# Chapter 2
# Constructing Instruments to Achieve Measurement

In Chap. 1, assessment was defined as the engagement of an entity with some instrument and the recording of observations of the engagement according to some protocol. Assessment is a *precursor to measurement*, which is a *transformation of assessments*. How can assessment instruments be designed so they can be transformed successfully into measurements?

Instruments that provide measurements have to be constructed empirically and experimentally. They require knowing what demonstrates more or less of the relevant property. For example, in the measurement of temperature, there are ways of heating an object to increase its temperature or cooling an object to decrease its temperature. Then an instrument has to demonstrate that it reacts consistently with this increase or decrease with an increase or decrease in heat. Another familiar example is the measurement of mass. Mass reacts to gravity and so gravity can be exploited to measure different amounts of mass. A more complicated example is the measurement of the amount of sugar in a juice using the specific gravity of the juice.

The same theoretical understanding of a variable in the social sciences needs to be present in constructing ways of measuring it. Although often not explicit, educational intervention (teaching) has the intention of changing the relevant property of the student—perhaps the understanding of mathematics, literature and so on. The tasks and the marking keys, designed to reflect the understanding, need to be constructed carefully.

In constructing and then administering an instrument, great care needs to be taken so that the many aspects of administration that can go wrong do not go wrong. For example, sufficient time to reveal understanding needs to be available. Anomalies appear, that is results contrary to expectation, when something has gone wrong. In these cases, we do not change the criterion of measurement—instead, we look for a substantive or administrative reason for the anomaly. This may include the poor design or functioning of a particular item or task, or some broader problem. The methods of analysis that you learn in this book are about diagnosing such anomalies. These anomalies then need to be referenced to the test, questionnaire or other aspects of the instrument or its administration. The statistical anomalies tell you where to



D. Andrich and I. Marais, *A Course in Rasch Measurement Theory*,
Springer Texts in Education, https://doi.org/10.1007/978-981-13-7496-8_2

look for sources of problems but they hardly ever explain the problem. Typically the process of instrument design consists of a number of stages. During these stages, answers are sought to questions such as what range of content the instrument should cover, what format the items should be in, how many items should be included, how the items should be scored, etc.

(i) During an initial *conceptualizing or planning stage* the goal is to define what needs to be measured. This is then the conceptual definition of the variable to be measured and the measuring instrument becomes its operational definition. This stage typically includes a diagram or conceptual map which should include the different aspects of the variable to be measured. For a test of mathematics proficiency, this may be the content areas, e.g. algebra, measurement, geometry, etc. For a quality of life questionnaire, these may include aspects such as cognitive, motor and affective functioning of a person. The map should include the content of each of these aspects that the items are expected to cover. It will then typically result in a set of instrument specifications that will include the number of the items to measure each aspect or content area.

(ii) After items have been developed they are typically refined through a number of *stages*. Items should be reviewed by experts and/or administered to a small group. Typically it is not the actual responses of the small group members that are of interest but their interpretation of the wording and response format of each item. However, extremely unpredictable responses which invalidate the assessment may also emerge. Items may be modified and then trialled on a larger group. It is recommended that these trial responses be analysed according to the Rasch model, anomalies identified and items then modified as necessary.

(iii) Sometimes items are discarded. However, it is important *not* to discard items solely on statistical grounds, that is, that they misfit statistically to the model. Instead, each item that is seen as misfitting, which means it is not operating as consistently with the other items as indicated by the model, needs to be studied to understand *why* it might be misfitting. Only if it is understood why it is misfitting, and if it cannot be improved, should it be discarded. For example, it may be that some distractor in a multiple-choice item is not functioning as intended, that a response in an item of an attitude questionnaire captures some other aspect of the variable as is usually the case with the *undecided* category, or that there are too many categories in an ordered category format for raters to be able to use them consistently. In each case, the diagnosis provides an opportunity to not only improve the item but also to learn more about the construct to be measured and how it might be measured.

