# CHAPTER 3. THE IMPERFECT CUMULATIVE SCALE

## 3.1 Model Violations

If a set of items does not form a perfect Guttman scale but contains a few “wrong” responses, we do not necessarily need to discard it. A “wrong” response, or a “model violation,” “model error,” or simply “error,” is a response that is inconsistent with the implications of the model. Guttman’s scaling model is very restrictive. With $k$ dichotomous items, there are $2^k$ possible response patterns in total, but only $k + 1$ response patterns form a perfect Guttman scale. So with 8 dichotomous items, there are $2^8 = 256$ possible response patterns but only 9 acceptable patterns. It is highly unlikely, even with the best possible set of questions, that a dataset contains only these nine acceptable response patterns. We therefore need to consider how to define *model violation* and how many model violations can be accepted.

In the 1950s and 1960s there was much discussion about how to define the number of errors or model violations for response patterns with more than two items. Take, for instance, the response pattern ABCD,1101 to the items A, B, C, and D, in which A, B, C, and D are ordered from the easiest to the most difficult. Are there two errors, because item C should be answered positively and item D should be answered negatively, if the number of positive responses is kept constant? Or is there only one error, because changing the response of either C or D would make the pattern perfect?

To add to this confusion, we also have to ask whether we still can use the response pattern ABCD,1101 for measurement purposes and, if so, how. Should such a person get the scale value 2 because her response pattern can be made perfect by changing the response to item D to 0? Should she get the scale value 4 because her response pattern can be made perfect by changing the response to item C to 1? Or should she get scale value 3 because she gave the positive response to three items? An excellent review of this debate can be found in chapter 2 of Mokken (1971). We return to this question later in this chapter.

## 3.2 Error: Violation of a Transitivity Relation Between Items and Subjects

The answer a person gives to a question is interpreted as a dominance relation between the person and the question (the item). The person dominates the item if she gives the positive response, and the item dominates the person if she gives the negative response. If we have more people and more items, and if all people and items can be represented along a single dimension, we can infer two other types of dominance relationships, namely dominance relations between people and dominance relations between items. This is the case because when all people and items can be represented along a single dimension, all these dominance relations are transitive. And if all dominance relations are transitive, then we have managed to order both the people and the items with respect to each other along that dimension. The concept of a model violation, or an error, can therefore best be explained as the violation of a transitivity relation.

Let us assume three items, A, B, and C, in order of difficulty, and four subjects, W, X, Y, and Z, in order of ability (Figure 3.1):

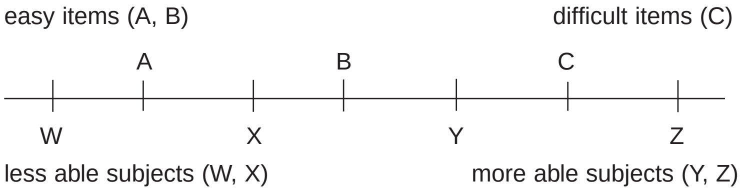

**Figure 3.1** Four subjects (W, X, Y, Z) and three items (A, B, C) along a cumulative scale.

1. We can specify a transitivity relationship *between three items* as follows: If item B is more difficult than item A, and if item C is more difficult than item B, then item C is more difficult than item A. This transitivity relationship is logically true if we define the "easiness"—and hence also the "difficulty"—of an item by the frequency (or proportion, as the relative frequency) of people who give the positive response to that item. If two or more items have exactly the same proportion of positive answers, we cannot order them, and they are tied. In terms of Table 2.1, both cells (0,1) and (1,0) should then be empty. But if the proportion of people agreeing with each item differs, then we have found the order of the items based on their "popularity" or, conversely, "difficulty."

Now we can specify three other transitivity relationships.

2. *Between two people and one item.* If person Y is more able than item B is difficult (i.e., if person Y is represented to the right of item B on

the dimension), and if person X is less able than item B is difficult (i.e., person X is represented to the left of item B), then person Y is more able than person X. This transitivity relationship (Y > B, B > X, and therefore Y > X) is used to order two people in terms of their ability, or scale value, on the dimension. The outcome of this transitivity relationship can then be used in the third transitivity relation.

