2004) the researchers distributed their survey to all of the teachers in what seems to be the Asia University Freshman English program. If all the 23 participants (or all but one) had responded, it would have been possible to create an acceptably accurate portrait of the teachers in the program.

"Random selection" here means that every person in the population of interest has an equal chance of being selected to participate in the survey. If the population of interest in your survey is Japanese university students of English, then every student in Japan enrolled in an English class must have an equal chance of being selected. I know of no cases in which such random selection has been done (or even attempted) in applied linguistics.

Instead, a much more common method of sampling, Lunneborg (2001) points out—I would contend this is the only method of sampling in applied linguistics studies—is randomization of available cases or, applied to survey research, random selection from available cases. Generally, even if researchers use random choice procedures to select participants, they are limited to those participants who are accessible, such as students enrolled in intact classes, students from one university, interested colleagues, or volunteers.

When true random selection is not possible, the only practical recommendation to make regarding the number of participants to have in a survey research project is: as many as possible. At times patience is necessary: for a recently completed study employing survey methods I was involved in (Molloy & Shimura, 2004), we needed three years to collect our data. We were collecting data from a convenient group of available cases, our students, and had to collect data over three academic years before we had enough participants to be sure our statistics would be representative.

An exception to the recommendation of interviews or observation over surveys might be a case like that in Connolly (2004). All members of Connolly's target population responded, and he presented the results from his short open-ended response survey almost in their entirety. There are times when, because of scheduling problems or physical separation, observations or interviews with open-ended questions become difficult. Provided one can get a high enough response rate, surveys can be useful even for small groups.

Even when all indications are that a survey can be done well with an adequate number of respondents, using *only* a survey is not the best choice. Any research method distorts the object of interest to some extent. A multimethod approach to any research question is always to be preferred: if you are conducting a survey, combining the findings with findings from interviews, observations, tests, participant diaries, or other research approaches will allow compensation for the distortions of any single research approach and make your final report more trustworthy and a more useful contribution to the literature.

GENERATING PROMPTS

Surveys comprise prompts and response collectors: "prompts" refers to what participants read or hear, such as questions or statements; "response collectors" refers to how the participants register their reactions to the prompts, such as by writing a short answer, choosing one of several options, or marking a scale.

Prompts are stimuli related to the behavior or construct the researcher is interested in. Again, if the research question concerns behavior, the relationship between the prompt and the behavior in question is relatively straightforward: if you ask me how many books I read each month, there is a reasonable chance that my answer would accord reasonably with the same point of interest investigated in a different way, such as direct observation.

If the research question concerns some sort of mental construct, however, an extra step is added to the reasoning process. If the researcher is interested in students' motivations, for

example, he or she uses prompts that, if agreed to, are held to indicate the participant has at least some of a certain kind of motivation. For example, in the well-established and –regarded Motivation/Attitude Test Battery (see, for example, Gardner & Lambert, 1972; Gardner, 1985; or Gardner, Masgoret, & Tremblay, 1999), there are prompts such as "Studying [French] is important because it enables people to better understand [French Canadian] life and culture." (Words in brackets are changed appropriately in different versions of the battery.) Participants rate this statement on a scale from –3 (strong disagreement) to 3 (strong agreement). Participants who rate this and several other statements in the battery highly are said to have or evince "integrative motivation," or a desire to learn language to fit into (in this case) French-Canadian society.

Note that asking a participant to agree or disagree with several statements is an indirect way of measuring this construct and that the Motivation/Attitude Test Battery does not include questions like "Do you want to fit into [French-Canadian] society?" This question might be considered a more direct measurement (akin to "How many books did you read last month?") of integrative motivation than asking several indirect questions, but, for reasons explained below, this "indirect" way of measuring constructs is more common in applied linguistics research. It is the approach taken in such widely used surveys as those on willingness to communicate (see the work of MacIntyre and colleagues, 1998, 2001), learning strategies (*e.g.,* Brown, Robson, & Rosenkjar, 2001), and language-learning anxiety (*e.g.,* Cheng, Horwitz, & Schallert, 1999; or, especially, Horwitz, 2001).

This indirect measurement is based upon the same reasoning as is behind language tests. In a language test, the ability to correctly answer a high proportion of particular questions (or perform well a high proportion of language tasks) is held to indicate the presence of a high degree of proficiency with the particular language. It is impossible to completely assess the construct of language proficiency: to do so would necessitate checking a student's performance in every possible language use situation. Tests work instead by checking a reasonable, representative subset of those situations. The consumers of the test (teachers, administrators, personnel departments, and so on) then are left to decide whether the test results do indeed adequately predict the language proficiency. It is likewise with surveys.

Deciding how many prompts would be best in a survey and what kind of prompts to use takes some planning and, if you decide to make your own prompts, considerable thought and preliminary work.

HOW MANY PROMPTS?

Before being given to participants, prompts should be put through many of the checks and tests described below; those checks often show that prompts are unsuitable. For example, prompts may prove to be ambiguous; they may not solicit different reactions from different people; if translated, they may not translate well. It is recommended that researchers begin with a preliminary prompt pool three to four times the size of the final expected total. Working on a survey to collect 12 examples of English-language complaint initiations, we (Molloy & Shimura, 2002) began with a prompt pool of 72; for a study on requesting for which we needed 6 prompts, we began with a pool of 126 (Shimura & Molloy, 2003).

Very generally, a survey that will be given to participants for testing or as a final instrument needs more prompts if (*a*) one is attempting to study a mental construct, rather than a behavior, (*b*) the construct may have manifestations in many behaviors, but not necessarily all areas for all participants, and (*c*) the researcher wants to assure relatively high reliability.

The difference in number because of the object of research (*a* above) is because of the relative directness of the measurement the survey prompts are doing. If a researcher is

interested in asking about particular behaviors (*e.g.*, how many books in an L2 a respondent reads), one question (*e.g.,* "How many books do you usually read in a month?" _____) may suffice. However, if the researcher is interested in something that cannot be directly measured (*i.e.*, in what is usually referred to as a "mental construct"), it is usually necessary to ask several questions about the same topic.

Asking multiple questions about mental constructs is common in part because people vary so much (*b* above). A personality test such as the Minnesota Multiphasic Personality Index (MMPI) or the Myers-Briggs Type Indicator (MBTI) (or any of the others in the list of those currently popular available in Viswesvaran & Ones, 2000) have several prompts contributing to each overall dimension of personality because no one behavior, reaction, or opinion can be considered a perfect indicator of a participant's placement on a particular continuum.

Both the MMPI and the MBTI have so many prompts the surveys can take an hour or more to complete. When I took the MBTI in 2001, I noted that I became bored with the test after about 10 questions. Longer surveys are only relatively more accurate assuming perfect patience and endurance on the part of the participants, something that is probably never the case. Having more prompts is usually better, but having the most questions is not necessarily the best. Nevertheless, it is almost always a conceptual mistake to have only one prompt for a given mental construct: if you are interested in a construct, ask about it more than once.

Regarding reliability (*c* above), generally the longer a survey is, the more reliable it will be. Reliability means the tendency of the survey to measure the same way every time. If your prompts are not good, your survey will probably not be reliable, but if they are, the longer the survey, the more reliable it will be.

A final point is that it may be helpful to have more prompts than you are interested in in your final survey. Such "extra" prompts, not directly related to your object of interest, are called "distractors" and can be especially useful when you think that knowing the object of study might affect the participants' answers. A good overview is available in Miller and Cann (1998). Kuha (1997) compared two surveys, one of which included a notation about the object of study and one of which did not, and presented evidence to show that the two types of surveys may elicit different results. Other cases in which distractors were used well can be found in Beebe and Takahashi (1989) and Olshtain and Weinbach (1993).