The problem with deleting items using statistical grounds only is that it risks eliminating items by chance that are sound, or some sound items may be affected in their fit by other items that really do have problems. Sometimes it is suggested that, for example, twice as many items should be constructed as is finally used. This is sound advice if no item is eliminated only on statistical grounds but on grounds of understood misfit, representation of items along the continuum, redundant items and so on. Sometimes having more items than required for one administration of the instrument can lead to a parallel form being constructed.

Two examples of the development of instruments used in health are provided by Doward et al. (2003) and Gilworth et al. (2003). The data analyses described in the rest of this book form part of the refinement stages of an assessment instrument.

Guides to test and questionnaire construction have a long history (e.g. for questionnaires Sudman & Bradburn, 1982; Oppenheim, 1992; and for achievement tests Bloom, Hastings, & Madaus, 1971). Because there are guides that describe the whole assessment design *process* in detail, the focus of this chapter is not the process of design. Instead, the chapter is a summary of some *observations on item and response format and the scoring of items* that have arisen from our research. In particular, there is an emphasis on successive response categories that are typically defined in a rating scale or the marking key of a test item. They are supposed to reflect successively more of the property to be measured. However, there is no guarantee that the categories will operate as intended. The ordering should be checked. The book provides a mechanism for doing so using the Rasch measurement model.

Instruments typically consist of items that require respondents to *generate* a response and/or those that require respondents to *choose* a response from among alternatives. The former is called a *constructed response* item and the latter a *selected response* item. In the rest of this chapter, we discuss, first, how both item types can be used to achieve measurements in tests of proficiency, and secondly, how selected response items can be used in rating scales to achieve measurements.

## Constructing Tests of Proficiency to Achieve Measurements

Three basic types of selected response items are typically used in tests of proficiency: alternate choice, multiple-choice and matching items. Table 2.1 shows examples of each type.

In a *matching* item, respondents are required to match each option in the right-hand column with one in the left-hand column. In an *alternate choice* item, a stem is followed by two response alternatives, typically TRUE/FALSE or YES/NO. An advantage of alternate choice items is that they are easy to write but a disadvantage is that respondents have a 50% chance of a correct answer if they guess randomly as opposed to a 25% chance of a correct answer on a four-alternative multiple-choice item. A *multiple-choice* item consists of a stem, followed by a number of response alternatives including the key (correct answer) and some distractors (incorrect answers).

Distractors are an integral part of a multiple-choice item. They should be plausible and should attract responses from those who do not have the required level of understanding to choose the correct answer (Smith, 1987). The quality of the distractors can make an item more or less difficult. For the same content of an item, distractors that are dismissed easily as incorrect responses by even the least able respondents

**Table 2.1** Examples of basic types of selected response items typically used in tests of proficiency

| | | |
|---|---|---|
| Alternate choice | An alternate choice item is an example of a selected response item | |
| | TRUE FALSE | |
| Multiple choice | Examinees have a 25% chance of randomly guessing the correct answer on a multiple-choice item with | |
| | a. 2 response alternatives | |
| | b. 4 response alternatives | |
| | c. 5 response alternatives | |
| | d. 6 response alternatives | |
| Matching | In the left column below are four different numbers of response options for a multiple-choice item. For numbers 1–4 listed in the left column record the letter from the right column that best matches an examinee's chance of randomly guessing the correct answer for that item | |
| | 1. 3 response options | a. 25% |
| | 2. 4 response options | b. 33.3% |
| | 3. 5 response options | c. 16.7% |
| | 4. 6 response options | d. 20% |

contribute to making an item easy; distractors that cannot be dismissed easily by even the most able respondents make an item difficult. If even one distractor cannot be readily dismissed by the moderately and very able respondents, then this distractor will contribute to the item being more difficult. Generally, not all distracters are equally plausible for a given proficiency of the respondents. In particular, one way of making a distractor plausible is to have it include aspects of a correct response (Andrich & Styles, 2009). However, there are those who argue that distractors should be plausible but completely wrong (e.g. Bertrand & Cebula, 1980).