3. *Between three people.* If person Y is more able than person X, and if person Z is more able than person Y, then person Z is more able than person X. This transitivity relationship allows us to give rank orders as scale values to all people (i.e., to give scale values on an ordinal scale to these people).

There is a fourth kind of transitivity relationship that is crucially important.

4. *Between one person and two items.* If person Y is more able than item B is difficult, and if item B is more difficult than item A, then person Y is also more able than item A is difficult. In other words, if person Y can give the positive response to the difficult item B, then he should also be able to give the positive response to the easy item A. Note that I purposely write *should,* because it is not always true that if a person gives the positive response to a difficult item, he also gives the positive response to an easier one.

Only the last of these four transitivity relations can logically be violated. We therefore define a model violation as *the violation of the transitivity relationship between one person and a pair of items.* Each time a person gives the positive response to a difficult item and the negative response to any easier item, she violates the cumulative model; that is, she makes an error with respect to the deterministic cumulative model. That is why this error is sometimes called a *Guttman error.* The number of errors a person makes in the response pattern ABCD,1101 is therefore 1, because only item pair (pair CD) violates the model.[1]

Let us see how to calculate the number of errors in the following response patterns (in which the items are given in order of difficulty from

[1] Establishing the number of model violations is possible only if we have established the difficulty order of the items. Previous definitions of model violations in terms of changing responses such that perfect patterns occur imply that such changes may affect the difficulty order of the items. However, this is not the case when model violations are defined in terms of a transitivity relationship between a subject and two items. In practical applications of this model the researcher may either use a preestablished theoretical order of the items or experiment with different possible orders of difficulty. We will not pursue this theme further in this monograph and will assume that the difficulty order of the items is known.

left to right). With four items there are six—namely 4*(4–1)/2—item pairs. The number of model violations is the number of item pairs that violate the model. Table 3.1 gives some examples. The order of the items was previously established by their order of difficulty in the sample, where A is the easiest and D is the most difficult item. In the first response patterns (ABCD,0011) four of the six item pairs violate the model. The number of model violations in this response pattern, or of a subject who gives this response pattern, therefore is four.

| *ABCD* | *AB* | *AC* | *AD* | *BC* | *BD* | *CD* | |
|---|---|---|---|---|---|---|---|
| 0 0 1 1 | 0 | 1 | 1 | 1 | 1 | 0 | 4 errors: item pairs AC, AD, BC, BD |
| 0 1 1 1 | 1 | 1 | 1 | 0 | 0 | 0 | 3 errors: item pairs AB, AC, AD |
| 0 1 0 1 | 1 | 0 | 1 | 0 | 0 | 1 | 3 errors: item pairs AB, AD, CD |
| 0 0 0 1 | 0 | 0 | 1 | 0 | 1 | 1 | 3 errors: item pairs AD, BD, CD |
| 1 0 1 1 | 0 | 0 | 0 | 1 | 1 | 0 | 2 errors: item pairs BC, BD |
| 0 1 1 0 | 1 | 1 | 0 | 0 | 0 | 0 | 2 errors: item pairs AB, AC |
| 1 0 0 1 | 0 | 0 | 0 | 0 | 1 | 1 | 2 errors: item pairs BD, CD |
| 0 0 1 0 | 0 | 1 | 0 | 1 | 0 | 0 | 2 errors: item pairs AC, BC |
| 1 1 0 1 | 0 | 0 | 0 | 0 | 0 | 1 | 1 error: item pair CD |
| 1 0 1 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 error: item pair BC |
| 0 1 0 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 error: item pair AB |
| # of violations | 4 | 4 | 4 | 4 | 4 | 4 | 24 model violations in this (hypothetical) dataset |

**Table 3.1** Comparison of item pairs that violate the cumulative model of four items in order from least to most difficult (1: violation; 0: no violation).

## 3.3 Extension of the Definition of Error to a Larger Dataset

When the number of errors in the response pattern of one subject has been defined as the number of item pairs in that response pattern that violate the cumulative model, it is easy to define the total number of errors in the dataset: It is simply the sum of the number of errors over all subjects. So if the dataset consists of the 11 response patterns in Table 3.1 then the total number of errors in the dataset would be 24, calculated either from the last column or from the last row.