GETTING PROMPTS FROM THE LITERATURE

This is one of the easiest and most helpful ways of finding prompts. It is the strategy taken by, for example, by Ford (1984) and (at least in part) by Matsuura, Chiba, and Hildebrandt (2001) and Chiba and Matsuura (2004). In research I am involved in, the published literature is always the first place searched for prompts (see, *e.g.,* Shimura & Molloy, 2003; 2004; Molloy & Shimura, 2004).

Taking prompts from the literature is easy because it makes inventing prompts from scratch unnecessary. (The prompts still have to be checked for reliability and validity, however.) It is helpful to the research community in general, as well, because it allows comparison of results across populations. Replication is common in many fields (particularly in the physical sciences), but not at all common in applied linguistics.

In some cases, the best-known survey forms in applied linguistics must be purchased (such as some surveys on learning strategies), and most of the best-known psychological and language aptitude tests as well can be used only for a fee (and in some cases only after taking a training course). However, many surveys are published in full in journal articles or book chapters; in nearly every case, at least a few prompts are reproduced. Writing directly to the authors is recommended should you wish to use all or parts of a published survey.

Getting prompts from the target audience

Prompts that come from the eventual consumers of the research (such as school administrators or the whole of a department faculty) are often used in surveys conducted for purposes other than pure research. Examples familiar to most include political polls conducted by parties or advocacy groups and consumer research. In school settings, examples include faculty evaluation surveys given to students and surveys used in needs analyses. Brown (2001) describes in interesting detail a survey he worked on at the behest of TESOL in which the prompts were designed to elicit information of interest to the group that commissioned the survey.

Interviewing or having discussions with colleagues, those likely to be reading your research, is immensely helpful as well. Connolly (2004) describes well how this might be done informally, but with sufficient rigor, with interviewing. Using expert resources is also an integral part of developing Thurstone scales, as described below.

GETTING PROMPTS FROM THE TARGET POPULATION

Kachru (1994) implied that the question of whether prompts described situations familiar to participants is particularly important in L2 or cross-cultural situations. Most applied linguistics studies involving surveys do not mention whether the prompts were familiar to participants or not, although some studies in the communication studies area (*e.g.,* Cai & Wilson, 2000; Wilson & Kunkel, 2000) do. One of the few general applied linguistics studies test the question of whether prompts were familiar to participants was one done by Kreuz and Roberts (1993) and one I was involved in was designed to directly assess the issue (Shimura & Molloy, 2003) of whether it was useful to try to guarantee familiarity by soliciting prompts from the target population, finding no decisive results.

Getting prompts from members of the population of interest, however, is a strategy that is likely to benefit the researcher by making the research more valid. The area of cross-cultural psychology gives hints about the problems of generating prompts without taking into account the world-views of the target population. Brislin and his co-workers (1973) give a fine overview of the bulk of the findings on cross-cultural prompt effects, and Church (2000) can provide the beginnings of an update in the field.

Interviews with target population members or observation of similar people can help researchers determine what is likely to be salient to participants. Summaries of several studies in which prompts were gathered with such procedures are given in the Cohen, Manion, and Morrison text (2000). Shibuya and Mimura (2002) found it useful to observe the behavior of their students to get ideas for prompts in a survey they did on the interaction of language learning, anxiety, and gender.

In cases in which the survey is designed to elicit language (such as the production questionnaires, or discourse completion tests, used in pragmalinguistics or sociopragmatics research), it is possible to simply have people like the eventual participants write prompts. In studies of Japanese and American English users initiation of complaints and Japanese requests, Mika Shimura and I (Molloy & Shimura, 2002, 2003; Shimura & Molloy, 2003, 2004) created a majority of the prompts we used from descriptions of language use situations solicited from our students.

GENERATING PROMPTS FROM THEORY

With regard to development of the field, this is perhaps the best strategy because testing published theories can allow those theories to be extended or modified. In this case, you will develop prompts that will allow testing of the implications of the theory. An example of such

an approach can be found in Hullett and Tamborini (2001), who used a survey as a way of refining an expectancy theory of communicative behavior. In my subfield of L2 pragmatics, the politeness theory of Brown and Levinson (1987) is influential and productive. In this theory, language behavior is held to vary according to three factors: the social distance between interlocutors, the relative social power of the interlocutors, and the degree of imposition on the hearer. In most assessments of situational language behavior, one or more of these factors is systematically varied in prompts. Examples can be found in virtually all L2 speech act studies published in the last 20 years (*e.g.,* Blum-Kulka & Olshtain, 1984, and many of the papers cited in the reviews by DuFon *et al.*, 1994, and Kasper & Rose, 2000).

CHECKING PROMPTS

Once you have a candidate list of prompts, they must be checked so that you can avoid some of the well-known problems with them.

Much of the general work on prompts comes from psychology and consumer research. Though survey questions or prompts should be "so clear and precise that the participants will know exactly what is being asked of them" (Brown, 1997, p. 115), even the clearest questions can be subject to "unreliable" answers from respondents. Backstrom and Hursh-César (1981, pp. 122-123) provide a useful inventory of ways respondents can reply to questions or prompts either inaccurately or untruthfully, ways that can affect the reliability of a survey even if every prompt is "clear and precise." From their inventory, five in particular seem relevant for self-administered surveys in applied linguistics.

The first is "acceptability," the tendency to answer questions in what the respondent thinks might be the most popular way. A language learner taking the Strategy Inventory for Language Learning as reported in Wharton (2000) might, for example, mark that he or she "learns from language mistakes" simply because he or she heard one should or people do learn from mistakes. Were I to take the SILL, I may well mark an answer high on a Likert scale for that question, even though on reflection or asked the same question in an interview I might more accurately answer "no," or "I don't know." Procter & Gamble in a decades-old survey on tooth brushing found a lovely example of the "acceptability" bias. Asked why they brush their teeth, most people answered "to get rid of food that might rot and cause cavities." Asked when they brushed their teeth, however, most people answered "when I get up." The firm concluded that people really brushed their teeth to make their mouths taste good.

Prompts or questions susceptible to acceptability unreliability may interact with "saving face" bias, the tendency to avoid embarrassing or socially inadequate answers. In a report of a survey administered several times to the same groups, Long and Russell (1999), report that participants' reactions to the prompt "student will speak better from class" became significantly more positive upon repeated administrations of the instrument. The drive to save face, to report the class not to have been a waste of time (or detrimental) in this respect may well have driven some of the change.

Many questions in the instrument of Long and Russell may too have been subject to the "partisan" bias, by which respondents are motivated to try to match their responses to the administrator's needs or biases. In this case, the surveys were given to intact classes of university students by their teachers.

Both the SILL survey (of 80 items) and the Long and Russell survey (of 32 items) might have suffered because of the "irrelevance" bias: the instruments may have seemed irrelevant to the concerns of some of the respondents, so they may have answered randomly simply to get the test over with.

Finally, the "halo effect" may affect reliability, especially in surveys that do not allow opting out (Bonikowska, 1988; see Silva, 2000, for a good use of opting out). The halo effect

describes the tendency to answer a question the respondent has never thought of before simply because the question was asked. Nakamura (1992), for example, asked respondents their opinion of the importance of Japanese English students' "certainty of the grammatical accuracy," "sureness of the phonological accuracy," and "confidence in the choice of words." Though I have been teaching mostly Japanese students for some time, it had never occurred to me to consider the importance of any given student's or students' in general "sureness" in phonology, rather than ability in phonology.

The common problems listed above are recounted with the assumption that the wording of the prompt is as clear as possible. Perfect clarity, however, is impossible, as different people can react differently to the same words (*e.g.,* Helfrich, 1986) and the same people may react differently to the same prompts at different times (*e.g.,* Molloy, 2004b).

