2.4 'CLOSED-ENDED' QUESTIONNAIRE ITEMS

Let us start our exploration of the various types of questionnaire items by first examining the most frequent question type: *closed-ended* (or simply *'closed'*) *questions*. Although this category subsumes several very different item types, these all share in common the fact that the respondent is provided with ready-made response options to choose from, normally by encircling or ticking one of them or by putting an 'X' in the appropriate slot/box. That is, these items do not require the respondents to produce any free writing; instead, they are to choose one of the alternatives, regardless of whether their preferred answer is among them.

The major advantage of closed-ended questions is that their coding and tabulation is straightforward and leaves no room for rater subjectivity. Accordingly, these questions are sometimes referred to as 'objective' items. They are particularly suited for quantitative, statistical analyses (cf. Section 4.3) because the response options can easily be numerically coded and entered into a computer database.

2.4.1 Rating scales

Ratings scales are undoubtedly the most popular items in research questionnaires. They require the respondent to make an evaluative judgement of the target by marking one of a series of categories organized into a *scale*. (Note that the term 'scale' has, unfortunately, two meanings in measurement theory: one referring to a cluster of items measuring the same thing – cf. Section 2.3.2 on 'multi-item scales' – and the other, discussed in this section, referring to a measurement procedure utilizing an ordered series of response categories.) The various points on the continuum of the scale indicate different degrees of a certain category; this can be of a diverse nature, ranging from various attributes (e.g., frequency or quality) to intensity (e.g., very much → not at all) and opinion (e.g., strongly agree → strongly disagree). The points on the scale are subsequently assigned successive numbers, which makes their computer coding a simple task.

The big asset of rating scales is that they can be used for evaluating almost anything, and accordingly, as Aiken (1996) points out, these scales are second only to teacher-made achievement tests in the frequency of usage of all psychological measurement procedures. Indeed, I believe that few people in the teaching profession are unfamiliar with this item format: we are regularly asked to complete rating scales in various evaluation forms (of students, teachers, coursebooks, or courses), and outside the school context we also frequently come across them, for example when asked about our opinions of certain services (e.g., in hotels, transport).

Likert scales

The most commonly used scaling technique is the *Likert scale*, which has been named after its inventor, Rensis Likert. Over the past 70 years (Likert's original article came out in 1932) the number of research studies employing this technique has certainly reached a six-digit figure, which is due to the fact that the method is simple, versatile, and reliable.

Likert scales consist of a series of statements all of which are related to a particular target (which can be, among others, an individual person, a group of people, an institution, or a concept); respondents are asked to indicate the extent to which they agree or disagree with these items by marking (e.g., circling) one of the responses ranging from 'strongly agree' to 'strongly disagree.' For example:

Hungarians are genuinely nice people.

Strongly agree	Agree	Neither agree nor disagree	Disagree	Strongly disagree

After the scale has been administered, each response option is assigned a number for scoring purposes (e.g., 'strongly agree' = 5, 'strongly disagree' = 1). With negatively worded items the scores are usually reversed before analysis. Finally, the scores for the items addressing the same target are summed up or averaged. Thus, Likert scales are multi-item scales, following a 'summative model.'

The statements on Likert scales should be 'characteristic,' that is, expressing either a positive/favorable or a negative/unfavorable attitude toward the object of interest. Neutral items (e.g., *"I think Hungarians are all right"*) do not work well on a Likert scale because they do not evoke salient evaluative reactions, and extreme items are also to be avoided. An important concern of questionnaire designers is to decide the *number of steps* or response options each scale contains. Original Likert scales contained five response options (as just illustrated), but subsequent research has also used two-, three-, four-, six-, and seven-response options successfully. The most common step numbers have been five or six, which raises a second important questions: Shall we use an even or an odd number of steps?

Some researchers prefer using an even number of response options because of the concern that certain respondents might use the middle category ('neither agree nor disagree,' 'not sure,' or 'neutral') to avoid making a real choice, that is, to take the easy way out. Although according to research, this may be true of roughly 20% of the respondents, it appears that the inclusion or exclusion of a middle

category does not affect the *relative* proportions of those actually expressing opinions and thus does not modify the results significantly (Nunnally, 1978; Robson, 1993). My personal preference in the past has been to omit the 'undecided' category and to use a six-point scale such as the one illustrated in Sample 2.3 (on page 29).