We can also calculate the number of errors contributed by each item separately. Knowing the number of errors in each item will turn out to be helpful in a later stage of developing a cumulative scale or evaluating whether all items in a cumulative scale are equally good. Although we need an item pair to define an error, we can distinguish the members of the pair by attributing an error in the pair to each of the two items separately. The number of errors for each item is then defined as the total number of item pairs that involve that item and that contain an error, summed over all people.

A small example is given in Table 3.2. Whereas in Table 3.1 we compared each item pair, in Table 3.2 we look at individual items by combining the item pairs that contain them. Item A consists of item pairs AB, AC, and AD, item B consists of item pairs AB, BC, and BD, and so on. For subject 1 with response pattern 0001, three item pairs violate the model: AD, BD, and CD. So items A, B, and C are each involved once, and item D is involved three times. Rather than showing all 11 response patterns with four items that contain one or more errors, here we show only five such patterns.

| *Errors in item (i.e., in item pairs containing that item)* | | | | | | | |
|---|---|---|---|---|---|---|---|
| | *ABCD* | *A* | *B* | *C* | *D* | *Total* | *Item pairs* |
| Person 1 | 0 0 0 1 | 1 | 1 | 1 | 3 | 3 | AD, BD, CD |
| Person 2 | 0 1 1 1 | 3 | 1 | 1 | 1 | 3 | AB, AC, AD |
| Person 3 | 1 0 0 1 | 0 | 1 | 1 | 2 | 2 | BD, CD |
| Person 4 | 1 0 1 1 | 0 | 2 | 1 | 1 | 2 | BC, BD |
| Person 5 | 0 0 1 0 | 1 | 1 | 2 | 0 | 2 | AC, BC |
| Total | | 5 | 6 | 6 | 7 | 12 | |

**Table 3.2** Calculating the number of errors in each item separately.

## 3.4 How Can We Evaluate the Amount of Model Violation in a Dataset?

Now that we have established what we mean by a model violation or a model error, and we can calculate the number of such errors, how do we evaluate the amount of error? What do we mean by “not too many errors”? Here again, a number of different answers have been given in the literature, and again Mokken’s (1971) book provides an excellent overview. If we can compare the number of errors we observe with some benchmark, we may be able to express the amount of model error in a measure of model fit. Several measures of model fit, which we may call criteria for scalability, have been proposed. We will discuss some of the major ones now.

What is the benchmark against which we want to compare the number of errors in the dataset? One answer that comes to mind is to use the maximum number of errors possible for this dataset. But what is the maximum number of errors that people can make?

The worst situation would be the one in which every person gave a positive response to all the difficult items and a negative response to all the easy items, as in ABCD,0011. But in almost all empirical situations the difficulty of the item is established from the very dataset under study. So if everybody gave the response ABCD,0011, items C and D would automatically be defined as the easy items and A and B as the difficult ones. In such a situation the maximum number of errors could never be as high as the total number of responses, the number of people times the number of items ($N * k$). In fact, it is not easy to determine the maximum number of errors possible in a dataset. A number of procedures give us estimates of that number, but these can be proven to overestimate the real maximum number of errors.

If we were able to determine the maximum number of errors possible, then we could develop a criterion for scalability by comparing the number of errors observed to the maximum number of errors. The first criterion for scalability developed by Guttman was called *the coefficient of reproducibility,* **Rep**.

$$\mathbf{Rep} = 1 - \frac{\mathrm{Err(obs)}}{N * k} \tag{3.1}$$

Here, Err(obs) is the observed number of errors and $N * k$ is the total number of responses. **Rep** can therefore be interpreted as the proportion of responses not in error.

In later proposed criteria for scalability, $N * k$ was replaced by a better estimate of the maximal number of model violations, Err(max). *The coefficient of scalability* **S** was then defined as follows:

$$\mathbf{S} = 1 - \frac{\text{Err(obs)}}{\text{Err(max)}} \tag{3.2}$$

But even if we had a proper estimate of the maximum number of errors, we have another reason for not wanting to compare the number of errors observed with the maximum number of errors. This problem can best be understood if we think of establishing the criterion for scalability as a form of hypothesis testing. We generally test one hypothesis by comparing it with another hypothesis. One hypothesis is that our dataset does indeed form a cumulative scale. But what should the other hypothesis be? When we compare the number of errors that we have observed with the maximum number of errors possible, given a candidate ordering of the items, we are saying in essence that the other hypothesis is that the data conform to a model that is as different from a cumulative scale as possible. But how sensible is this as the other hypothesis?