Helfrich (1986) points out some problems that can be caused by the use of specific words in English-language surveys. (I assume that similar problems obtain with other languages, but have not seen research on the point.) Helfrich provides a useful model of how participants understand questionnaire items and a convenient review of the psychologic literature regarding the effects of some aspects of wording and word placement in questionnaire prompts, pointing out that the implicit assumption of researchers using questionnaires is that "the understanding and the mental representation of items is similar for all individuals" (p. 179). This implies that differences in responses from participant to participant will be attributable to differences in the construct in each participant, these differences in turn attributable to, for example, personal history, personality type, or mood (see, for example, Molloy, 2004b, for an example of how mood can affect responses to L2-research questionnaires).

Negatively worded items in surveys are well-studied. Barnette (2000) provides a review of the literature, as well as empiric evidence that negatively worded items should be avoided if possible. The alternative Barnette proposes has the additional benefit of possibly ameliorating the effect of response sets as well. (It should be noted that other writers, *e.g.*, Dolle, 2001, argue against Barnette's position, but I find Barnette more convincing.) It probably would have been a wise decision for Chiba and Matsuura (2004) to not have used negatively worded prompts in their surveys, given what is known about the threats to validity negatively worded items pose. A consideration of the possible effects of negative wording of prompts would seem to be especially important for the reader of the Chiba and Matsuura article, as the wording of some items was changed between the first and second surveys they report. An example would be the change in the item referring to grammar teaching: from the first survey (item 4), "It is not necessary that English be a required course at university level in Japan" (p. 102, Table 1); from the second survey (item 3) "It is necessary for English to be a required course at university level in Japan" (p. 115, Appendix).

Besides literally negative statements (i.e., those containing the words "no," "not," or variations thereupon), there are also statements that contain "psychological negations," or negative connotations (Helfrich, 1986; Cohen, Manion, & Morrison, 2000). Hence, it was good for Chiba and Matsuura (2004) change the wording of the "game" item between their first and second surveys, substituting a psychological negative for a literal negative: in the first survey (item 14), "Game-oriented activities are not appropriate for university level students in Japan" (p. 102, Table 1); in the second survey (item 16), "Game-oriented activities are childish for university level students in Japan" (p. 116, Appendix).

An additional complication is instantiated with the effective double negatives (and consequent greater difficulty in processing) of items such as numbers 24 from both surveys in the Chiba and Matsuura (2004) article: "Students' reticence is not a problem for me to teach them." The reader wonders what the effect of using a perhaps less-negatively connoted word than "reticence," such as "reserve" would have had on the participants' responses.

The surveys Chiba and Matsuura (2004) have, in some ways, good use of items designed to elicit information about participants' behavior as teachers: they use vague, rather than specific, comparison words in those prompts that seem to be to elicit reports about teaching practices, at least insofar as the prompts they use do not seem to force the participants to assume the additional burden of recalling specific episodes (*cf.* Gaskell, Wright, & O'Muircheartaigh, 1993 on the difficulties involved in using more specific prompts to ask about behavior in the past).

However, a problem with the Chiba and Matsuura (2004) surveys, as wholes would seem to be the intermixing of implicitly comparative items and absolute items. It is well known that survey respondents tend to manifest response sets (Backstrom & Hursh-César, 1981). Response sets are habitual answering styles (such as circling only "agree") that have nothing to do with the prompt content. Response sets are especially likely when the same response scale is used throughout a survey (Barnette, 2000). The response set makes it seem an unwise idea for researchers to mix on a single survey prompts such as "I seldom correct students' grammatical mistakes" (second survey, item 9, p. 116, Appendix) and "Japanese students need to learn communication skills such as interrupting and turn-taking" (second survey, item 21, p. 116, Appendix). Note that the former item asks respondents to evaluate or report on their own behavior, presumably by comparing it with other aspects of their behavior and contains the comparison word "seldom," while the second item is written so that it contains no such word: it is an absolute statement. The danger with the latter item is that of what we might call "scale appropriation" or "scale construction" (*cf.* Low, 1999): respondents may treat the response scale for the latter item not as degree of agreement with the statement, but as a measure of how important the skill(s) in question are.

Item 21 from the second Chiba and Matsuura (2004) survey is an example of another common class of problematic prompts: the "double-barreled" (Brown, 2001) question. This is a form of prompt that can be considered to be asking two different questions. With "Japanese students need to learn communication skills such as interrupting and turn-taking," for example, it may be that participants were being asked to respond to two different ideas: "students [in general] need to learn" and "*Japanese* students [as opposed to students of other nationalities] need to learn." A similar ambiguity may be in found in a survey by Ito (2004). Ito surveyed parents who had children in French immersion programs in Canada; one item (Q1, p. 126) concerned reasons for the parents' choice. One response choice was "intrinsic value in learning the francophone culture in Canada." The possibility here is that some participants may have interpreted the question as referring to "francophone culture in Canada [and not in other places]" an some as "francophone culture in Canada [as an example of possible francophone cultures]."

AVOIDING PROBLEMS

It must be mentioned again that any prompt is potentially problematic, given the ambiguity in language and the differences in how people can react from day to day to the same language. Nevertheless, survey writers can do some things to ameliorate potential problems. Following standard procedures and discussing possible problems in your report can go far to instill confidence on the part of your readers.

THINK-ALOUD PROTOCOLS AND RETROSPECTIVE DEBRIEFING

Potential prompts should be discussed with other researchers, colleagues, and members of the target population. This is especially important when participants can be considered to

belong to a culture, subculture, age group, or other group different from that or those of the researchers.

Besides informal discussion, two formalized elicitation procedures have proved useful in detecting potential problems or ambiguities in survey prompts. These are known as think-aloud protocols and retrospective debriefing.

In a think-aloud protocols, a person from the population of interest is asked to respond to the prompts while attempting to articulate everything that goes through his or her mind. The process is most often used in L1 studies of mathematical reasoning or writing research (see Bracewell & Breuleux, 1994, Smagorinsky, 2001, or Williamson *et al.,* 2000, for reviews), but has been advocated for language studies (Ericsson & Simon, 1985) and second-language studies (Ericsson & Simon, 1987). In applied linguistics surveys, think-aloud protocols have been used by Robinson (1992) and Molloy (2004a) to particularly useful effect for finding otherwise unidentifiable factors that might complicate responses to prompts. Robinson found factors associated with might be called the culture of his participants (Chinese women using English). In my study, listening to participants talk and timing their pauses allowed me to identify confusing or unnecessary information that may have put a needless cognitive burden on participants. The presence of needless information was something that had not become apparent with interviews and discussions about the prompts I had had before: it was only because think-aloud protocol participants had hesitated at similar times that the potential problem became apparent.

Retrospective debriefing (for a review, see Taylor & Dionne, 2000) resembles think-aloud protocols, but takes place after, not during, responding to stimuli and allows the researcher to collect responses that are deeper than those in think-aloud protocols. In retrospective debriefing, participants are asked to talk their way through their responses immediately after responding. It is most productive to audio- or videorecord participants while they react to the stimuli and use the recording as a reminding device, playing and pausing the recording while asking "What were you thinking here?"

PILOTING

When a final groups of prompts has been selected, the survey should be piloted with a group of participants similar to those in the population of interest and analyzed as if it were a final survey. Piloting allows further potential problems, particularly those in administering the survey, to be identified. For example, in a pilot administration for a survey I used last year (Molloy, 2004b), I found about a third of the surveys were returned half completed simply because some participants did not realize they were supposed to turn the paper over.

