

2 Constructing your own psychometric questionnaire

When you see a questionnaire in a popular magazine or newspaper, it is often no more than a series of items that are not necessarily related to each other and that are scored and interpreted individually. This chapter is a guide to the construction of psychometric questionnaires, where items can be combined to produce an overall scale.

Questionnaires are used to measure a wide variety of attributes and characteristics. The most common examples are knowledge-based questionnaires—i.e., questionnaires of ability, aptitude, and achievement—and person-based questionnaires, i.e., questionnaires of personality, clinical symptoms, mood, and attitudes. Whatever type of questionnaire you wish to develop, this guide will take you through the main stages of construction and show you how to tailor your questionnaire to its particular purpose. Throughout the guide, the construction of the Golombok Rust Inventory of Marital State (GRIMS; Rust et al. 1988) will be described (in italics) as a practical example.

The purpose of the questionnaire

The first step in developing a questionnaire is to ask yourself: “What is it for?” Unless you have a clear and precise answer to this question, your questionnaire will not tell you what you want to know.

With the GRIMS, we wanted to develop a questionnaire to assess the quality of the relationship of couples who are married or living together. We intended that the GRIMS would be of use in research, to help therapists or counselors either evaluate the effectiveness of therapy for couples with relationship problems or investigate the impact of social, psychological, medical, or other factors on a relationship. In addition, we hoped that it would be used clinically as a quick and easy-to-administer technique for identifying the severity of a problem, for finding out which partner perceives a problem in the relationship, and for identifying any improvement or lack of improvement in either partner or both partners over time.

Write down clearly and precisely the purpose of your questionnaire.

Making a blueprint

A blueprint, sometimes known as the test specification, is a framework for developing the questionnaire. A grid structure is generally used with content areas along the horizontal axis and manifestations (ways in which the content areas may manifest) along the vertical axis (see Table 2.1). For practical reasons, either four or five categories are usually employed along each axis. Fewer often results in too narrow a questionnaire, and more can be too cumbersome to deal with.

Table 2.1 A test blueprint with four content areas and four manifestations

Content Areas		A	B	C	D
Manifestations	A				
	B				
	C				
	D				

Content areas

A clear purpose will enable you to specify the content of your questionnaire. The content areas should cover everything that is relevant to the purpose of the questionnaire.

The many different ideas about what constitutes a good or bad relationship posed a problem when we tried to specify the content areas of the GRIMS. For this reason, we used the expertise of relationship therapists/counselors and their clients. The therapists/counselors were asked to identify areas that they believed to be important in marital harmony as well as the areas that they would assess during initial interviews. Information from clients was obtained by asking them to identify their targets for change. The views of these experts were collated to provide the following content areas that were generally considered to be important for assessing the state of a relationship: (i) interests shared (work, politics, friends, etc.) and degree of dependence and independence; (ii) communication (verbal and nonverbal); (iii) gender; (iv) warmth, love, and hostility; (v) trust and respect; (vi) roles, expectations, and goals; (vii) decision-making; and (viii) coping with problems and crises.

Write down the content areas to be covered by your questionnaire. If these are not clear-cut, consult experts in the field.

Manifestations

The ways in which the content areas may manifest themselves will vary according to the type of questionnaire under construction. For example, questionnaires designed to measure educational attainment may use Bloom’s (1956) taxonomy of educational objectives to tap into different forms of knowledge. For questionnaires that are more psychological in nature, behavioral, cognitive, and affective manifestations of the content areas may be more appropriate. For personality questionnaires, you will need to balance socially desirable and socially undesirable aspects of the trait, as well as for acquiescence. The latter is achieved by allowing half of the items to manifest positively (e.g., “I am outgoing” in an extraversion scale) and half to manifest negatively (e.g., “I am shy” in an extraversion scale). In specifying manifestations, it is important to ensure that different aspects of the content areas will be elicited.

In constructing the GRIMS, we again took account of the experts’ information to obtain the following manifestations: (i) beliefs about, insight into, and understanding of the nature of dyadic relationships; (ii) behavior within the actual relationship; (iii) attitudes and feelings about relationships; (iv) motivation for change, understanding the possibility of change, and commitment to a future together; and (v) extent of agreement within the couple.

As you can see from the GRIMS blueprint, what is described as a content area and what is described as a manifestation may not always be clear-cut.

Write down ways in which the content areas of your questionnaire may become manifest.

You will now be able to construct your blueprint. The number of cells will be the number of content areas multiplied by the number of manifestations. Between 16 and 25 cells (i.e., 4×4 , 4×5 , 5×4 , or 5×5) is generally considered ideal for sufficient breadth while maintaining manageability.

Draw your blueprint, labeling each content area (column) and each manifestation (row).

Each cell in the blueprint represents the interaction of a content area with a manifestation of that content area. By writing items for your questionnaire that correspond to each cell of the blueprint, you will ensure that all aspects that are relevant to the purpose of your questionnaire are covered.