Also typically used in tests of proficiency are *constructed response* items, which include items with a short answer up to essay type items. Van Wyke (2003) provides some guidelines for polytomous scoring of constructed response items. The paper shows how a Rasch model analysis confirmed that the marking keys of some mathematics items were working whereas the marking keys of other items were not working as required. In the cases of marking keys working as required, a higher score on an item required a greater proficiency to achieve than did a lower score. In the cases of marking keys not working as required, a higher score did not require a greater proficiency to achieve than did a lower score. Figure 2.1 shows the format and content of two mathematics items analysed in the paper.

**Item S016**

Six children are going to have a kayak race. They draw coloured marbles out of a hat, each representing a different kayak.

Doug chooses first. What are the chances that he will get a blue kayak?

________________

Doug gets his blue kayak. Amanda is to choose second. She wants either a blue or a black kayak. What are the chances that she gets what she wants?

________________

**Marking Key**

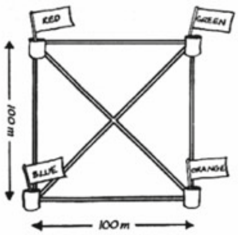
| | |
|---|---|
| 2 | 2/5 |
| 1 | 1/5 |
| 0 | incorrect response |

**Item S019**

Later you and your friends have a bike racing competition. Three courses are available. The turning points are marked by coloured flags. The courses are shown on the following map.

They are:

RED-GREEN-ORANGE-RED
RED-ORANGE-BLUE-RED
GREEN-ORANGE-BLUE-GREEN




(b) If you ride the GREEN-ORANGE-BLUE-GREEN course, through what angle does your bike turn at

(ii) the BLUE mark? ________________

**Marking Key**

| | |
|---|---|
| 2 | 135° |
| 1 | 45° |
| 0 | incorrect response |

**Fig. 2.1** Items S016 and S019 in van Wyke (2003)

A Rasch model analysis showed that the marking key of item S019 worked as required, whereas item S016's did not. To arrive at the partially correct response for item S016 the student correctly recognizes that there are five canoes to choose from, but incorrectly states that only one will suit Amanda (1/5). To arrive at the fully correct response, the student must recognize that two of the canoes will suit Amanda (2/5). The problem here is that there is no logical or empirical evidence to suggest that these responses form a developmental continuum, such that it would be meaningful to place them along an achievement scale. As a result, the item really functions dichotomously as correct or incorrect. Students who have little understanding of simple probability are not likely to even get the *out of five* part right, and will score 0 marks. Students with sufficient understanding to get the *out of five* part right are also likely to get the number of chances right as well. As a result, the middle response category, with a score of 1, fails to function properly.

In contrast, for item S019 the difference between a partially correct response (45°) and a fully correct response (135°) is quite substantial. The marking key for this item has identified two levels of response that do form part of a developmental continuum which you will learn about; understanding that the turned angle is really 135° is more difficult, and comes later developmentally, than recognizing that the drawn angle is 45°. Here there is logical and empirical evidence that it is meaningful to place these responses in this order on an achievement scale. Students who are awarded 1 mark for this item know something about angle measure, whereas students who are awarded 2 marks know this, but also know something more.

The Rasch model analysis of these and other constructed response items reflect how marking keys should operate to identify a hierarchy of responses; a higher level response should require more proficiency than a lower level response.

## Constructing Rating Scales to Achieve Measurements

Questionnaires often include *selected response* items with ordered response options. Table 2.2 shows some examples from Bock (1975) of selected response formats typically used in rating scales.

The format of the kind shown in example (*d*) of Table 2.2 where response options range from Strongly approve or Strongly agree to Strongly disapprove or Strongly disagree, is often called a Likert (1932) format.