More important, analysis of the pilot data allows identification of prompts that do not behave as expected or are useless. Prompts that do not behave as expected are those that elicit response patterns that do not match those of prompts that are ostensibly similar. Because surveys are used to investigate behaviors or attitudes that are characteristic of groups of people, prompts that elicit responses that form unique patterns are probably prompts that actually measure something besides the point of interest. For example, in a survey designed to elicit different responses from women and men, women should respond roughly similarly to prompts and in a way different from the way men respond. If one prompt elicits a response pattern that does not differ with the gender of the participants, that prompt should be considered for deletion. (A analogous discussion in the testing area is available in Molloy, 2004c.)

Prompts that are useless are those that elicit response patterns identical to those of another response or that only elicit one kind of response. It would be useless, for example, to ask both "Do you prefer dogs to cats?" and "Do you prefer cats to dogs?" as the responses to the two

prompts (with a yes-no response format) would probably be mirror images of one another, allowing the answers from one question to be predicted from the answer to the other. The second question does not yield any further useful information. Likewise, questions that elicit only one response (such as when everyone responds "strongly agree") from participants add no further information and should be eliminated to reduce the burden on participants.

TRANSLATING PROMPTS

In L2 or any cross-cultural survey research, instructions and prompts should be translated from the language in which the material is originally written into a language that participants can understand. This can be a complicated process in, for example, studies of second (not foreign) language classrooms, immigrants, or other populations with various strongest languages. Some of the surveys summarized in the translation chapter by Brislin and co-workers (1973) involve situations in which no one who knows both the original language and the target language can be found. Fortunately, in Japan many surveys are used with persons with a single strongest language.

Even in cases in which the survey involves the ability to comprehend an L2, "translation" may still be necessary. In piloting a recent survey I conducted (Molloy, 2004b), participants commented that an e-mail message I was using as a prompt and that was ostensibly written by a university student was too formal. I finally gave the letter to a much younger colleague for "translation" into informal English.

Besides logistic difficulties, the main problem in translating for surveys is assuring that the translation and the original are equivalent. The most commonly used translation method in second-language research is back translation. This involves asking two or more bilinguals to translate each piece of the survey from the original language to the target language, then getting two more bilinguals to translate the target language material back into the original language. It is best if the translators do not know the specific purpose of the materials. Using more than one translator is necessary at each step to ensure that idiosyncratic translations are not mistakenly used. The original writing and the twice-translated language are then compared.

If the original and the twice-translated passages match, the next step is to check if the translations are idiomatic. This can be done by asking someone proficient in the target language (and, ideally, with no ability in the original language) to rewrite the translated passage into idiomatic language, which is then checked by another proficient bilingual.

When the two pieces compared do not match, the researcher can rewrite the original and go through the process again (using different translators, if possible). Another option, and one that I frequently use, is to simply discard the problematic prompt.

In cases (such as instructions or prompts involving essential construct words) in which pieces cannot be discarded, the two further options are to present the materials in both languages or, if the number of participants is large enough, to use a counterbalanced design in which rival translations are used in the survey instruments. In the e-mail study I did (Molloy, 2004b), the word "appropriate," which has several possible translations in Japanese, proved problematic. I tried to counteract the possible effects of the difference between "appropriate" and "*tekisetsuna*," the Japanese equivalent most often suggested by the translators, statistically. I made 100 survey forms with both "appropriate" and "*tekisetsuna*," 100 with "*tekisetsuna*" only, and 100 with "appropriate" only. These were randomly distributed to the participants, which enabled me to see if the response patterns in the three different versions of the prompt differed more than could be expected by chance. Luckily, they did not. If they had, I would have attempted a different approach to checking the notion of appropriate.

Brislin and co-workers (1973) summarize research on English that is most successfully translated (pp. 33-35) and note that, generally, the more specific and concrete the original English is, the more often translation will be successful. This means using the active voice, repeated nouns rather than pronouns, and sentences as short as possible.

There are other methods of attempting to ensure translation equivalence. One, for example, is translation by committee. The empirical research I have seen seems to indicate, however, that the back-translation method is most reliable.

CHOOSING A PROMPT STYLE

Several prompt and response styles mentioned as seldom or never used in the literature would be fruitful areas for research; however, I recommend that such research be undertaken in the course of replicating previously published research. This is because the error and crudity of focus in survey research means that isolating the effects of prompt style or response format is nearly impossible with new survey foci, as it is difficult or impossible to attribute variation convincingly.

SHORT ANSWER

Using short, free response answers circumvents problems of scale or response interpretation to some extent. However, short answers are difficult to code, score, and (if necessary) translate, not to mention expensive, as at least two coders must go through all of the data independently, and it is best to use at least four independent translators. Large-scale written surveys with short answers are commonly undertaken, particularly in the field of public health, but only with plenty of funding. Even one of the smaller short answer surveys I have been involved in (Molloy & Shimura, 2003), involving only 2766 written answers, took two people about a solid month to process, before any analysis.

The short answer format can yield much richer data than other formats, but does necessitate a great deal of work. One potentially useful format (which I have not yet seen used in the applied linguistic area) might be to combine a short-answer format with a more restricted response format, such as asking participants to agree or disagree with a statement and then explain their choice. This might be an especially good option when it seems likely that participants will have a "yes, but…" response in mind, as pointed out by Carrick (personal communication, 12 November 2004).

MULTIPLE-RESPONSE CHECKLISTS

Multiple-response checklists are lists of prompt responses from which participants can choose two or more options (see, *e.g.,* Ito, 2004). These are particularly useful when the researcher's intent is to describe the characteristics of one group. A drawback to multiple-response checklists is that they make cross-group comparisons conceptually and analytically difficult.

RANKING TASKS

Ranking tasks involve ordering series of statements or other prompts, for example, from "most important" to "least important." The ranking approach has been used often in consumer research (see, for a thorough explanation and some relevant examples, Schiffman, Reynolds, & Young, 1981), but seems to have fallen out of favor for two reasons. First, it is not amenable to more familiar and more powerful statistical analyses. Second, it has been

critiqued because the lists of stimuli are often longer than can be held in short term memory at once, allowing critics to argue that ranking tasks are psychologically a series of two-item preference tasks. One well-known study using ranking tasks, however, was published by Carrell and Konneker (1981), who asked participants to rank stimuli in order of politeness.

I do not know that multiple-response checklists have yet been used with rating tasks. In studies of motivation, for example, it may be illuminating to have participants choose several responses from a multiple-response checklist and then rank their selections by importance.

CODINGS

A possibility that I have not seen used in applied linguistics survey research is to ask participants to categorize stimuli by assigning codes to them. Codes differ from ratings or rankings in that they do not involve comparison of one or more prompts with others: prompts are assigned one by one to predetermined categories. Codings in surveys have been used extensively in anthropological linguistics. Further information can be found in that area, such as the references given by Wierzbicka (1994) and her critics (*e.g.,* Harré, 1993).

LIKERT-SCALE

Likert scales are series of statements about which participants indicate to what degree they agree. Busch (1993) and Turner (1993) give good summaries of Likert scales in applied linguistics, and recent uses of Likert scales can be found in much of the applied linguistics research on motivation (*e.g.,* Sawada, 2004 or Cheng, Horwitz, & Schallert, 1999).

Likert scales and Likert-type scales are immensely popular, but tend to be misused. One way in which true Likert scales have been questionably used is discussed by Clason and Dormody (no date): most of the foundational research done by Likert himself was done on the assumption that the entire series of statements would form a scale and that the sum of scores on all items related to a single construct would be used to indicated the degree to which that single construct was held.

That is, true Likert scales are when each statement is related to one and only one construct. Their practical advantage over Thurstone scales is that they allow fewer prompts to be used to measure the same construct.

In practice, Likert (and Likert-type) scales are often used as if individual statements or subgroups of statements indicate the strength of particular constructs. The research that originally made Likert scales popular, however, did not take this possibility into account.