A decision that has to be made when designing the blueprint is whether to give different weightings to the cells, i.e., whether to write more items for some cells than for others. This will depend on whether or not you feel that some content areas or manifestations are more important than others. In the blueprint in Table 2.2, it has been decided that content area A should receive a weighting of 40%, content area B a weighting of 40%, content area C a weighting of 10%, and content area D a weighting of 10%. For the manifestations, a weighting of 25% has been allocated to each.

For the GRIMS, equal weightings were assigned to each cell, as we had no reason to believe that any of the content areas or manifestations were more important than the others.

Assign percentages to each content area of your blueprint so that the total of the percentages across the content areas adds up to 100%.

Assign percentages to each manifestation in your blueprint so that the total of the percentages down the manifestations adds up to 100%.

Insert these percentages into your blueprint.

Assigning weightings will tell you what proportion of all items in the questionnaire should be written for each cell. The next step is to decide on the total number of items to include. You must consider factors such as the size of your blueprint (a large blueprint with many content areas and manifestations will need a greater number of items than a small one) and the amount of time available for administering the questionnaire. There is no point in asking people with little time to spare to complete a lengthy inventory, as the quality of their responses will be poor and items may be omitted. The characteristics of the respondents are also important. Children and people who are older or have a physical or mental illness may be slow and unable to maintain concentration. Although it is

Table 2.2 Assignment of percentages of items to content areas (columns) and manifestations (rows)

Content Areas					
		A (40%)	B (40%)	C (10%)	D (10%)
Manifestations	A (25%)				
	B (25%)				
	C (25%)				
	D (25%)				

important to include a sufficient number of items to ensure high reliability, compliance among respondents is crucial, and a balance must be struck between the two. A minimum of 12 items per scale is usually required to achieve adequate reliability. In the plan, however, a minimum of 20 items should be aimed for, and a fairly straightforward questionnaire of this length should take an average respondent no longer than six minutes to complete. As it is necessary to construct a pilot version of your questionnaire in the first instance, you must remember to allow for at least 50% more items in the blueprint than you intend to include in the final version.

The GRIMS was intended as a short questionnaire for use with both distressed and non-distressed couples. As we hoped to achieve a final scale of about 30 items, we planned a pilot version with 100 items.

Decide on how many items to include in the pilot version of your questionnaire by taking into account the desired number of items in the final version, the size of your blueprint, the time available for testing, and the characteristics of the respondents.

Once you have assigned weightings to the cells and decided on the total number of items that you require for your pilot questionnaire, you will be able to work out how many items to write for each cell. The blueprint in Table 2.3, with given weightings, shows the number of items that have to be written for each cell to obtain a pilot questionnaire with 80 items. The first step is to work out the total number of items for each content area and for each manifestation. The blueprint specifies that 40% of the items (32 items) should be on content area A, 40% on content area B (32 items), 10% (eight items) on content area C, and 10% (eight items) on content area D. These numbers are entered into the bottom row of the blueprint. Similarly, the blueprint specifies that 25% of the items (20 items) should concern each of the manifestations, and this is entered into the right-hand column of the blueprint. To calculate the number of items in each cell of the blueprint, multiply the total number of items in a content area by the percentage assigned to the manifestation in each row. For example, the number of items for the top left-hand cell (content area A/manifestation A) is 25% of 32 items, which is eight items. The number of items to be written for each cell is calculated in the same way. If you do not obtain an exact number of items for a cell, approximate to the number above or below while trying to maintain the same total number of items as you had originally planned.

The 100 items in the equally weighted 40-cell GRIMS blueprint allowed two or three items per cell.

Enter the number of items to be written for each cell into your blueprint.

Table 2.3 Assignment of number of items per cell, per content area, and per manifestation to a test blueprint

		Content Areas				No. of Items
		A (40%)	B (40%)	C (10%)	D (10%)	
Manifestations	A (25%)	8	8	2	2	20
	B (25%)	8	8	2	2	20
	C (25%)	8	8	2	2	20
	D (25%)	8	8	2	2	20
No. of Items		32	32	8	8	80

Writing items

There are several types of items that are used in questionnaires, the most common of which are alternate-choice, multiple-choice, and rating-scale items. Different item types are suitable for different purposes, and consideration of the attribute or characteristic that you wish your questionnaire to measure will guide you toward an appropriate choice.

Alternate-choice items

An item for which the respondent is given two choices from which to select a response, e.g., true or false, yes or no. Most commonly used in knowledge-based questionnaires, e.g., “Bogota is the capital of Colombia: true or false?” Sometimes used in personality questionnaires, e.g., “I never use a lucky charm: yes or no?” Generally considered inappropriate for clinical-symptom, mood, or attitude questionnaires—but used occasionally.

Advantages

Good for assessing knowledge of facts and comprehension of materials presented in the question. Fast and easy to use.