The final question regarding Likert scales concerns the format of the respondents' answers: How do various physical appearances such as encircling options or ticking boxes compare to each other? Nunnally (1978) states that such variations appear to make little difference in the important psychometric properties of ratings as long as the layout of the questionnaire is clear and there are sufficient instructions and examples to orientate the respondents.

Likert scales have been used successfully with younger children as well; in such cases the number of the response options is often reduced to three and the options themselves are presented in a pictorial format instead of words. For example, in a three-point 'smilegram' children are asked to check the box under the face that best expresses how they feel toward a target:

Variations on Likert scales

Likert scales use response options representing the degree of agreement. This standard set of responses (i.e., strongly agree → strongly disagree) can be easily replaced by other descriptive terms that are relevant to the target. For example, Oxford's (1990) "Strategy Inventory in Language Learning" uses categories ranging from 'Never or almost never true of me' to 'Always or almost always true of me.' Or, in Dörnyei and Clément's (2001) "Language Orientation Question-

naire" a five-point scale ranging from "Not at all true" to "Absolutely true" has been used to assess attitudes toward language learning.

While these variations usually work well, we need to be careful about how to aggregate item scores to obtain multi-item scale scores. Likert scale items that measure the same attitude can simply be summed up because they refer to the same target and it is assumed that a higher total score reflects a stronger endorsement of the target attitude. However, not every variation on Likert scales is summative in the psychometric sense. For example, in Oxford's (1990) learning strategy inventory just mentioned, the various items within a group ask about the frequency of the use of different strategies. In this case, summing up the items would imply that the more strategies a person uses, the more developed his/her strategic skills are in the particular area. However, with regard to learning strategies this is *not* the case, since it is the *quality* rather than the quantity of the strategies a person utilizes that matters: One can be a very competent strategy user by consistently employing one single strategy that particularly suits his/her abilities and learning style. Thus, in this case, the summation of different item scores is not related linearly to the underlying trait.

Semantic differential scales

Instead of Likert scales we can also use *semantic differential scales* for certain measurement purposes. These are very useful in that by using them we can avoid writing statements (which is not always easy); instead, respondents are asked to indicate their answers by marking a continuum (with a tick or an 'X') between two bipolar adjectives on the extremes. For example:

Listening comprehension tasks are:

difficult ___:___:___:___:___:_X_:___ easy

useless ___:_X_:___:___:___:___:___ useful

These scales are based on the recognition that most adjectives have logical opposites and where an opposing adjective is not obviously available, one can easily be generated with 'in-' or 'un-' or by simply writing 'not ...'. Although the scope of semantic differential scales is more limited than that of Likert scales, the ease of their construction and the fact that the method is easily adaptable to study virtually any concept, activity, or person, may compensate for this. Oppenheim (1992) raises an interesting point concerning the content of semantic differential scales. He argues that it is possible and often useful to include adjective pairs that are seemingly inappropriate to the concept under consideration, such as masculine/feminine (with respect to a brand of cigarettes, for example), or rough/smooth (with respect to, say, Socialism): "By their more imaginative approach, such scales can be used to cover aspects that respondents can hardly put into words, though they do reflect an attitude or feeling" (p. 239). An additional bonus of semantic differential scales is that because they involve little reading, very little testing time is required.

Semantic differential scales are similar to Likert scales in that several items are used to evaluate the same target, and multi-item scores are computed by summing up the individual item scores. An important technical point concerning the construction of such bipolar scales is that the position of the 'negative' and 'positive' poles, if they can be designated as such, should be varied (i.e., the positive pole should alternate between being on the right and the left sides) to avoid superficial responding or a position response set (Aiken, 1996).

Semantic differential scales have been around for almost 50 years and during this time several factor analytic studies examined their content structure. The general conclusion is that there are three major factors of meaning involved in them:

- *evaluation*, referring to the overall positive meaning associated with the target (e.g., good-bad, wise-foolish, honest-dishonest);

- *potency*, referring to the target's overall strength or importance (e.g., strong-weak, hard-soft, useful-useless);

- *activity*, referring to the extent to which the target is associated with action (active-passive, tense-relaxed, quick-slow).

Scales are normally constructed to contain items focusing on each of the three dimensions; however, the items measuring the three evaluative aspects tend to correlate with each other.