LIKERT-TYPE SCALE

Likert-type scales, as the name implies, resemble Likert scales, except that the solicited response is not simple agreement or disagreement. For example, the response may be a report on learning strategies (*e.g.,* Isoda, 2004) asking how often a participant does something or a test of grammatical acceptability (*e.g.,* Gass, 1994) asking how acceptable a sentence is.

Likert-type scales likewise are immensely popular survey tools, but in practice (if not in theory) seem to be particularly associated with prompt-response mismatches (as discussed above) and with inappropriate statistical analyses (as discussed below).

Likert-type scales have been used in combination with yes-no response formats. Bardovi-Harlig and Dörnyei (1998) and the replication of their study done by Niezgoda and Röver (2001) asked participants asked participants whether particular pieces of language were correct or appropriate and, with negative responses, asked participants to rate how bad the problem was.

SEMANTIC DIFFERENTIAL SCALES

Semantic differential scales are most often used to discover attitudes, as with the AMTB created by Gardner and Lambert (Gardner & Lambert, 1972; Gardner, 1985; or Gardner, Masgoret, & Tremblay, 1999). With the semantic differential, participants read a prompt stem, such as "my teacher is" and a series of opposites such as "strict" and "lenient" separated by some kind of scale. Participants indicate where along the continuum the prompt stem is.

Semantic differential scales may be misleading when used with word pairs that are literal, but not psychologic, opposites. For example, it would be difficult to see how to interpret differing marks on a scale separating "correct" and "incorrect."

THURSTONE SCALES

Thurstone scales are seldom used in applied linguistics (or any other field) these days. This is mostly because (a) they take much more work to construct than other scales and (b) many older studies have been taken to indicate that the easier Likert and Likert-type scales are acceptably accurate (Andrich, 1988). However, recent research by Roberts, Laughlin, and Wedell (1999) seems to show that Thurstone scales may be more accurate in measuring extreme responses, such as those near the ends of a "strongly agree" to "strongly disagree" scale.

Garson (2004) gives a summary of some of the steps in constructing Thurstone scales. The complicated scale development begins with having expert judges rank series of statements regarding how much the statement is indicative of the construct of interest. The most consistently agreed upon statements from each rank are then used to construct the scale.

Thurstone scales, then, necessitate having a very large preliminary prompt pool, as statements the judges do not agree upon will be eliminated. They may also necessitate the participants' reading a larger number of statements, as the degree to which a particular participant holds the construct of interest is measured by the number of (and which) statements they agree or disagree with.

Nevertheless, it would seem that Thurstone scales should be investigated in applied linguistics contexts: they might be particularly useful in research (such as that done by Chiba and Matsuura, 2004, or Matsuura, Chiba, and Hilderbrandt, 2001) that involves constructs that tend to elicit extreme reactions, such as the use of an L1 in L2 classrooms.

CHOOSING A RESPONSE FORMAT

There are several response formats that can be used with survey prompts. Most of them need to be fitted with labels of one sort or another. The researcher should pay particular attention to the interaction between the prompt statement and the response format labels. Low (1999) presents convincing evidence that scales are sometimes "appropriated" by participants and used in ways not intended by the researcher. Example might be found with some of the prompt and response format pairs in the surveys conducted by Chiba and Matsuura (2004). For example, item 11 of the first survey the researchers report is "I seldom correct students' grammatical mistakes" (Table 1, p. 102). The response format is a Likert-type scale anchored with the phrases "strongly agree" and "strongly disagree." It is a little difficult to understand what might be meant with a response to this item of "strongly disagree." Does it mean "I strongly disagree that I seldom correct students' grammatical mistakes"? How is this different from "I disagree that I seldom correct students' grammatical mistakes"? The researchers here are literally asking for an opinion on a statement about the frequency of a particular behavior

of the participant: it would seem that a "yes" or "no" response format (that is, confirmation rather than degree of agreement) would be more appropriate for such a prompt. The pitfall Low (1999) points out is that in situations in which there is a mismatch between the statement and the response format, participants may respond to a different scale than the one the researcher intended. In the Chiba and Matsuura item 11 case, perhaps participants were responding to this prompt as if the scale were one asking for the frequency of behavior, converting "strongly agree" to "I always correct students' grammatical mistakes" and "strongly disagree" to "I never correct students' grammatical mistakes" and responding to that idea. Perhaps they were instead converting the prompt to "It is important to correct students' grammatical mistakes." Whatever the cases may have been (and they may have differed from participant to participant), the mismatch between the statement and the response format makes it difficult to have much confidence in the results obtained.

VISUAL ANALOGUE

The visual analogue response format consists of an unbroken line anchored by two alternatives. For example, in a study I did (Molloy, 2004a), I used a visual analogue scale anchored by "same" and "different" to ask participants how much alike two written passages were in terms of appropriateness. In another study (Molloy, 2004b), I used eight visual analogue scales with Likert scale items, anchoring the line with "I agree" and "I disagree." Visual analogue response formats are often seen with semantic differential scales.

To use a visual analogue scale, the participant simply makes a mark (a circle, a check, or, best, an "x") on the line to indicate, for example, how much he or she agrees or disagree with a statement. The distance from one end of the scale to the mark is then measured and converted to a proportion to give an interval-scale statistic for degree of agreement.

The visual analogue scale has the advantage of being a true interval scale, which means that it can properly be used with many of the more familiar statistical tests, such as *t*-tests or Pearson's correlation coefficient.

Many researchers, however, argue that visual analogue response formats suffer from the same general limitation that ranking tasks do: that is, that the human mind simply cannot usefully make as many distinctions as a visual analogue scale allows. For example, with a 68 mm line measured with a ruler that resolves to 1 mm, it is possible to distinguish 132 different places on the line. It is a stretch to imagine that most people can distinguish 132 degrees of opinion about anything.

A second, practical, drawback with visual analogue responses is that, if they are given on paper, they take a great deal of tedious hand measurement. In a study I did in 2003 (Molloy, 2004b), I wound up having to make more than 36,000 separate measurements with a ruler, a task that consumed most of my free time for more than three months.

MAGNITUDE ESTIMATION

Magnitude estimation is a response format that may circumvent the memory-limitation criticism and the tedium associated with visual analogue response formats. It also releases the researcher from the analytic constraints associated with ratings that use pre-determined scales.

In magnitude estimation, participants are instructed to assign a number (such as 100) to the first of a series of statements to which he or she is to indicate his or her agreement; the next statement is compared with the first and assigned a different number to indicate the participant's greater or lesser agreement. Which numbers the participants choose are up to the participant.

The technique is described and demonstrated conveniently by Allard (2002), who also provides some background information on it. Magnitude estimation has most often been used and tested with estimation of physical properties, such as weight. In these cases, it has been found to be reliable and convenient. In applied linguistics, magnitude estimation has seldom been used (*e.g.,* Bard, Robertson, & Sorace, 1996; Bond *et al.*, 1998), but the method bears investigation. However, until research is done to assess the relative reliability of magnitude estimation compared with more traditional response formats, the method is likely to remain less popular.

LIKERT-TYPE

This is the type of response format most often associated with Likert and Likert-type scales. It includes a range of discrete options anchored by extremes. Examples can be found in Chiba and Matsuura (2004), Isoda (2004), and Sawada (2004). Abdel-Khalek (1998) used a seven-point Likert-type response scale to collect responses to the statement "I am afraid of death," anchoring the scale with "strongly agree" and "strongly disagree." In most cases, researchers use from five to eleven discrete options; more than eleven choices having been found to not add much accuracy to the response format.

The principal choices a researcher must make in using a Likert-type response format are whether to use an odd or an even number of options and whether to label each discrete option. Using an even number of options (as in Chiba & Matsuura, 2004) with a agree-disagree scales, for example, forces participants to choose between agreement and disagreement, while odd numbers can be said to allow a neither agree nor disagree option. The former makes responses easier to analyze.