A more appropriate other hypothesis is the null hypothesis, according to which the items are simply unrelated. They do not form a cumulative scale, nor do they conform to some other weird extreme model. If the items are unrelated, we cannot predict whether someone who gives the positive response to a supposedly difficult item will give the positive response to a supposedly easier item. Or, put differently, our prediction about whether a person gives the positive response to an easy item does not change if we know that he responded positively to a more difficult item.

The idea that we should compare the number of errors observed with the number of errors expected *under statistical independence*, Err(exp), was suggested by Loevinger in 1948 but not taken up until the end of the 1960s by Mokken. Mokken reintroduced Loevinger's *coefficient of homogeneity* **H** as a criterion for scalability:

$$\mathbf{H} = 1 - \frac{\text{Err(obs)}}{\text{Err(exp)}} \tag{3.3}$$

Here, Err(exp) for item pair (**i**,**j**) is the number of errors expected under statistical independence, and Err(obs) is the number of subjects who give both a positive response to the more difficult item **j** and a negative response to the easier item **i**.[2] **H** = 1 implies perfect model fit, because then there are no errors to be observed. That is also true when **Rep** = 1 or when **S** = 1. **H** = 0 implies that we cannot distinguish our dataset from a completely random dataset. **H** can be negative when we observe more errors then we would expect under statistical independence. This might happen, for instance, if we used the wrong order of difficulty of the items.

So let us compare the following two hypotheses:

| | | |
|---|---|---|
| The null hypothesis | $H_0$ | The items are unrelated. |
| The model hypothesis | $H_1$ | The items form a cumulative scale. |

It is common practice in statistics to want to reject or falsify the null hypothesis in favor of the model hypothesis. The model hypothesis is generally called the alternative hypothesis. To put this strategy of model testing more informally, everything worse than perfect is bad, but everything better than random is good.

The easiest way to describe the number of errors expected under statistical independence is by reference to a cross table of two items. As an example, let us return to the two questions about religious beliefs:

| | |
|---|---|
| Question A: Do you believe in heaven? | yes/no |
| Question B: Do you believe in hell? | yes/no |

Let us assume that in interviews with 100 people, we find that 24 people believe in heaven and hell, 34 believe in neither heaven nor hell, 36 believe in heaven but not hell, and 6 believe in hell but not heaven (Table 3.3a). Do these two items form a cumulative scale? Let us compare the number of errors observed with the number of errors expected under statistical independence.

[2] Err(exp) is calculated as $[1 - p(i)]*p(j)*N$, in which $p(i)$ and $p(j)$ are the relative frequencies with which the positive response to these items were given, and $N$ is the sample size of the dataset. **H** has the same interpretation as Goodman and Kruskal's (1979) coefficient λ (lambda) as a proportional reduction in error measure. It can also be interpreted as the ratio of the correlation between the two variables over the highest possible correlation, given the marginal distributions of the two variables.

| | | Heaven | | |
|---|---|---|---|---|
| | | Yes | No | Total |
| Hell | Yes | 24 | 6 | 30 |
| | No | 36 | 34 | 70 |
| | Total | 60 | 40 | 100 |

**Table 3.3a** Hypothetical empirical situation: Belief in heaven versus belief in hell.

How do we find the number of errors observed? In the 2 × 2 cross table in which the row item is the difficult one (only 30 positive answers) and the column item is the easier one (60 positive responses), the upper right top cell (hell yes, heaven no) can be called the error cell. In a perfect cumulative scale, this cell is empty, as in the second cross table (Table 3.3b), called the perfect situation. In the empirical situation, however, the error cell contains 6 people, that is, Err(obs) = 6.

| | | Heaven | | |
|---|---|---|---|---|
| | | Yes | No | Total |
| Hell | Yes | 30 | 0 | 30 |
| | No | 30 | 40 | 70 |
| | Total | 60 | 40 | 100 |