Disadvantages

For ability, aptitude, and achievement items, the correct response is often not clear-cut, i.e., completely true or completely false. Another problem is that the respondent has a 50% chance of obtaining the correct response by guessing. For personality, clinical-symptom, mood, and attitude questionnaires, there are no right or wrong answers. However, respondents often consider the narrow range of possible responses to be too restrictive.

Multiple-choice items

An item for which the respondent is given more than two choices from which to select a response. It consists of two parts: (i) the stem—a statement or question that contains the problem; and (ii) the options—a list of possible responses, of which one is correct or the best and the others are distractors. Often four or five possible responses are used to reduce the probability of guessing the answer. The most widely used item type in knowledge-based questionnaires, for example:

What is the capital of Colombia?

- A. La Paz
- B. Bogotá
- C. Lima
- D. Santiago

Not used in person-based questionnaires.

Advantages

Well suited to the wide variety of materials that may be presented in ability, aptitude, and achievement questionnaires. Challenging items that are easy to administer and score can

be constructed. The effects of guessing are also reduced with multiple-choice items. For example, an item with five options gives someone a 20% chance of guessing the correct answer, compared with 50% in alternate-choice items.

Disadvantages

Time and skill are needed to write good multiple-choice items. A common problem is that not all of the options are effective, i.e., they are so unlikely to be correct that they do not function as possible options. This can reduce what is intended as a five-choice item to a three- or four-choice item, or even to an alternate-choice item.

Rating-scale items

An item for which the possible responses lie along a continuum, e.g., yes, don't know, no; true, uncertain, false; strongly disagree, disagree, agree, strongly agree; always, sometimes, occasionally, hardly ever, never. Up to seven options are generally used, as it is difficult for respondents to differentiate meaningfully among more than that number. Although rating-scale items are similar to multiple-choice items in giving several response options, the options in rating scales are ranked, whereas multiple-choice item options are independent of each other. Not used in knowledge-based questionnaires. The most widely used item type in person-based questionnaires, for example:

I am not a superstitious person

- A. strongly disagree
- B. disagree
- C. agree
- D. strongly agree

Advantages

Respondents feel able to express themselves more precisely with rating-scale items than with alternate-choice items.

Disadvantages

Respondents differ in their interpretations of the response options, e.g., “frequently” has a different meaning to different individuals. Some respondents tend to always choose the most extreme options. When an uneven number of response options is used, many respondents tend to choose the middle one, e.g., “don't know” or “occasionally.”

The type of options should be chosen to suit the material to be presented in the questionnaire. There are no fixed rules about which type of options is best. A personality or mood questionnaire might require responses in terms of the options “not at all,” “somewhat,” and “very much.” Attitude questionnaires generally consist of statements about an attitude object followed by the options “strongly agree,” “agree,” “uncertain,” “disagree,” and “strongly disagree.” For clinical-symptom questionnaires, you might find that options relating to the frequency of occurrence—such as “always,” “sometimes,” “occasionally,” “hardly ever,” and “never”—are the most suitable.

The most appropriate number of options to choose from will also depend on the nature of the questionnaire. It is important to provide a sufficient number for respondents to feel able to express themselves adequately while ensuring that there are not so many that they have to make meaningless discriminations. In questionnaires using rating-scale items, where strength of response should be reflected in the respondent's score, it is common for at least four options to be used.

It is sometimes necessary to use different types of items in a questionnaire because of the nature of the material to be included. However, it is preferable to use only one item type wherever possible, to produce a neatly presented questionnaire.

Rating-scale items are the most appropriate for a scale of relationship state. The GRIMS items are presented as statements to which the respondents are asked to respond with "strongly agree," "agree," "disagree," or "strongly disagree." This spread of options allows strength of feeling to affect scores. The items are forced choice, i.e., there is no "don't know" category.

Decide which item type is most appropriate for your questionnaire. In general, multiple-choice items are best for knowledge-based questionnaires, and rating-scale items are best for person-based questionnaires unless you have good reason, such as speed or simplicity, for choosing alternate-choice items. A good method for deciding which to choose is to try to construct items of each type using different options. The most appropriate choice for your questionnaire will soon become clear.

Before beginning to write items for your questionnaire, read the following summary of important points to remember. For a more detailed discussion of how to write good items, see Thorndike and Thorndike-Christ (2014).

All questionnaires

Make sure that your items match your blueprint. The allocation of items to specific cells may become a bit fuzzy, as some items may be appropriate for more than one cell. If you find that some cells are inappropriate and you decide to omit them, do not do so without proper consideration. Remember, however, that the blueprint is a guide and not a straitjacket.

Write each item clearly and simply. Avoid irrelevant material, and keep the options as short as possible. Each item should ask only one question or make only one statement. Where possible, avoid subjective words such as "frequently," as these may be interpreted differently by different respondents. It is also important that all options function as feasible responses, i.e., that none be clearly wrong or irrelevant and therefore unlikely to be chosen.

After writing your items, read them again a few days later. Also, ask a colleague to look at them to ensure that they are easily understood and unambiguous.