Labeling each option would seem a good choice for making the intent of the scale more clear, and it could be argued that doing so may forestall or ameliorate scale appropriation (Low, 1999). However, the researcher must take care to assure that the scale only contains points *on* the scale. Gorsuch (2001), for example, investigated teachers' responses to statements about the appropriateness of various classroom activities using a scale with labels of "strongly agree," "agree," "don't know," "disagree," and "strongly disagree." The "don't know" option in this case does not fit into a continuum from "strongly agree" to "strongly disagree" (although "neutral" might). Using labels that do not all belong to the same scale may save space, but may tempt the researcher to analyze the responses as if they all were part of the same scale.

Likert-type response formats are convenient and familiar. The principal drawback with them is not theoretic, but the way they are usually treated analytically. At their most accurate, Likert-type response formats will yield information on an ordinal (or rank-order) scale. Rank-order scales tell the researcher that one option should be ranked higher or lower than another, but not how much higher or lower. This, in turn, means that responses to Likert-type responses cannot be summed (to give an average response) and that most of the familiar and more powerful statistical tests (such as *t*-tests or Pearson correlation coefficients) cannot be used for analysis.

Another potential drawback with Likert-type response formats, discussed earlier, is that they seem to be particularly susceptible to scale appropriation (Low, 1999). Whether this is because Likert-type scales are simply used most often or because they are inherently constituted to allow scales appropriation has not yet been investigated.

Likert-type response formats are essentially similar to the ratings that are most commonly used in testing (*e.g.,* Nunn & Lingley, 2004, or the Asia University oral placement interview discussed by Hansford, 2004). As with ratings used in testing, the reliability of Likert and Likert-type scales depends heavily on having a range of responses to each prompt. If

participants in pilot studies tend to respond all in the same way, it may be a good idea to pilot test a yes-no, binary response format to see if that eliminates the reliability problem.

YES-NO (BINARY RESPONSE)

Binary response formats have the great advantage of being analytically simple and psychologically easy to process. For example, yes-no or agree-disagree response formats can be checked for reliability using the more accurate Kuder-Richardson formulas. At least some research (*e.g.*, Abdel-Khalek, 1999) has showed binary-response formats to yield as much useful information as ordinal response formats such as Likert-type scales; in the applied linguistics area, Bardovi-Harlig and Dörnyei (1998) have used binary response formats to good effect, and Upshur and Turner (1995) have given an illuminating account of constructing such a scale.

The principal drawback of binary response format is that more prompts are necessary to collect the same information that might be collected with a Likert-type response scale. This means not only that will the survey contain more prompts, but also that checking the prompts for ambiguity, translating, and other piloting procedures will take correspondingly longer.

OPTING OUT

The concerns of survey participants and those of researchers are frequently different, and what is important to researchers may not even be an issue to participants or may not be applicable to participants. For example, Isoda (2004) used a survey to study learning strategies. One of the items in the survey was "I work with my neighbors to check the answers" (no. 27, p. 13). Although it would be a simple matter for me to choose one of Isoda's response options, the language lessons I have are one-to-one and the entire question irrelevant to my situation. Likewise, in Sawada's survey on motivation (2004), one of the items is "If I learn [a language] better I will be able to avoid senility" (item 29, pp. 62). Before reading this item, I had never thought of avoiding "senility" before and certainly not in connection with language learning. As a participant, I would be easily able to mark a response option for either of these prompts, but in neither case would my response be related to my actual attitude.

In any survey in which the "halo," "irrelevance," or "partisan" biases discussed above seem likely to obtain, response options should include opting out, or the choice of not registering an opinion about the topic. An extensive and convincing argument for the importance of opting out in applied linguistics is given by Bonikowska (1998), but in summary allowing participants to opt out will give a more realistic portrait of the object of study. A drawback to participants' opting out is that the analysis of results becomes somewhat more difficult; fortunately, computers have ameliorated much of the computational tedium in analysis, so that increased analytic complexity probably should not be an excuse for forcing responses to every item in a survey.

HOW MUCH SPACE TO USE FOR SHORT ANSWERS

Because so many things may vary in short-answer responses, it is best to leave more space than the physically biggest response collected in pilot studies takes. If space-limitation effects are of particular concern, consider piloting items one to a page to allow maximal blank space. In the only (unpublished) empirical studies I know of concerning the use of white space, I found that none of 77 participants in a pilot study used more than 160 mm of white space in writing in response to a prompt soliciting written advice; in a later study, in only one response from among about 3000 was 160 mm not enough space for a response.

ORDER EFFECTS

"Order effects" is a general term that refers to earlier prompts affecting responses to later ones. Subsumed in this idea is the idea of the effect of fatigue: in responding to a long survey, participants may become tired, bored, or distracted toward the end of the survey and pay less attention to the prompts, making the answers toward the end of the survey. With fatigue, it is not the contents of the earlier prompts that cause problems, but simply that they *are* earlier.

If you believe that this may be a factor in a survey you create, in some cases you can ameliorate the statistical effects of order effect by using different randomly ordered forms. For example, in a survey I have been working on for the past three years (described in greater detail in Molloy, 2004b), I asked participants to read 12 pairs of letters (about half an A4 printed page) in a second language and answer 8 questions (in Japanese) about one letter from each pair (a total of 96 responses). I worried that the task would tire the participants and make the later responses less reliable than the earlier ones. To prevent systematic effects of fatigue (though not, of course, the fatigue as such), I systematically ordered the survey instruments so that each of the 192 participants received a survey with the prompts in a different order. In other cases, I have simply randomly distributed the same prompts, but in different orders, using five to ten different survey forms.

CONSENT FORMS AND PRIVACY

In current practice in the United States, research that involves human participants requires that the participants be made aware of what their participation in the study means. This includes the purpose of the study, the information that will be collected from the participants, who will have access to the information, possible benefits and detrimental effects of the study, and that their participation is voluntary. In many cases, funding or journal publication is not possible without such consent having been obtained. More importantly, it is simple ethical behavior to assure that participants be informed of the consequences of participating in a study and of their rights with regard to the information they provide.

Each survey should be accompanied by an informed consent agreement signed or sealed by the researcher or researchers. No survey response should be used unless the participant has indicated in writing that he or she is aware of the consequences of participating in the study, and it is up to the researcher to allow participants to withdraw from the study even if they have given written permission before. Written information about the study should be given to participants at least once, although I have at times used as many as three written, bilingual informed-consent agreements with each participant in conjunction with follow-up interviews to reemphasize that participants can drop out at any time.

If the researcher is prepared to compensate for partly, incorrectly, or ambiguously filled-out survey forms, then having participants respond to surveys anonymously is an option. This necessitates assigning randomly generated participant codes and using physically separate informed-consent agreements. In my own practice, I usually use confidential participation agreements, in which the researcher(s) and the participant are aware of which participant contributed which response, but the information is kept confidential and any information that might be used to identify a particular participant is altered or deleted. In one case I was involved in, information that could have been used to identify the participant (a well-known public figure) could not be altered without changing the substance of the participant's responses. That participant was dropped from the study. Allow at least 30 minutes to sign 100 informed consent agreements.

PILOTING AND RELIABILITY

Once a preliminary survey form has been constructed, a final procedure that will allow potential problems to be identified is piloting the entire survey instrument with a group of participants as similar as possible to those who will respond to the final version of the survey. The group does not have to be as large as possible, but it should be large enough to allow for a realistic simulation of the conditions in which the survey will be used.

Those conditions include conditions of analysis: because surveys have to be checked for reliability it may be necessary to use a minimal number of pilot participants, depending on the reliability check you will be using. Reliability (to be discussed in the companion article) must be assessed when the pilot administration is completed. Because unreliable surveys give unreliable results, it saves much time, effort, and frustration to abandon or redesign an unreliable survey as soon as possible.