Sometimes each response category is defined descriptively and sometimes only the end points are described leaving the intervening points without verbal description (*b* and *c*). In Table 2.2, all the verbal descriptions are qualitative. Sometimes quantitative

**Table 2.2** Examples of selected response formats from Bock (1975)

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| a | Dislike extremely | Dislike very much | Dislike moderately | Dislike slightly | Neither like nor dislike | Like slightly | Like moderately | Like very much | Like extremely |
| b | Weak | | | | | | | | Strong |
| c | Practically identical | 1 | . . . . | . . . . | 10 | Totally different | | | |
| d | Strongly approve | | Approve | | Undecided | | Disapprove | | Strongly disapprove |

descriptions like ‘Once a week’ are used. In order to achieve measurement, the *number, order and wording of response categories* need to be carefully considered and checked with data.

## *Number, Order and Wording of Response Categories*

### Number and Wording of Response Categories

Hagquist and Andrich (2004) analysed responses to a measure of self-reported health administered to adolescents in Sweden for 3 years of investigations. Table 2.3 shows the item characteristics at different years of investigations. They describe how Rasch analyses revealed that response categories were not working as intended. Table 2.3 shows the experimental changes made to the number and wording of the response categories in different years of investigations.

The table not only shows that three items were removed after the 1985/86 investigation, but also that the response categories used in 1985/86, 1993/94 and 1997/98

**Table 2.3** Item characteristics at different years of investigations from Hagquist and Andrich (2004)

| | 1985/86 | 1989/90 | 1993/94 and 1997/98 |
|---|---|---|---|
| Initial question | How often do you have the following complaints? | In the last 6 months, how often have you had the following complaints? | In the last 6 months, how often have you had the following complaints? |
| Items | Headache | Headache | Headache |
| | Stomach ache | Stomach ache | Stomach ache |
| | Backache | Have been irritable or in a bad temper | Backache |
| | Feel low | Feel nervous | Feel low |
| | Being irritable or in a bad temper | Have had difficulty getting to sleep | Have been irritable or in a bad temper |
| | Feel nervous | Felt dizzy | Feel nervous |
| | Difficulty getting to sleep | | Have had difficulty getting to sleep |
| | Feel dizzy | | Feel dizzy |
| Response categories | About every day | Often | About every day |
| | More than once a week | Sometimes | More than once a week |
| | About once a week | Seldom | About once a week |
| | About once a month | Never | About once a month |
| | Seldom or never | | Seldom or never |

were quantitative expressions (e.g. 'About once a week') whereas those used in 1989/90 were qualitative (e.g. 'Often'). Also, there were five response categories in the 1985/86, 1993/94 and 1997/98 investigations, whereas there were four in the 1989/90 investigation.

**Order of Response Categories**

Successive response categories should reflect successively more of the property measured. In the well-known Likert format, the 'undecided/not sure' response category is placed in the middle between the other categories, as shown in example (*d*) of Table 2.2. Being placed in the middle of the response continuum, it is intended that this category imply an attitude somewhere in between Agree and Disagree or in between Approve and Disapprove. However, whether this category works as intended is rarely checked.

Using the methods you learn in this book, Andrich, de Jong, & Sheridan (1997) showed that an 'undecided/not sure' response category did not operate as intended when placed in the middle between the other categories in an analysis of responses to an instrument measuring teachers' attitude towards a new teaching strategy. They recommend that the 'undecided/not sure' category, if needed, be placed separately to the side, e.g. strongly disagree, disagree, agree, strongly agree and undecided/not sure.

## *An Example of the Assessment of Writing by Raters*

Raters or judges often rate persons according to one or multiple criteria on some performance. When rating on only one criterion it is called *holistic* rating, whereas *analytical* rating is rating according to multiple criteria. In the latter case, a decision has to be made regarding the number of rating criteria and the number of response categories for each criterion.

Humphry and Heldsinger (2014) described how a rating scheme was revised after analyses showed that the rating criteria and categories were not working as required. The rating scheme was used in the assessment of writing of school children in Western Australia in year levels 3, 5 and 7. Raters used the same rating scheme to rate the writing of children in all three year levels. The criteria and number of response categories for each criterion are shown in Table 2.4. On the left of the table are the original rating criteria and numbers of response categories and on the right the revised criteria and numbers of categories.