**Table 3.3b** Perfect Guttman scale for items from Table 3.3a.

How do we find the number of errors expected under statistical independence? In the case of statistical independence the probability of a given response to two items is simply the product of the probability of the response to each item taken separately. So if the probability of a positive response to the difficult item is 0.30 and the probability of a negative response to the easy item is 0.40 (1 – 0.60), the probability of the two responses taken together is 0.3 * 0.4 = 0.12. Because there are 100 people in the dataset, we expect 100 * 0.12 = 12 people to give the error response to these two items if the items are statistically independent. Therefore, Err(exp) = 12 (Table 3.3c).

| | | Heaven | | |
|---|---|---|---|---|
| | | Yes | No | Total |
| Hell | Yes | 18 | 12 | 30 |
| | No | 42 | 28 | 70 |
| | Total | 60 | 40 | 100 |

**Table 3.3c** Cross table if response to items from Table 3.3a were statistically independent.

Now we can compute the coefficient of homogeneity for this scale of two items:

$$\mathbf{H} = 1 - \frac{6}{12} = 0.50$$

Note that if the denominator Err(exp) becomes 0, it will be impossible to calculate the **H** value. Err(exp) becomes 0 if all subjects give the positive response to the easier item or if all subjects give the negative response to the more difficult item. Items to which every subject gives the same response are therefore not included in the analysis.

The coefficient of homogeneity is 0.50. Now the next question is, “Is this high or low?”

## 3.5 Evaluating the Coefficient of Homogeneity

There are two ways to approach the question, “How high or low is a coefficient of homogeneity **H** = 0.50?” Informally, we might ask, “How close is 0.50 to 0.00?” or no homogeneity, or we might ask, “How close is 0.50 to 1.00?” or perfect homogeneity. There is a statistical answer to the first question. It is possible to estimate the probability that a certain value of the coefficient of homogeneity (say, our 0.50) in a sample of size *N* is found in a population that has an **H** coefficient of 0.00, which means that the responses to all items are unrelated. This statistical answer is given in Appendix 1. It depends on finding the distribution of the **H** coefficient, in the case that all responses are statistically independent, and deriving a (one-sided) confidence interval between 0 and a positive value, given a particular exceedance probability α (generally

5%). If **H** falls within this confidence interval, we accept the null hypothesis and reject the model hypothesis that our items form a cumulative scale. This decision is formulated in terms of a *Z* and a *Z*(*i*) statistic, *Z* for the whole scale and *Z*(*i*) for item *i:* If *Z* (or *Z*(*i*)) is high enough (roughly >3), then the homogeneity of the whole scale (or of item *i*) cannot be explained by chance.

But accepting the model hypothesis—that the coefficient of homogeneity is higher than 0.00 in the entire population—does not provide an answer to the second question: How high is high, or how close is our dataset to perfect homogeneity? Unfortunately, there is no simple answer to this question. Correlations, or coefficients of homogeneity, that are statistically significant may still not be very important. Mokken has suggested that datasets with coefficients below 0.30 are not homogeneous enough to form a cumulative scale. He based this suggestion on substantive experience and informal comparison with scales accepted on the basis of reliability and factor analysis.

## 3.6 Using the Coefficient of Homogeneity in Scales With More Than Two Items

Now that we can test whether two items form a cumulative scale by interpreting the coefficient of homogeneity, we can extend this test to scales with more than two items. In such a scale we sum the number of errors observed, Err(obs), in each item pair. So with four items, we add over the six pairs. We can also calculate the amount of error expected under statistical independence, Err(exp), for each item pair and sum over all item pairs. The coefficient of homogeneity for the whole scale, **H**, is as follows:

$$\mathbf{H} = 1 - \frac{\sum_{i=1}^{k-1} \sum_{j=i+1}^{k} \mathrm{Err(obs)}}{\sum_{i=1}^{k-1} \sum_{j=i+1}^{k} \mathrm{Err(exp)}} \qquad (3.4)$$