Knowledge-based questionnaires

Make sure that alternate-choice items can undoubtedly be classified as true or false; otherwise some respondents will think of exceptions to the rule.

For multiple-choice items, ensure that each item has only one correct or best response. Ideally, each distractor option should be used equally by respondents who do not choose the correct response. Remember that the more similar the options, the more difficult the item.

Person-based questionnaires

Sometimes respondents will complete a questionnaire in a certain way irrespective of the content of the items.

Acquiescence

Acquiescence is the tendency to agree with items regardless of their content. This can be reduced by ensuring that an equal or almost equal number of items is scored in each direction. To do this, it is usually necessary to reverse some of the items. For example, the item “I am satisfied with our relationship” can be reversed to “I am dissatisfied with our relationship.” When reversing items, it is important to check that the reversed item really does mean the opposite of the original item. It is best to avoid double-negative statements, as these cause confusion. Acquiescence is less likely to occur with items that are clear, unambiguous, and specific.

Social desirability

Social desirability is the tendency to respond to an item in a socially acceptable manner. This can be reduced by excluding items that are clearly socially desirable or undesirable. If this is unavoidable due to the nature of your questionnaire, try to ask the question indirectly to evoke a response that is not simply a reflection of how the respondent wishes to present themselves. For example, an item to measure paranoia may be subtly phrased as “there are some people whom I trust completely” rather than “people are plotting against me.” Social desirability can also be reduced by asking respondents to give an immediate response rather than a careful consideration of each item.

Indecisiveness

Indecisiveness is the tendency to use the “don’t know” or “uncertain” option. This is a common problem that can easily be eliminated by omitting the middle category. It is advisable to do so unless respondents are likely to become irritated by items that they feel are unanswerable.

Extreme response

Extreme response is the tendency to choose an extreme option regardless of direction. Some respondents will use one direction for a series of items and then switch to the other direction, and so on. Again, this can be reduced by the use of clear, unambiguous, and specific items.

It is important to bear in mind these habitual ways of responding when writing items. However, a careful item analysis will eliminate items that are biased toward a particular response.

Examples of GRIMS items:

“We both seem to like the same things” was written for the blueprint cell representing content area (i) and manifestation (ii).

“I wish there was more warmth and affection between us” was written for the blueprint cell representing content area (iv) and manifestation (iv).

Write each of your items on a small card so that you can easily make changes to wording and ordering. To order the items for your questionnaire, pick an interesting and unthreatening item to start with, and then shuffle the cards to randomize the rest. Make adjustments if too many similar-looking items occur together. For knowledge-based questionnaires that have items of increasing difficulty, order the items from easy to hard.

Designing the questionnaire

Good design is crucial for producing a reliable and valid questionnaire. Respondents feel less intimidated by a questionnaire that has a clear layout and is easy to understand, and take their task of completing the questionnaire more seriously.

Background information

Include headings and sufficient space for the respondent to fill out their name, age, gender, or whatever other background information you require. It is often useful to obtain the date on which the questionnaire is completed, especially if it is to be administered again.

Instructions

The instructions must be clear and unambiguous. They should tell the respondent how to choose a response and how to indicate the chosen response in the questionnaire. Other relevant instructions should be given, e.g., respond as quickly as possible, respond to every item, or respond as honestly as possible. Information that is likely to increase compliance—e.g., regarding confidentiality—should be stressed.

Sample instructions for a knowledge-based multiple-choice questionnaire:

INSTRUCTIONS: Each item is followed by a choice of possible responses: A, B, C, D, or E. Read each item carefully and decide which choice best answers the question. Indicate your answer by circling the letter responding to your choice. Your score will be the number of correct answers, so respond to each question even if you are unsure of the correct answer.

Sample instructions for a person-based rating-scale questionnaire:

INSTRUCTIONS: Each statement is followed by a series of possible responses: strongly disagree, disagree, agree, or strongly agree. Read each statement carefully and decide which response best describes how you feel. Then put a tick over the corresponding response. Please respond to every statement. If you are not completely sure which response is most accurate, tick the response that you feel is most appropriate. Do not spend too long on each statement. It is important that you answer each question as honestly as possible. **ALL INFORMATION WILL BE TREATED WITH THE STRICTEST CONFIDENCE.**

Layout

The following tips will help you to arrange items on the page so that they are easy to read:

- (a) Number each item.
- (b) Keep each line short, with no more than 10 or 12 words per line.
- (c) Ensure that the items produce a straight vertical margin down the left-hand side of the page.
- (d) Arrange the response options to produce a straight vertical margin down the right-hand side of the page. Insert headings at the top and symbols next to each item. There should be a clear visual relationship between each item and its response options. This can be done by inserting a dotted line from the item stem to its response option.