Sample 2.4 Instructions for semantic differential scales

The following section of the questionnaire aims at finding out about your ideas and impressions about SOMETHING. In answering the questions we would like to ask you to rate these concepts on a number of scales. These all have pairs of opposites at each end, and between these there are 7 dashes. You are to place a check mark on one of the seven positions, indicating how you feel about the particular concept in view of the two poles. For example, if the scales refer to "listening comprehension tasks" and you find these rather useless and fairly easy, you can place your check marks as follows:

LISTENING COMPREHENSION TASKS ARE:

difficult ___:___:___:___:___:_X_:___ easy

useless ___:_X_:___:___:___:___:___ useful

In the following items please place your check marks rapidly and don't stop to think about each scale. We are interested in your immediate impression. Remember, this is not a test and there are no right or wrong answers. The "right" answer is the one that is true for you. Be sure to make only one check mark on each scale. Thank you!

Numerical rating scales

Teenagers sometimes play a rating game whereby they evaluate the appearance and 'sexiness' of the various girls/boys they see passing by in the street on a scale of 1-10. They would be surprised to hear that what they are doing is applying *numerical rating scales*. These scales involve 'giving so many marks out of so many,' that is, assigning one of several numbers corresponding to a series of ordered categories describing a feature of the target. The popularity of this scaling technique is due to the fact that the rating continuum can refer to a wide range of adjectives (e.g., excellent → poor; conscientious → slapdash) or adverbs (e.g., always → never); in fact, numerical ratings can easily be turned into semantic differential scales and vice versa. Sample 2.2 on page 28 provides an example.

True-false items

In some scales the designers only set two response options: true versus false (or 'yes' or 'no'), resulting in what is usually referred to as a *'true-false item.'* While generally it is true that the more options an item contains, the more accurate evaluation it yields, there might be cases when only such a polarized, yes-no decision can be considered reliable. For example, little children are sometimes seen as incapable of providing more elaborate ratings, and some personality test items also follow a true-false rating to ensure reliability in domains where the respondent may not be able to properly evaluate the degree to which a particular feature is present/true or not. In addition, with certain specific areas such as study habits, it may also be more appropriate to apply true/false items when the questions ask about occurrences of various behaviors in the past.

The key sentence (i.e., the one to be judged) in a good true-false item is relatively short and contains a single idea that is not subject to debate (i.e., it is either true or false). Due to the nature of the responses, the *acquiescence bias* (cf. Section 1.2.2) – that is, the tendency to respond in the affirmative direction when in doubt – may be a problem (Aiken, 1997). Because offering a polarized, black-and-

white judgment can often be perceived as too forced, some scales include a middle position, involving an 'undecided,' 'neutral,' or 'don't know' option.

2.4.2 Multiple-choice items

Language researchers will be very familiar with the multiple-choice item format because of its popularity in standardized L2 proficiency testing. The item type is also frequently used in questionnaires with respondents being asked to mark – depending on the question – one or more options. If none of the items apply, the respondent may have the option to leave the question unanswered, but because this makes it difficult to decide later whether the omission of a mark was a conscious decision or just an accident, it is better to include a *"Don't know"* and a *"Not applicable"* category (and sometimes even a *"No response"* option). Also, it is often desirable to ensure that an exhaustive list of categories is provided, and for this purpose it may be necessary to include an *"Other"* category, typically followed by an open-ended question of the *"Please specify"* sort (cf. Section 2.5.2).

Multiple choice items are relatively straightforward. It makes them more reader-friendly if we can make the response options shorter by including as much information in the stem as we can without repeating this every time. It also makes it easier to answer them if the response options have a natural order; otherwise they should be arranged in a random or alphabetical order. It is an obvious yet often violated rule that all options should be grammatically correct with respect to the stem. Finally, the use of negative expressions, such as "not," should be avoided in both the stem and the response options – a rule that generally applies to all question types (cf. Section 2.6.2).

Interestingly, multiple-choice items can also produce ordinal rather than nominal (categorical) data (cf. Section 4.3.4), that is, the various alternatives can represent *degrees* of an attitude, interest, and belief. Respondents are, then, instructed to choose only one of these options and their answers will be coded according to the value of the particular option they chose: e.g., Option A may be assigned '2' and Option D '3'. Obviously the value of each option cannot be set in ad-

vance on a purely theoretical basis but can only be deduced from extensive pilot testing (cf. Section 2.9) whereby the items are administered to a group of respondents and the value of each response option is calculated on the basis of their answers (for examples of such 'graded' multiple choice items, see Sample 2.5 below).