References

Abdel-Khalek, A. M. (1998). Single- versus multi-item scales in measuring death anxiety. *Death Studies*, 22(8), 763-772.

Allard, F. (2002). *An example of magnitude estimation* [WWW document]. URL http://www.ahs.uwaterloo.ca/~kin356/example.html.

Andrich, D. (1988). Thurstone scales. In J. P. Keeves (Ed.). *Educational research, methodology, and measurement: An international handbook* (pp. 303-306). Oxford: Pergamon.

Backstrom, C. H., & Hursh-César, G. (1981). *Survey research (second edition)*. New York: Macmillan.

Bard, E. G., Robertson, D., & Sorace, A. (1996). Magnitude estimation of linguistic acceptability. *Language*, 72, 32-68.

Bardovi-Harlig, K., & Dörnyei, Z. (1998). Do language learners recognize pragmatic violations? Pragmatic versus grammatical awareness in instructed L2 learning. *TESOL Quarterly*, 32(2), 233-262.

Barnette, J. J. (2000). Effect of stem and Likert response option reversals on survey on survey internal consistency: If you feel the need, there is a better alternative to using those negatively worded stems. *Educational and Psychological Measurement*, 60(3), 361-370.

Beebe, L. M., & Takahashi, T. (1989). Do you have a bag?: Social status and patterned variation in second language acquisition. In S. Gass, C. Madden, D. Preston, & L. Selinker (Eds.), *Variation in second language acquisition: Discourse and pragmatics* (pp. 103-125). Clevedon, Avon: Multilingual Matters Ltd.

Blum-Kulka, S., & Olshtain, E. (1984). Requests and apologies: A cross-cultural study of speech act realization patterns (CCSARP). *Applied Linguistics*, 5, 196-213.

Brislin, R. W., Lonner, W. J., & Thorndike, R. M. (1973). *Cross-cultural research methods.* New York: John Wiley & Sons.

Bond, Z. S., Fucci, D., Stockmal, V., & McColl, D. (1998, 30 November-3 December). *Multi-dimensional scaling of listener responses to complex auditory stimuli.* Paper presented at the 5th International Conference on Spoken Language Processing, Sydney, Australia. [WWW document]. URL http://wwwhome.cs.utwente.nl/~taaltool/Icslp98/HTML/SL98S106.HTM#p0163.

Bonikowska, M. P. (1988). The choice of opting out. *Applied Linguistics*, 9(2), 169-181.

Bracewell, R. J., & Breuleux, A. (1994). Substance and romance in analyzing think-aloud protocols. In. P. Smagorinsky (Ed.), *Speaking about writing: reflections on research methodology* (pp.55-88). Thousand Oaks, CA: Sage.

Brown, J. D. (1997). Designing surveys for language programs. In D. T. Griffee & D. Nunan (Eds.), *Classroom teachers and classroom research* (pp. 109-121). Tokyo: The Japan Association for Language Teaching.

Brown, J. D. (2001). *Using surveys in language programs.* Cambridge: Cambridge University Press.

Brown, J. D., Robson, G., & Rosenkjar, P. R. (2001). Personality, motivation, anxiety, strategies, and language proficiency of Japanese students. In Z. Dörnyei & R. Schmidt (Eds.), *Motivation and second language acquisition* (pp. 361-398). Honolulu: University of Hawaii.

Brown, P., & Levinson, S. (1987). *Politeness: Some universals in language usage.* Cambridge: Cambridge University Press.

Busch, M. (1993). Using Likert scales in L2 research: A researcher comments… *TESOL Quarterly*, 27(4), 733-736.

Cai, D., & Wilson, S. R. (2000). Identity implications of influence goals: A cross-cultural comparison of interaction goals and facework. *Communication Studies*, 51, 307-328.

Carrell, P. L., & Konneker, B. H. (1981). Politeness: Comparing native and nonnative judgments. *Language Learning*, 31, 17-30.

Cheng, Y., Horwitz, E. K., & Schallert, D. L. (1999). Language anxiety: Differentiating writing and speaking components. *Language Learning*, 44(3), 471-446.

Chiba, R., & Matsuura, H. (2004). Diverse attitudes toward teaching communicative English in Japan: Native vs nonnative beliefs. *Asia University Journal of International Relations*, 13, 97-118.

Church, A. T. (2000). Culture and personality: Toward an integrated cultural trait psychology. *Journal of Personality*, 68, 651-703.

Clason, D. L., & Dormody, T. J. (No date). Analyzing data measured by individual Likert-type items. *Journal of Agricultural Education*, 35(4), 31-35.

Cohen, L., Manion, L., & Morrison, K. (2000). *Research methods in education* (5th ed.). London: Routledge/Falmer.

Connolly, M. (2004). Revisiting the special challenges of teaching lower-level Freshman English. *CELE Journal*, 12, 5-23.

Dolle, R. (2001). Who wants to try a questionnaire? *Journal of Environmental Health*, 63, 38-40.

Dörnyei, Z. (2003). *Questionnaires in second language research: Construction, administration, and processing.* Mahwah, NJ: Lawrence Erlbaum Associates.

DuFon, M. A., Kasper, G., Takahashi, S., & Yoshinaga, N. (1994). Bibliography on linguistic politeness. *Journal of Pragmatics*, 21, 527-578.

Ericsson, K. A., & Simon, H. A. (1985). Protocol analysis. In T. J. van Dijk (Ed.), *Handbook of discourse analysis: Vol. 2: Dimensions of discourse* (pp. 259-268). London: Academic Press.

Ericsson, K. A., & Simon, H. A. (1987). Verbal reports on thinking. In C. Færch & G. Kasper (Eds.), *Introspection in second language research* (pp. 24-53). Clevedon, Avon, England: Multilingual Matters.

Ford, C. E. (1984). The influence of speech variety on teachers' evaluation of students with comparable academic ability. *TESOL Quarterly*, 18, 25-40.

Gardner, R. C. (1985). Appendix A: Instructions and items from the attitude/motivation test battery. In *Social psychology and second language learning* (pp. 177-183). London: Edward Arnold.

Gardner, R. C., & Lambert, W. E. (1972). *Attitudes and motivation in second-language learning.* Rowley, MA: Newbury House.