Let us give an example with four items in Tables 3.4 and 3.5:

|  |  | *Item B* |  |  |  |  | *Item C* |  |  |  |  | *Item D* |  |  |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
|  |  | Yes | No |  |  |  | Yes | No |  |  |  | Yes | No |  |
|  | Yes | 25 | 5 | 30 |  | Yes | 24 | 6 | 30 |  | Yes | 26 | 4 | 30 |
| *Item A* |  |  |  |  | *Item A* |  |  |  |  | *Item A* |  |  |  |  |
|  | No | 25 | 45 | 70 |  | No | 36 | 34 | 70 |  | No | 54 | 16 | 70 |
|  |  | 50 | 50 | 100 |  |  | 60 | 40 | 100 |  |  | 80 | 20 | 100 |

|  |  | *Item C* |  |  |  |  | *Item D* |  |  |  |  | *Item D* |  |  |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
|  |  | Yes | No |  |  |  | Yes | No |  |  |  | Yes | No |  |
|  | Yes | 40 | 10 | 50 |  | Yes | 43 | 7 | 50 |  | Yes | 52 | 8 | 60 |
| *Item B* |  |  |  |  | *Item B* |  |  |  |  | *Item C* |  |  |  |  |
|  | No | 20 | 30 | 50 |  | No | 37 | 13 | 50 |  | No | 28 | 12 | 40 |
|  |  | 60 | 40 | 100 |  |  | 80 | 20 | 100 |  |  | 80 | 20 | 100 |

**Table 3.4** Example with four items.

| *Item pair* | *AB* | *AC* | *AD* | *BC* | *BD* | *CD* | *Sum* |
|---|---|---|---|---|---|---|---|
| Err(obs) | 5 | 6 | 4 | 10 | 7 | 8 | 40 |
| Err(exp) | 15 | 12 | 6 | 20 | 10 | 12 | 75 |
| **H** | 0.67 | 0.50 | 0.33 | 0.50 | 0.30 | 0.33 | 0.47 |

**Table 3.5** Summary of the results from Table 3.4.

The sum of the errors *observed* over all item pairs is 5 + 6 + 4 + 10 + 7 + 8 = 40.

The sum of the errors *expected* over all item pairs is 15 + 12 + 6 + 20 + 10 + 12 = 75.

$$\mathbf{H} = 1 - \frac{40}{75} = 0.47$$

It is now also possible to determine the coefficients of scalability of each of the four items, A, B, C, and D. In this case the number of errors observed and expected have to be summed over the item pairs that include the item under scrutiny.

$$\mathbf{H}_{\mathbf{A}} = 1 - \frac{5+6+4}{15+12+6} = 1 - \frac{15}{33} = 0.55 \quad \mathbf{H}_{\mathbf{B}} = 1 - \frac{5+10+7}{15+20+10} = \frac{22}{45} = 0.51$$

$$\mathbf{H}_{\mathbf{C}} = 1 - \frac{6+10+8}{15+20+12} = 1 - \frac{24}{47} = 0.49 \quad \mathbf{H}_{\mathbf{D}} = 1 - \frac{4+7+8}{6+10+12} = \frac{19}{28} = 0.32$$

Each of the four items has a coefficient of homogeneity that is higher than the proposed lower boundary of 0.30. It is easy to show that when each of the items in a cumulative scale has a coefficient of homogeneity larger than some value *c*, the scale as a whole will also have a coefficient of homogeneity that is larger than *c*. The inspection of the coefficient of homogeneity of individual items allows the researcher to evaluate these items for their inclusion in the cumulative scale. Items that are not sufficiently homogeneous should not be part of a cumulative scale. Item D, for instance, just barely makes the lower boundary of 0.30. Table 3.6 gives another numerical example.