	STRONGLY DISAGREE	DISAGREE	AGREE	STRONGLY AGREE
1. _____	SD	D	A	SA
2. _____	SD	D	A	SA
3. _____	SD	D	A	SA
4. _____	SD	D	A	SA
5. _____	SD	D	A	SA

- (e) Separate each item with a space rather than a horizontal line. If your items, instructions, and background information all fit on one page, that is good. However, it is better to produce a neat two- or three-page questionnaire than one page that looks cramped.
- (f) If using more than one type of item, group similar items together. Each type will need different instructions and response options.
- (g) Remember that different PCs, laptops, and smartphones have different layouts, and it is important for your questionnaire to look good on all of them. Having the questionnaire printed using a high-quality printer is a sensible practice in case anyone still wants to use pencil-and-paper administration. Ensure that whatever the medium, the type is large enough to be read easily. Use your computing skills creatively to plan the layout. Experiment with different fonts, colors, type sizes, and spacings to see which look best.
- (h) You can use design as a tool to portray or disguise the purpose of your questionnaire. For example, small, closely set type can make a questionnaire look very formal, while larger type with items spaced well apart on colored paper is friendlier. Design can set an atmosphere, so use it!

The GRIMS was designed with simplicity of administration in mind. The respondent must answer 28 questions on one scrollable page, with the same response options for each question. This makes it quick and uncomplicated to complete.

Try different layouts of your questionnaire using different media until the arrangement looks logical. Then experiment with font, color, type size, spacing, and number of pages to see what looks best.

To score your questionnaire, allocate a score to each response option, and then add up the scores for all the items to give a total score for the questionnaire.

For knowledge-based questionnaires, it is common to give the correct or best option for each item a score of 1 and the distractor options a score of 0. The higher the total score, the better the performance.

For person-based questionnaires, scores should be allocated to response options according to a continuous scale, e.g., always = 5, usually = 4, occasionally = 3, hardly ever = 2, never = 1; yes = 2, uncertain = 1, no = 0; true = 1, false = 0. For reversed items, it is necessary to reverse the scoring (e.g., always = 1, usually = 2, occasionally = 3, hardly ever = 4, never = 5), so that each item is scored in the right direction. After reversing the scores for reversed items, add up the scores for all items to obtain the total score for the questionnaire. Depending on the way in which you have allocated scores to response options, the higher the total score, the greater or lesser the presence of the characteristic being measured.

A scoring key that fits over the questionnaire to identify which option the respondent has chosen for each item and its score can be useful for quick and easy scoring. In the following example, the respondent has obtained a total score of 12 (2 + 2 + 3 + 5):

	ALWAYS	USUALLY	OCCASIONALLY	HARDLY EVER	NEVER
1. _____	A(5)	U(4)	O(3)	HE(2)	N(1)
2. (reversed item)_____	A(1)	U(2)	O(3)	HE(4)	N(5)
3. _____	A(5)	U(4)	O(3)	HE(2)	N(1)
4. (reversed item)_____	A(1)	U(2)	O(3)	HE(4)	N(5)

Simple scripts can also be written for scoring questionnaires, although it is good practice to keep a backup of the item-level data in case you ever decide to change the scoring process.

Piloting the questionnaire

The next stage in constructing your questionnaire is the pilot study. This involves having the questionnaire completed by people who are similar to those for whom the questionnaire is intended. Analysis of these data will help you to select the best items for the final version of your questionnaire.

If, for example, your questionnaire is intended for women with preschool-age children, you might carry out the pilot study at a baby clinic or a mothers-and-toddlers club. If it is for use with the general population, you would need to find a group of people who are representative of the population at large. This is often more difficult than finding a more specific group. You could make use of the electoral register, but this is usually too

time-consuming to be worthwhile for a pilot study. When a truly representative group is impossible to find, an approximation is usually good enough. A common strategy is to hand out questionnaires in public places such as shopping centers, train and bus stations, airport lounges, doctors' waiting rooms, or cafeterias of large organizations. The respondents who take part in the pilot study should vary in terms of demographic characteristics such as age, gender, and social class. There is little point in piloting a questionnaire intended for any gender only with men, or a questionnaire to be used throughout an industry only with managers and not manual workers. It is important to obtain relevant demographic information from the respondents in the pilot study to help with the validation of your questionnaire at a later stage.

The pilot version of your questionnaire should be administered to as many people as possible. The minimum number of respondents required is one more than the number of items. If it is not possible to obtain this many, it is better to use fewer people than to omit the piloting stage altogether.

The pilot version of the GRIMS was administered to both partners in 60 client couples from relationship therapy and relationship guidance clinics throughout the country.

Administer your questionnaire and obtain relevant demographic information from a group of people who are similar to those for whom the final questionnaire is intended.

Item analysis

Item analysis of the data collected in the pilot study to select the best items for the final version of your questionnaire involves an examination of the *facility* and the *discrimination* of each item. For knowledge-based multiple-choice items, it is also important to look at *distractors*.