Gardner, R. C., Masgoret, A-M., & Tremblay, P. F. (1999). Home background characteristics and second language learning. *Journal of Language and Social Psychology*, 18, 419-437.
Garson, G. D. (2004). Scales and standard measures. In *Statnotes: An online textbook* [WWW document]. URL http://www2.chass.ncsu.edu/garson/pa765/standard.htm.
Gaskell, G., Wright, D., & O'Muircheartaigh, C. (1993). Reliability of surveys. *The Psychologist*, 11, 500-503.
Gass, S. M. (1994). The reliability of second-language grammaticality judgments. In E. E. Tarone, Cohen, & S. M. Gass (Eds.), *Research in Second Language Learning* (pp. 303-322). Mahwah, NJ: Lawrence Ehrbaum.
Gorsuch, G. (2001). Japanese EFL teachers' perceptions of communicative, audiolingual and *yakudoku* activities: The plan versus the reality. *Education Policy Analysis Archives*, 9(10) [ISSN 1068-2341]. [WWW document]. URL http://epaa.asu.edu/epaa/v9n10.html.
Hansford, V. (2004). CELE's OPI training system. *CELE Journal*, 12, 103-106.
Harré, R. (1993). Universals yet again: A test of the 'Wierzbicka Thesis.' *Language Sciences*, 15, 231-238.
Helfrich, H. (1986). On linguistic variables influencing the understanding of questionnaire items. In A. Angleitner & J. S. Wiggins (Eds.), *Personality assessment via questionnaires* (pp. 178-188). Berlin: Springer-Verlag.
Horwitz, E. K. (2001). Language anxiety and achievement. *Annual Review of Applied Linguistics*, 21, 112-126.
Hullett, C. R., & Tamborini, R. (2001). When I'm within my rights: An expectancy-based model of actor evaluative and behavioral responses to compliance-resistance strategies. *Communication Studies*, 52, 1-16.
Isoda, T. (2004). exploring learners' thoughts and attributes affecting learning strategy use. *JACET Bulletin*, 39, 1-14.
Ito, H. (2004). Parental evaluative perceptions of immersion education in Canada. *JACET Bulletin*, 39, 123-135.
Kachru, Y. (1994). Crosscultural speech act research and the classroom. *In Pragmatics and language learning. Monograph series Vol. 5* (pp. 39-31 [*sic*]). [Eric document E 398 739].
Kasper, G., & Rose, K. (Eds.) (2000). *Research methods in interlanguage pragmatics.* Mahwah, NJ: Erlbaum.
Kreuz, R. J., & Roberts, R. M. (1993). When collaboration fails: Consequences of pragmatic errors in conversation. *Journal of Pragmatics*, 19, 239-252.
Kuha, M. (1997, April). *The influence of naming the target speech act in instructions on production questionnaires.* Paper presented at the 11th Annual International Conference on Pragmatics and Language Learning at the University of Illinois at Urbana-Champaign.
Long, R. W. III & Russell, G. (1999). Looking back: Student attitudinal change over an academic year. *The Language Teacher*, 23(10), 17-27.
Low, G. (1999). What respondents do with questionnaires: Accounting for incongruity and fluidity. *Applied Linguistics*, 20, 503-533.
Lunneborg, C. E. (2001). Random assignment of available cases: Bootstrap standard errors and confidence intervals. *Psychological Methods*, 6, 402-412.
MacIntyre, P. D., Baker, S. C., Clément, R., & Conrod, S. (2001). Willingness to communicate, social support, and language-learning orientations of immersion students. *Studies in Second Language Acquisition*, 23, 369-388.
MacIntyre, P. D., Clément, R., Dörnyei, Z., & Noels, K. A. (1998). Conceptualizing willingness to communicate in a L2: A situational model of L2 confidence and affiliation. *The Modern Language Journal*, 82, 545-62.
Matsuura, H., Chiba, R., & Hilderbrandt, P. (2001). Beliefs about learning and teaching communicative English in Japan. *JALT Journal*, 23, 69-89.

Miller, J., & Cann, R. (1998). Data collection in linguistics. In J. L. Mey (Ed.), *Concise encyclopedia of pragmatics* (pp. 194-196). Amsterdam: Elsevier.

Molloy, H. P. L. (2004a). Is appropriate appropriate? An investigation of interpersonal semantic stability. *Proceedings of the 2003 JALT PAN-SIG Conference.* Tokyo: Japan Association of Language Teachers.

Molloy, H. P. L. (2004b, February). *Approaches to reliability calculations in pragmatic acceptability tests.* Paper presented at the Temple University Japan Research Colloquium, Tokyo.

Molloy, H. P. L. (2004c). How reliable is the Asia University Freshman English Placement Test? A classical internal reliability study. *CELE Journal*, 12, 64-86.

Molloy, H. P. L., & Shimura, M. (2002, September). *Production and recognition difference in Japanese university students' English-language complaining.* Paper presented at the JACET 41st Annual Convention, Tokyo.

Molloy, H. P. L., & Shimura, M. (2003, September 5). *Approaches to a theory of complaint interactions.* Paper presented at the JACET 42nd Annual Convention, Sendai, Japan.

Molloy, H. P. L., & Shimura, M. (2004). [Partial replication and reanalysis of Beebe, Takahashi, & Uliss-Weltz]. *Pragmatics Matters* (in press).

Nakamura, Y. (1992). Differences in N/NN teachers' evaluation of Japanese students' English speaking ability. *Cross Currents*, 19, 161-165.

Niezgoda, K., & Röver, C. (2001). Pragmatic and grammatical awareness: A function of the learning environment. In K. R. Rose & G. Kasper (Eds.), *Pragmatics and language teaching* (pp. 63-79). Cambridge: Cambridge University Press.

Nunn, R., & Lingley, D. (2004). Formative placement testing and its impact on ELT curriculum. *JACET Bulletin*, 39, 73-86.

Olshtain, E., & Weinbach, L. (1993). Interlanguage features of the speech act of complaining. In G. Kasper & S. Blum-Kulka (Eds.), *Interlanguage pragmatics* (pp. 108-122). Oxford: Oxford University Press.

Roberts, J. S., Laughlin, J. E., & Wedell, D. H. (1999). Validity issues in the Likert and Thurstone approaches to attitude measurement. *Educational and Psychological Measurement*, 59(2), 211-233.

Robinson, M. (1992). Introspective methodology in interlanguage pragmatics research. In G. Kasper (Ed.), *Pragmatics of Japanese as a native and target language* (Technical report 3) (pp. 27-82). Honolulu: University of Hawaii, Second Language Teaching & Curriculum Center.

Sawada, M. (2004). Adult EFL learner motivation: Learning English as lifelong learning. *JACET Bulletin*, 39, 59-71.

Schiffman, S. S., Reynolds, M. L., & Young, F. W. (1981). *Introduction to multidimensional scaling: Theory, methods, and applications.* New York: Academic Press.

Shibuya, A., & Mimura, C. (2002). Motivation and anxiety of Japanese female EFL students. *KOTESOL Proceedings 2002* (pp. 51-64). Seoul: KOTESOL.

Shimura, M., & Molloy, H. P. L. (2003, November). *The reality and realism of production questionnaire prompts*. Paper presented at the JALT National Meeting, Shizuoka, Japan.

Shimura, M., & Molloy, H. P. L. (2004, September 5). *How do complaint plans differ across cultures and languages?* Paper presented at the JACET National Meeting, Nagoya.

Silva, R. S. (2000). Pragmatics, bilingualism, and the native speaker. *Language & Communication*, 20, 161-178.

Smagorinsky, P. (2001). Rethinking protocol analysis from a cultural perspective. *Annual Review of Applied Linguistics*, 21, 233-245.

Taylor, K. L., & Dionne, J.-P. (2000). Accessing problem-solving strategy knowledge: The complementary use of concurrent verbal protocols and retrospective debriefing. *Journal of Educational Psychology*, 92, 413-425.

Turner, J. (1993). Using Likert scales in L2 research: Another researcher comments… *TESOL Quarterly*, 27(4), 736-739.

Upshur, J. A., & Turner, C. E. (1995). Constructing rating scales for second language tests. *ELT Journal*, 49, 3-12.

Viswesvaran, C., & Ones, D. S. (2000). Measurement errors in "Big Five Factors" personality assessment: Reliability generalization across studies and measures. *Educational and Psychological Measurement*, 60, 224-235.

Wharton, G. (2000). Language learning strategy use of bilingual foreign language learners in Singapore. *Language Learning*, 50, 203-243.

Wierzbicka, A. (1994). "Cultural scripts": A semantic approach to cultural analysis and cross-cultural communication. *Pragmatics and Language Learning*, 5, 1-24.

Williamson, J., Ranyard, R., & Cuthbert, L. (2000). A conversation-based process tracing method for use with naturalistic decisions: An evaluation study. *British Journal of Psychology*, 91, 203-221.

Wilson, S. R., & Kunkel, A. W. (2000). Identity implications of influence goals: Similarities in perceived face threats and facework across sex and close relationships. *Journal of Language and Social Psychology*, 19, 195-221.