The rating scheme was revised because raters using the original scheme were prone to give similar ratings on all the criteria arising from a holistic impression of the writing. This is also known as a *halo* effect. The effect resulted not so much from a bias of individual raters, but because of crude and arbitrary levels in the criteria of

**Table 2.4** Original and revised classification schemes for the assessment of writing (Humphry & Heldsinger, 2014)

| Original classification scheme | | Revised classification scheme | |
|---|---|---|---|
| Criterion | Score range | Criterion | Score range |
| On-balance judgement | 0–7 | On-balance judgement | 0–6 |
| Spelling | 0–5 | Spelling | 0–5 |
| Vocabulary | 0–7 | Vocabulary | 0–6 |
| Sentence control | 0–7 | Sentence structure | 0–6 |
| Punctuation | 0–6 | Punctuation of sentences | 0–2 |
| Form of writing | 0–7 | Punctuation within sentences | 0–3 |
| Subject matter | 0–7 | Narrative structure | 0–4 |
| Text organization | 0–7 | Paragraphing | 0–2 |
| Purpose and audience | 0–7 | Characterisation and setting | 0–3 |
| | | Ideas | 0–5 |
| Total score range | 0–60 | Total score range | 0–42 |

the rating scale that did not match aspects of the writing task. In particular, the rating categories were relatively crude and arbitrary and did not arise from the task.

The solution in this case, also described in the paper, involved a rewriting of criteria and the number of rating categories for each criterion so that the criteria and the number of categories for each criterion arose naturally from each task. New data collected with the revised criteria showed that the halo effect was eliminated. The differences in the number of categories among the criteria helped reduce the tendency to give the same rating for each criterion.

It is evident from Table 2.4 that some, not all, criteria were changed, and that the numbers of categories for the revised classification system were different across criteria. This variation exemplifies making the criteria relevant to each task and to the performances of the students engaged in each task. Both the chosen criteria and the number of categories for each criterion reflected the evidence that could be obtained from the writing.

## *An Example of the Assessment of the Early Development Indicator Instrument*

Andrich and Styles (2004) assessed the psychometric properties of the early development indicator (EDI) instrument. The authors aimed to establish the validity and reliability of each of the five subscales (physical health and well-being; social competence; emotional maturity; language and cognitive development; communication skills) using the Rasch measurement model.

The results showed that in sets of items the ordering of the categories was not working as intended. It was recommended that these items, with originally five ordered response categories, be reduced to two or three categories. Andrich and Styles (2004) explain that having more categories than teachers could use was a problem for reliability because assessors were not able to use the categories consistently, and for validity because it raises the question of whether successive categories indicate more of the property. Table 2.5 shows the items in each of the five subscales which were recommended to have a reduced number of response categories and descriptors. To confirm that reducing the number of categories was the relevant improvement, it was noted that the items in those subscales that had only three ordered categories worked correctly. These features indicated that five categories were too many categories for early childhood teachers to respond to consistently, while they could do so with only three categories. You will learn about the method they used to diagnose the problems in this book.

**Table 2.5** Items recommended to have a reduced number of response categories (Andrich & Styles, 2004)

| Subscale | Items | Original number of categories | Recommended number of categories | Suggested descriptors |
|---|---|---|---|---|
| PHWB | A2–A5 | 5 | 2 | Never/rarely,<br>Usually/always |
| | A9–A13 | 5 | 3 | Very poor/poor,<br>Average,<br>Very good/excellent |
| SC | C1, C2 | 5 | 3 | Very poor/poor,<br>Average,<br>Very good/excellent |
| EM | No change | | | |
| LCD | No change | | | |
| CS | B1–B7 | 5 | 3 | Very poor/poor,<br>Average,<br>Very good/excellent |

## The Measurement of Attitudes: Two Response Mechanisms

We conclude this chapter with a note on two response mechanisms at work in the measurement of attitudes. One is the so-called cumulative mechanism and the other is the unfolding mechanism. We include this section because they can be confused. However, in this book, we only deal with the cumulative mechanism.