The first step is to create an item-analysis table with each column (a, b, c, d, e, etc.) representing an item and each row (1, 2, 3, 4, 5, etc.) representing a respondent. For knowledge-based items, insert "1" in each cell for which the respondent gave the correct answer and "0" for each incorrect answer. Add up the scores to give total scores for each row (i.e., each respondent) and each column (i.e., each item).

Table 2.4 shows a sample item-analysis table for a knowledge-based questionnaire.

Facility

Most questionnaires are designed to differentiate between respondents according to whatever knowledge or characteristic is being measured (see discussion of standardization in Chapter 3). A good item, therefore, is one for which different respondents give different responses. The facility index gives an indication of the extent to which all respondents answer an item in the same way. If they do, then these items are redundant, and it is important to get rid of them. For example, if every respondent gives the correct response to a particular item, this simply has the effect of adding one point to the total score for each respondent and does not discriminate among them.

For knowledge-based questionnaires, the facility index is calculated by dividing the number of respondents who obtain the correct response for an item by the total number of respondents. Ideally, the facility index for each item should lie between .25 and .75, averaging .5 for the entire questionnaire. A facility index of less than .25 indicates that the item is too difficult, as very few respondents obtain the correct response; and a facility

Table 2.4 shows a sample item-analysis table for a knowledge-based questionnaire

		<i>Items</i>					
Respondents		a	b	c	d	e	Sum
	1	1	1	0	1	1	4
	2	0	1	0	0	1	2
	3	1	0	0	1	1	4
	4	1	0	0	0	1	2
	5	1	0	0	1	1	3
	Sum	3	2	0	3	5	
	Facility	.8	.4	0.0	.6	1.0	
	Ddiscrimination	.13	-.48	UNDEF	.67	UNDEF	

index of more than .75 shows that the item is too easy, as most respondents obtain the correct response. In Table 2.4, we would want to eliminate items c and e from the final questionnaire, as everyone has responded to these items in the same way.

If it is a person-based questionnaire, then items may have values of more than 1. For example, if the response options for each item are “strongly agree,” “agree,” “disagree,” and “strongly disagree,” then the item values may be 1, 2, 3, or 4. Insert the actual score for each item into the item-analysis table, remembering to ensure that reversed items are scored in the opposite direction to nonreversed items. The facility index for person-based items is calculated by summing the scores for the item for each respondent, then dividing this total by the total number of respondents. An item with a facility index that is equal to or approaching either of the extreme scores for the item should not be included in the final version of the questionnaire. It is also important to ensure by looking at the scores in the item-analysis table that a good facility index—i.e., one lying somewhere between the extreme scores—does not simply mean that everyone has chosen the middle option.

Discrimination

This is the ability of each item to discriminate among respondents according to whatever the questionnaire is measuring, i.e., respondents who perform well on a knowledge-based questionnaire or who exhibit the characteristic being measured by a person-based questionnaire should respond to each item in a particular way. Items should be selected for the final version of the questionnaire only if they measure the same knowledge or characteristic as the other items in the questionnaire. In a knowledge-based questionnaire, this means that for each and every item, those with higher total scores on the questionnaire should be more likely to get the item correct than those with lower scores on the questionnaire. Items that fail to do this are said not to discriminate between high and low scorers, and hence should be removed.

More usually, discrimination is measured by correlating each item with the total score from summing all the other items in the questionnaire (i.e., the total score with the item that was removed). You can use a spreadsheet program such as Excel to do this, although any statistical analysis package will do. In Table 2.4, the Pearson product-moment correlation coefficient was used (CORREL in Excel), but some prefer biserial correlations or point-biserial correlations. However, it is the relative size of the correlations rather than the actual size that matters, and these remain in the same order whichever is used. The higher

the correlation, the more discriminating the item. A minimum correlation of .2 is generally required. Items with negative or zero correlations are always excluded. In Table 2.4, only item d fully meets these criteria. Items b, c, and e would have to be removed, as they have either a negative or an undefined correlation (undefined because the formula would have led to a division by zero). There are no hard-and-fast rules about inclusion criteria for items in the final questionnaire. It is common to choose 70%–80% of the original items. The higher this discrimination index for the item, the better. But in Table 2.4, maybe keep item a—it is the only one left! The same procedure is used whether the data are from a knowledge-based or a person-based test.

Distractors

An examination of the use of distractor options by respondents who do not choose the correct or best option should be carried out for each item to ensure that each distractor is endorsed by a similar proportion of respondents. This can be done for each item, counting the number of times that each of its distractors has been endorsed. The number of endorsements should be similar for all these distractors. Items for which distractor options are not functioning properly should be considered for exclusion from the final questionnaire.