Sample 2.5. Multiple-choice attitude items from the 'Attitude/Motivation Test Battery' (Gardner, 1985, p. 181)

Scoring Key	
	During French class, I would like:
2	(a) to have a combination of French and English spoken.
1	(b) to have as much English as possible spoken.
3	(c) to have only French spoken.
	If there were a French Club in my school, I would:
2	(a) attend meetings once in a while.
3	(b) be most interested in joining.
1	(c) definitely not join.

2.4.3 Rank order items

It is a common human mental activity to rank order people, objects, or even abstract concepts, according to some criterion, and *rank order items* in questionnaires capitalize on our familiarity with this process. As the name suggests, these items contain some sort of a list and respondents are asked to order the items by assigning a number to them

according to their preferences. Wilson and McClean (1994) warn us that it may be very demanding to arrange items in order of importance whenever there are more than five ranks requested, and it has also been found, more generally, that rank order items impose a more difficult task on the respondent than single-response items. Furthermore, unlike in a rating scale in which a person can assign the same value to several items (e.g., one can mark 'strongly agree' in all the items in a multi-item scale), in rank order items each sub-component must have a different value even though such a forced choice may not be natural in every case.

In my own research, I have tended to avoid rank order items because it is not easy to process them statistically. We cannot simply count the mean of the ranks for each item across the sample because the numerical values assigned to the items are not the same as in rating scales: they are only an easy technical method to indicate *order* rather than the *extent* of endorsement. That is, if something is ranked third, the value '3' does not necessarily mean that the degree of one's attitude is 3 out of, say, 5 (which would be the case in a Likert scale); it only means that the particular target's relevance/importance is, in the respondent's estimation, somewhere between the things ranked second and fourth; the actual value can be very near to the second and miles away from the forth or vice versa. To illustrate this, let us take a short list of items that we may need for travelling abroad:

- passport
- credit card
- tickets
- plumbing manual.

'Plumbing manual' would probably be ranked by everybody as the least necessary item in the list but by assigning a value of '4' or '1' to it (depending on which end we start counting from) its value would be only one less (or more) than the next one is the list, whereas in reality its value for travelling purposes is next to zero (unless you are a plumber...).

2.4.4 Numeric items

One item type that is seemingly open-ended but is, in effect, closed-ended can be labeled as a *numeric item*. These items ask for a specific numeric value, such as the respondent's age in years, or the number of foreign languages spoken by a person. What makes these items similar to closed questions is that we can anticipate the range of the possible answers and the respondent's task is to specify a particular value within the anticipated range. We could, in fact, list, for example for the 'age' item, all the possible numbers (e.g., between 5 and 100) for the respondent to choose from (in a multiple-choice fashion) but this would not be space-economical. However, computerized, on-line questionnaires often do provide these options in a pull-down menu for the respondent to click on the selected answer.

2.4.5 Checklists

Checklists are similar to rank order items in that they consist of a list of descriptive terms, attributes, or even objects, and respondents are instructed to mark the items on the list that apply to the particular question. For example, students might be asked to mark all the adjectives in a list of personality characteristics that describe their teacher. This evaluation would, then, yield a score for the teacher on each characteristic, indicating how many raters checked the particular adjective; that is, the person's score on each item can be set equal to the number of judges who checked it. In the teacher's case, a score of '0' on the 'fairness' item would mean that nobody thinks that the teacher is fair (which would be problematic). Because – unless otherwise instructed – different respondents may check a different number of items (e.g., someone may check almost all the adjectives, whereas another rater might check only one), this response set can have a pronounced effect on the scores and therefore some sort of grouping or statistical control is frequently used (Aiken, 1996).

2.5 OPEN-ENDED QUESTIONS

Open-ended questions include items where the actual question is not followed by response options for the respondent to choose from but rather by some blank space (e.g., dotted lines) for the respondent to fill. As we have seen in the previous chapter (in Section 1.3), questionnaires are not particularly suited for truly qualitative, exploratory research. Accordingly, they tend to have few open-ended questions and even the ones included are relatively short, with their 'openness' somehow restricted. Questionnaires are not the right place for essay questions.