In both, statements or questions are asked and the persons are required to *agree* or *disagree* to them. Sometimes persons are asked to indicate their strength of agreement or disagreement.

The key element in the construction of these variables is that the statements themselves represent different degrees of intensities, and that these can in principle be placed on a line of increasing intensity. Recall that a variable indicates a construct in which the idea of more or less, greater or smaller, and the like, is involved. Perhaps the best way to take the construction of these variables a step further is to consider an example. The cumulative mechanism will be considered first.

### *An Example: The Cumulative Mechanism*

Consider the three statements below which refer to drug testing in the workplace.

| In employment in the public service, drug testing | | Agree | Disagree |
|---|---|---|---|
| 1. | Is acceptable in some settings | A | D |
| 2. | Is acceptable and may be compulsory in some settings | A | D |
| 3. | Should be compulsory in all settings | A | D |

A feature of the structure of these statements is that they are of increasing intensity for agreement with drug testing in a workplace. Their characteristic is that if you did agree to the third statement, then you would tend to agree to the other two as well.

On the other hand, you may agree to the first statement but not agree to the second. If you did not agree to the second, then you would tend not to agree with the third.

If, however, you agree with the second, then you would tend to agree with the first but may not agree to the third.

Finally, you may disagree with all three statements.

If 1 is coded as *agree*
and 0 is coded as *disagree*

then the structure of the responses that are acceptable takes the form of Table 2.6.

In this case, the *agree* responses can be summed to give a total score, and the greater the score, the stronger the attitude towards drug testing. Thus, a person with a score of 3 has a stronger attitude for drug testing than a person with a score of 2,

**Table 2.6** Cumulative mechanism—structure of responses

| | Statement | | | |
|---|---|---|---|---|
| | 1 | 2 | 3 | Total score |
| Typical response patterns | 0 | 0 | 0 | 0 |
| | 1 | 0 | 0 | 1 |
| | 1 | 1 | 0 | 2 |
| | 1 | 1 | 1 | 3 |
| Atypical response patterns | 0 | 1 | 0 | |
| | 0 | 0 | 1 | |
| | 1 | 0 | 1 | |
| | 0 | 1 | 1 | |

and so on. This kind of structure and mechanism appears also in tests of achievement or performance.

Because the acceptable responses accumulate as the intensity of an attitude increases, the structure of the response mechanism is said to be *cumulative*.

The key point here is that if person A has an attitude stronger than person B, then A should have agreed to all statements that B agreed to, and in addition, one more.

In general, this structure is found in performance assessments and achievement testing when different tasks or questions have different difficulty. We deal with an example in Chap. 5 where the cumulative mechanism is elaborated.

## *An Example: The Unfolding Mechanism*

| In employment in the public service, drug testing | | Agree | Disagree |
|---|---|---|---|
| 1. | Is not acceptable in any setting | A | D |
| 2. | Is acceptable only in a few settings | A | D |
| 3. | Is acceptable in any setting | A | D |

The three statements are also of increasing intensity in attitude, with the first not supporting the drug testing and the last supporting it.

In this case, it is most likely that only *one* statement would have an agree response. If one agreed to statement 1, it is unlikely the person would agree to statements 2 and 3. If one agreed to statement 2, it is unlikely the person would agree to statements 1 and 3, and likewise for statement 3.

The response structure, with agree being coded 1 and disagree 0, takes the form of Table 2.7.

In this case, the measurement of attitude *cannot* be obtained by simply summing the scores. Instead, values must first be given to the statements that locate them on the line. The procedure of obtaining these values themselves is a complicated process,

**Table 2.7** Unfolding mechanism—structure of responses

| | Statement | | |
|---|---|---|---|
| | 1 | 2 | 3 |
| Typical response patterns | 1 | 0 | 0 |
| | 0 | 1 | 0 |
| | 0 | 0 | 1 |
| Atypical response patterns | 0 | 0 | 0 |
| | 1 | 1 | 0 |
| | 0 | 1 | 1 |
| | 1 | 0 | 1 |
| | 1 | 1 | 1 |

but at this stage perhaps you can give them intuitively reasonable values. These should be of increasing intensity. For example, we might give the first statement a value of –1, the second a value of 0 and the third a value of +1.