When deciding which items to include in the final version of your questionnaire, you will have to take many factors into account and balance them against each other. In addition to facility, discrimination, and distractors, you will need to consider the number of items that you require for the final version (at least 12, and more usually 20, are necessary for a reliable questionnaire) and how well the items fit the blueprint. For example, you might include an item with fairly poor discrimination if you have very few items from that area of the blueprint, or you might include an item with poor facility if it has reasonable discrimination. In a personality questionnaire, it is also important to ensure that there are approximately equal numbers of reversed and nonreversed items. Ways of improving items may become clear at this stage. For example, changing the wording of an item may improve facility, or a distractor may be made more realistic. However, it is not a good idea to change many items, as you will not know how these changes affect the reliability and validity of the questionnaire. The procedures of item analysis will inform you about the characteristics of each item. It is then up to you to decide which criteria are most important for the purpose of your particular questionnaire.

Decide which items from the pilot version of your questionnaire to include in the final version—taking account of facility, discrimination, and, if appropriate, distractors. Order the items and design the questionnaire as before.

Obtaining reliability

Reliability is an estimate of the accuracy of a questionnaire, and is discussed in more detail in Chapter 3 (the next chapter). For example, a questionnaire is reliable if a respondent obtains a similar score on different occasions, provided that the respondent has not changed in a way that affects their response to the questionnaire. When you publish your questionnaire, you will be expected to report its reliability. Hence, you will want to have some information about the impact of your particular item selection on this. You have only so far given your respondents the questionnaire once. However, it is possible

to estimate the reliability from the data you already have. There are two ways of doing this. Although there are many arguments over which is the most appropriate, they both generally (and rather surprisingly) deliver very similar results.

Cronbach's alpha

The first method is calculating a statistic called Cronbach's alpha, which is a measure of the internal consistency of the questionnaire. Cronbach's alpha is widely accepted as a surrogate for reliability. Most statistical packages allow you to do this quite easily from the data in an item-analysis table. The second is a method called split-half reliability.

Split-half reliability

The second method is called split-half reliability. Here the questionnaire is divided into two halves (usually odd and even items), and the correlation between the halves is used to produce an estimate of reliability for the whole questionnaire. For split-half reliability, the Pearson product-moment correlation coefficient between the two halves of the questionnaire is used in the Spearman–Brown formula to give an estimate of reliability for the whole questionnaire:

Spearman–Brown formula

$$r_{11} = (2r_{\frac{1}{2}\frac{1}{2}})/(1+r_{\frac{1}{2}\frac{1}{2}})$$

r_{11} = estimated reliability for the whole questionnaire and

$r_{\frac{1}{2}\frac{1}{2}}$ = correlation between two halves of the questionnaire.

For example, if the Pearson product-moment correlation coefficient between two halves of a questionnaire is .80:

$$r_{11} = 2(0.80)/(1 + 0.80) = 0.88$$

The greater the number of respondents, the better the estimate of reliability. If fewer than 50 respondents were included in the pilot study, it is necessary to have the final version of the questionnaire completed by more people, ensuring once again that they are similar to those for whom the questionnaire is intended. The dual use of the pilot-study data for item selection and reliability estimation will mean that reliabilities are overestimated. Ideally, data from at least 200 respondents who were not part of the pilot study should be used in calculating reliability. Where the questionnaire is intended for different types of respondents, it is common to show that it is reliable for each type. In this case, a total of 200 respondents would be needed altogether. Whatever measure of reliability is used, a coefficient of at least .7 is generally required for person-based questionnaires, and at least .8 for knowledge-based questionnaires.

For the GRIMS, split-half reliabilities were obtained for men and women separately for respondents in the pilot study, relationship therapy clients, and a general population group. Reliabilities ranged from .81 to .94.

Calculate the split-half reliability for the final version of your questionnaire using data from the relevant items from all of the respondents in the pilot study, plus additional respondents if necessary. For each respondent, calculate the total score for the even items in the final version of your questionnaire and

the total score for the odd items. Correlate the odd items with the even items using the Pearson product-moment correlation. Use this correlation coefficient in the Spearman–Brown formula to obtain an estimate of reliability for the whole questionnaire.

Assessing validity

The validity of a questionnaire is the extent to which it measures what it is intended to measure. Validity is discussed in Chapter 3 (next chapter), but at this point there are only two of the many forms of validity that you should apply.

Face validity

This describes the appearance of the questionnaire to respondents, i.e., whether or not it looks as if it is measuring what it claims to measure. If not, respondents may not take the questionnaire seriously. Look carefully at your selection of items and the general layout of the questionnaire with this in mind.

Content validity

This is the relationship between the content and the purpose of the questionnaire, i.e., whether or not there is a good match between the test specification and the task specification. For example, the blueprint for a questionnaire used in job selection should match the job description. Content validity is generally taken care of in constructing the blueprint and in the item analysis. However, it is important to check that the balance of items in the final version of your questionnaire matches the original blueprint.