In spite of this inherent limitation of the questionnaire as a research instrument (namely that due to the relatively short and superficial engagement of the respondents it cannot aim at more than obtaining a superficial, "thin" description of the target) open-ended questions still have merits. Although we cannot expect any soul-searching self-disclosure in the responses, by permitting greater freedom of expression, open-format items can provide a far greater "richness" than fully quantitative data. The open responses can offer graphic examples, illustrative quotes, and can also lead us to identify issues not previously anticipated. Furthermore, sometimes we need open-ended items for the simple reason that we do not know the range of possible answers and therefore cannot provide pre-prepared response categories. Oppenheim (1992) also points out that in some cases there may actually be good reasons for asking the same question both in an open and closed form.

The other side of the coin is that open-ended questions have certain serious disadvantages, most notably the following two:

- They take up precious 'respondent-availability time' and thus restrict the range of topics the questionnaire can contain.
- They are difficult to code in a reliable manner.

Because of these considerations, professional questionnaires tend not to include any real open-ended items; yet, my recommendation is that it might be worth experimenting with including some. Research-

ers agree that truly open questions (i.e., the ones that require quite a bit of writing) should be placed at the end rather than at the beginning of the questionnaire. In this way, they are not answered at the expense of the closed items: they do not discourage people from completing the questionnaire and do not prevent those who get bogged down with them from answering the other questions.

In my experience, open-ended questions work particularly well if they are not completely open but contain certain guidance. In the following we will look at four techniques to provide such guidance.

2.5.1 Specific open questions

Specific open questions ask about concrete pieces of information, such as facts about the respondent, past activities, or preferences (e.g., *Which is your favorite television program/weekend activity? What languages have you studied in the past?*). They can normally be answered in one line, which is usually explicitly marked on the questionnaire (e.g., with dots). The answers can sometimes be followed up with a 'Why?' question.

2.5.2 Clarification questions

Certain answers may be potentially so important that it is worth attaching a clarification question to them, for example in a 'routed' form:

```
If you rated the coursebook you are using as
"poor" or "very poor," please briefly explain
why. Write your answer here:

_____ _____

_____ _____
```

Clarification questions are also appropriate when there is an "Other" category in a multiple-choice item. Typically, "Please specify" is used and some space is left for the respondent to provide a statement.

2.5.3 Sentence completion items

A simple question is often less effective in eliciting a meaningful answer than an unfinished sentence beginning that the respondent needs to complete. I have successfully used this technique on various feedback forms in particular. A good completion item should be worded so that it directs the respondent's attention to a well-defined issue/area. Sometimes respondents are asked not to 'agonize' over the answers but jot down the first thing that comes to mind. For example:

```
One thing I liked about this activity is ______
_____________________________________
One thing I didn't like about this activity is
_____________________________________
I found this activity ___________________
_____________________________________
```

2.5.4 Short-answer questions

The term *'short-answer questions'* is sometimes used to distinguish these questions from 'essay questions' (which are not recommended in ordinary questionnaires and therefore will not be discussed). Short-answer questions involve a real exploratory enquiry about an issue;

that is, they require a more free-ranging and unpredictable response. As Gillham (2000, pp. 34-35) concludes, these questions:

> can be motivating for the respondent, and they enable the researcher to trawl for the unknown and the unexpected. One or two questions of this type can be a good way of finishing a questionnaire, which can otherwise easily leave respondents with the impression that their personal opinions or experiences have to fit the straitjacket of prescribed answers.

Gillham even recommends the inclusion of a completely open concluding question, such as, *"We have tried to make this questionnaire as comprehensive as possible but you may feel that there are things we have missed out. Please write what you think below, using an extra page if necessary"* (pp. 34-35).

Good short-answer questions are worded in such a focused way that the question can be answered succinctly, with a 'short answer' – this is usually more than a phrase and less than a paragraph (and certainly no more than two paragraphs). That is, short-answer questions do not ask about things in general, but deal with only one concept or idea. For example, rather than asking, *"What did you like about the workshop?"* it might be better to narrow down the question by asking, *"What was it you found most useful about the workshop?"*

One type of questionnaire that is almost always concluded by a few open-ended questions is college forms for students to evaluate their teachers/courses. A typical final sequence of questions is as follows: *What were the most effective aspects of this course? What were the least effective aspects of this course? How could this course be further improved?*