Then the attitude for each pattern would be calculated as follows:

| | Statement | | | |
|---|---|---|---|---|
| | 1 | 2 | 3 | |
| Value | – 1.0 | 0.0 | 1.0 | Attitude Values |
| Typical response pattern | 1 | 0 | 0 | $(1)\ (-1) + 0(0) + (0)\ (1) = -1$ |
| | 0 | 1 | 0 | $0(-1) + 1(0) + 0(1) = 0$ |
| | 0 | 0 | 1 | $0(-1) + 0(1) + 1(1) = 1$ |

Usually, more than three questions are asked, and having more questions increases the precision of the measurement. When you look at questionnaires in the future, consider which of these two types of structures governs them. If you are required to construct a questionnaire that is an operationalization of a construct, then you need to think about which of these structures you wish to use. Perhaps you can consider a construct, and make up some statements that would form either or both structures. You could try to ask some friends to agree or disagree to the statements, and see if their responses conform to the expected patterns.

## *A Practical Approach: Likert Scales*

The history and methods of measurement of social variables is interesting, but the two principles described above are central.

A practical method for constructing questionnaires for assessing attitudes and opinions that was developed by Likert (1932) and which now goes under the name of *Likert-style*, involved the following two modifications to the above procedures.

First, statements that reflected ambivalent attitudes, such as statement 2 in the unfolding mechanism of drug testing in the workplace, were eliminated. Consider the following three statements which appeared in a questionnaire constructed to measure attitudes towards capital punishment:

1. Capital punishment is one of the most hideous practices of our time.
2. I do not believe in capital punishment but I am not sure it is not necessary.
3. Capital punishment gives the criminals what they deserve.

Statements 1 and 3 express a clear attitude, while statement 2 expresses an ambivalent one. Such a statement is excluded leaving just statements 1 and 3. With all three statements, the mechanism is unfolding.

Second, persons are asked to respond by agreeing (strongly or not) or disagreeing (strongly or not) to the statements as in the format below:

| | | | | | |
|---|---|---|---|---|---|
| 1. | Capital punishment is one of the most hideous practices of our time (reversed re capital punishment) | Strongly disagree | Disagree | Agree | Strongly agree |
| 3. | Capital punishment gives the criminals what they deserve (positive re capital punishment) | Strongly disagree | Disagree | Agree | Strongly agree |

Persons are required to circle the number that corresponds to the response that best reflects their opinion.

Statements 1 and 3 would then be scored in a reverse way relative to each other. Thus, if 1 is assigned to *strongly agree* in statement 1, then 1 would be assigned to *strongly disagree* in statement 3.

Of course, this would be done by the researcher, and not be indicated to the respondents.

*Thus suppose person A responded as below*:

| | | Strongly disagree | Disagree | Agree | Strongly agree |
|---|---|---|---|---|---|
| 1. | Capital punishment is one of the most hideous practices of our time (reversed) | 4 | 3 | 2 | 1 |
| 3. | Capital punishment gives the criminals what they deserve | 1 | 2 | 3 | 4 |

**Table 2.8** Four statements on capital punishment taken from a set described in Wohwill (1963)

| | | Strongly disagree | Disagree | Agree | Strongly agree |
|---|---|---|---|---|---|
| A | Capital punishment is one of the most hideous practices of our time | 4 | 3 | 2 | 1 |
| B | Capital punishment is not an effective deterrent to crime | 4 | 3 | 2 | 1 |
| C | Until we find a more civilized way to prevent crime, we must have capital punishment | 1 | 2 | 3 | 4 |
| D | Capital punishment gives criminals what they deserve | 1 | 2 | 3 | 4 |

Person A would score 4 (reversed scoring) on the first question and 3 on the second question giving a score of 7. This is a high score relative to the maximum possible of 8 and minimum of 2 on the two questions, and indicates a strong positive attitude towards capital punishment.