Content validity of the GRIMS is high with respect to its specification, and good face validity has been incorporated into the item selection. It is also important for the GRIMS to have good diagnostic validity. This was established by determining that couples who presented at marriage guidance clinics had significantly higher scores than a matched sample from the general population. Moreover, couples presenting for marital therapy had significantly higher scores than couples presenting for sex therapy. Because the GRIMS was intended as a measure of improvement after therapy, it was important to obtain a rating of the validity of the GRIMS as an estimator of change. Couples were asked to complete the GRIMS before and after therapy; and the therapists, who were unaware of their clients' GRIMS scores, were asked to rate the couple on a five-point scale ranging from 0 ("improved a great deal") to 4 ("got worse"). The GRIMS scores for the male and female partners were averaged for each couple. The average score before therapy was subtracted from the average score after therapy to give a change score representing change during therapy. The change scores were correlated with the therapists' ratings of change, producing a correlation coefficient of .77. This is firm evidence for the validity of change in the GRIMS score as an estimate of change in the quality of the relationship(s) or in the effectiveness of therapy.

Ensure that your questionnaire has good face validity and content validity. Consider carefully what other forms of validity will be required at a later stage, and draft plans for any necessary data collection.

Standardization

Standardization involves obtaining scores on the final version of your questionnaire from appropriate groups of respondents (see Chapter 3). These scores are called norms. Large numbers of respondents must be carefully selected for a standardization group according to clearly specified criteria in order for norms to be meaningful. Norms can be obtained from the data in the pilot study, but this is not the preferred method.

With good norms, it is possible to interpret the score of an individual respondent, i.e., whether or not their score on the questionnaire is typical. This is useful if, for example, you wish to know how an individual child performs on an ability test compared with other children of the same age, or if you wish to determine how a person with a suspected clinical disorder compares with people who have been diagnosed as having that disorder.

It is not always necessary to produce norms. If your questionnaire has been developed for research, which involves comparing groups of respondents, norms can be useful in interpreting the performance of a group as a whole, but they are not crucial. If, however, you wish to interpret the score of an individual, it is necessary to have good norms against which to compare an individual score.

It is important to include as many respondents as possible in the standardization group, and to ensure that they are truly representative. A minimum of several hundred is generally required, but this depends to a large extent on the nature of the respondents. Some are easier to find than others, and it is often better to obtain a smaller group of very appropriate respondents than a larger but less appropriate one. In some cases, it is necessary to obtain several standardization groups or to stratify the standardization group according to relevant variables such as age, gender, or social class. Ideally, there should be several hundred respondents in each group or stratification. Norms should be presented in terms of the mean and standard deviation for each group or stratification.

The mean score for the standardization group is simply the average of the scores for the respondents in that group. The standard deviation is a measure of the amount of variation in the standardization group. (It is the square root of the average of the squared deviations from the mean.) If you have all of the scores in an Excel spreadsheet, it can easily be calculated using the STDEV function. Alternatively, use any statistical package.

Once you know the mean and the standard deviation of the standardization group, also frequently called the norm group, you can calculate for each person by how many standard deviations their score differs from the mean. This figure ranges between about -3.00 and about $+3.00$, and is referred to as a standard score or z score. One advantage of the standard score is that anyone who understands how they are calculated can immediately interpret someone's standard score in terms of how they compare with everyone in the standardization sample. If their z score is 0 , they are right at the average. If their z score is 1.00 , they are one standard deviation above the mean. If their z score is -1.50 , they are one and a half standard deviations below the mean, and so on. However, it would not be an easy score to give as feedback to someone regarding their result. Hence, there are various techniques available for rescaling standard scores to make them more presentable. These are called standardized scores. The most frequent of these as far as knowledge-based tests are concerned is the T score. To obtain a T score, you simply multiply the standard score by 10 , add 50 , and round to the nearest whole number. For person-based tests, it is more common to multiply the z score by 2 , add 5 , then round the answer off. This produces a score between 1 and 9 , called a

Table 2.5 Example of standardized scores obtained for all the individuals in a standardization sample of seven people

Person	Score					
	Raw	<i>z</i>	<i>T</i>	Stanine	<i>IQ</i>	Percentile
1	44	−1.28	37	2	81	10.03
2	48	−1.04	40	3	84	14.92
3	57	−.49	45	4	93	31.21
4	66	.05	51	5	101	52.00
5	75	.60	56	6	109	72.57
6	76	.66	57	6	110	74.54
7	90	1.50	65	8	123	93.32
Mean	65.14	0.00	50	5	100	
S.D.	16.54	1.00	10	2	15	

stanine. Nearly all personality tests report stanine scores. Even IQ scores today are normally standardized in the same way, but this time by multiplying by 15 and adding 100.

Table 2.5 gives an example of standardized scores obtained for all the individuals in a standardization sample of seven people. It also has an additional column containing the percentile equivalent—that is, the percentage of people in the standardization sample who obtained a score at this level or less.

The GRIMS was standardized using two groups: (i) a random sample of people consulting their family doctor with the usual variety of medical problems (a general-population group); and (ii) clients attending relationship guidance clinics and sexual therapy clinics (a relationship-problems group).

Standardize your questionnaire using a relevant group or groups of as many respondents as possible. Present the norms in terms of the mean and standard deviation for each group or stratification.