*Now suppose person B responded as below*:

| | | Strongly disagree | Disagree | Agree | Strongly agree |
|---|---|---|---|---|---|
| 1. | Capital punishment is one of the most hideous practices of our time (reversed) | 4 | 3 | ② | 1 |
| 3. | Capital punishment gives the criminals what they deserve | 1 | ② | 3 | 4 |

Person B would score 2 (reversed scoring) on the first question and 2 on the second, giving a total of 4. This reflects a more moderate attitude towards capital punishment than a score of 7 obtained by person A.

In general, more than 2 questions would be asked, perhaps 10 or so. Although it may be difficult to construct in many cases, try also to have both the positively worded and the negatively worded statements themselves of different intensities.

For example, the two statements above on capital punishment, which are relatively extreme, may be supplemented with two other statements giving the set in Table 2.8. These are the kinds of instruments that are often analysed using Rasch models (Hagquist & Andrich, 2004).

## Exercises

Respond to the statements in Table 2.8 and give yourself a score. Are you for or against capital punishment and to what degree?

## References


Andrich, D., & Styles, I. (2004). Final report on the psychometric analysis of the Early Development Instrument (EDI) using the Rasch model: A technical paper commissioned for the development of the Australian Early Development Instrument (AEDI). http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.495.5875&rep=rep1&type=pdf.

Andrich, D., & Styles, I. (2009). Distractors with information in multiple choice items: A rationale based on the Rasch model. In E. V. Smith Jr. & G. E. Stone (Eds.), *Criterion referenced testing: Practice analysis to score reporting using Rasch measurement models* (pp. 24–70). Maple Grove: JAM Press.

Andrich, D., de Jong, J. H. A. L., & Sheridan, B. E. (1997). Diagnostic opportunities with the Rasch model for ordered response categories. In J. Rost & R. Langeheine (Eds.), *Applications of latent trait and latent class models in the social sciences* (pp. 59–70). Münster and New York: Waxmann.

Bertrand, A., & Cebula, J. P. (1980). *Tests, measurement, and evaluation: A developmental approach*. Boston, MA: Addison-Wesley.

Bloom, B. S., Hastings, J. T., & Madaus, G. F. (1971). *Handbook on formative and summative evaluation of student learning*. New York: McGraw-Hill.

Bock, R. D. (1975). *Multivariate statistical methods in behavioral research*. New York: McGraw-Hill.

Doward, L. C., et al. (2003). Development of the ASQoL: A quality of life instrument specific to ankylosing spondylitis. *Annals of the Rheumatic Diseases, 62,* 20–26.

Gilworth, G. et al. (2003). Development of a Work Instability Scale for Rheumatoid Arthritis. *Arthritis & Rheumatism* (Arthritis Care & Research)*, 49*(3), 349–354.

Hagquist, C., & Andrich, D. (2004). Measuring subjective health among adolescents in Sweden: A Rasch analysis of the HBSC instrument. *Social Indicators Research, 68,* 201–220.

Humphry, S. M., & Heldsinger, S. A. (2014). Common structural design features of rubrics may represent a threat to validity. *Educational Researcher, 43*(5), 253–263.

Likert, R. A. (1932). A technique for the measurement of attitudes. *Archives of Psychology, 140,* 5–55.

Oppenheim, A. N. (1992). *Questionnaire design, interviewing and attitude measurement*. London: Pinter.

Smith, R. M. (1987). Assessing partial knowledge in vocabulary. *Journal of Educational Measurement, 24*(3), 217–231.

Sudman, S., & Bradburn, N. M. (1982). *Asking questions: A practical guide to questionnaire design*. San Fransisco: Jossey-Bass.

van Wyke, J. F. (2003). Constructing and interpreting achievement scales using polytomously scored items: A comparison between the Rasch and Thurstone models. Professional Doctorate thesis, Murdoch University, Western Australia.

Wohwill, J. F. (1963). The measurement of scalability of non-cumulative items. *Educational and Psychological Measurement, 23,* 543–555.