*Number of errors in a cumulative scale*

| *A* | *B* | *C* | *D* | *Freq.* | *AB* | *AC* | *AD* | *BC* | *BD* | *CD* |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | | |
| 1 | 1 | 1 | 1 | 70 | | | | | | |
| 1 | 1 | 1 | 0 | 240 | | | | | | |
| 1 | 1 | 0 | 1 | 40 | | | | | | 40 |
| 1 | 0 | 1 | 1 | 20 | | | | 20 | 20 | |
| 0 | 1 | 1 | 1 | 8 | 8 | 8 | 8 | | | |
| 1 | 1 | 0 | 0 | 160 | | | | | | |
| 1 | 0 | 1 | 0 | 60 | | | | 60 | | |
| 1 | 0 | 0 | 1 | 28 | | | | | 28 | 28 |
| 0 | 1 | 1 | 0 | 16 | 16 | 16 | | | | |
| 0 | 1 | 0 | 1 | 14 | 14 | | 14 | | | 14 |
| 0 | 0 | 1 | 1 | 4 | | 4 | 4 | 4 | 4 | |
| 1 | 0 | 0 | 0 | 168 | | | | | | |
| 0 | 1 | 0 | 0 | 48 | 48 | | | | | |
| 0 | 0 | 1 | 0 | 24 | | 24 | | 24 | | |
| 0 | 0 | 0 | 1 | 10 | | | 10 | | 10 | 10 |
| 0 | 0 | 0 | 0 | 90 | | | | | | |
| --- | --- | --- | --- | ---- | -- | -- | --- | -- | -- | -- |
| 726 | 596 | 442 | 194 | 1,000 | 86 | 52 | 36 | 108 | 62 | 92 |
| No. errors expected: | | | | | 163 | 121 | 53 | 179 | 115 | 108 |
| **H**(*ij*): | | | | | .47 | .57 | .32 | .40 | .46 | .15 |

The number of errors expected in the cumulative scale for item pair AB was calculated as (1,000 – 726)*.596 = 163.30.
**H**(AB) was calculated as 1 – 86/163 = 0.47.

Item coefficients for the cumulative scale:

E(o)A: 86 + 52 + 36 = 174 E(e)A: 163 + 121 + 53 = 337
H(A) = 1 – 174/337 = .48

E(o)B: 86 + 108 + 62 = 256 E(e)B: 163 + 179 + 115 = 457
H(B) = 1 – 256/457 = .44

E(o)C: 52 + 108 + 92 = 252   E(e)C: 121 + 179 + 115 = 408
H(C) = 1 – 252/408 = .38

E(o)D: 36 + 62 + 92 = 190   E(e)D: 53 + 115 + 108 = 276
H(D) = 1 – 190/276 = .31

------   ------

Total Err(obs) = 872   Total Err(exp) = 1,478

**H**$_{\text{(scale)}}$= 1 – 872/1,478 = .41

For total Err(obs) and Err(exp) we can divide by 2, because each pair contributes twice to a model violation.

**Table 3.6** A small numerical example for the cumulative scale.

For scales larger than two items, we can inspect not only whether all **H**(*ij*)s are significantly higher than 0 but also whether all **H**(*i*)s and the overall **H** are significantly higher than 0 (See Appendix 1 for elaboration).

The use of 0.30 as a lower boundary for the homogeneity of each item and of the scale as a whole is generally far higher than the boundary for statistical significance. A homogeneity value of 0.30 could be insignificant only if the number of respondents is very small (say, below 50), the number of items is very small (say, 2 or 3), and the difficulty of the items is extreme (say, above 0.90 or below 0.10). In analyses where these conditions do not hold, this precaution of testing against the null hypothesis is generally not necessary. We want to keep the lower boundary higher than the boundary for statistical significance for reasons of interpretability or substantive relevance. In our experience scales or items with homogeneity values below 0.30 are difficult to interpret.

## 3.7 The “Cause” of Errors: Items or Subjects?

When a set of responses to items by subjects does not conform to a perfect Guttman scale, we have until now followed a procedure in which we “blame” the items: The items are not good enough indicators of the same latent trait. But in Guttman’s model, an error is simply a violation of the expected relationship between a subject and a pair of items. So the error could just as well be attributed to subjects who march to a different drummer or for whom the questions mean something different.

If we want a measurement instrument that can be used for different groups of subjects, at different time periods, or in different experimental conditions, then we would rather have a scale in which the items function in the same way for all subjects. If necessary, we may discard the items that function differently and work with the maximal subset of items that are still useful. These items can be considered as still sufficiently prototypical indicators for the latent variable.

It is more difficult to justify discarding subjects and working only with a maximal subset of subjects. In that case the researcher faces the problem of generalizing from the remaining sample to a larger population. Often the initial sample was drawn from a well-determined population with a well-specified procedure. If we now delete a number of subjects who do not fit the scale, the remaining sample may no longer represent the original population well. So it is generally preferable to remove items rather than discarding subjects. Nevertheless, there are occasions when one would like to identify and possibly remove the deviant subjects. Let us look at a method for doing this.

## 3.8 "Blaming" the Subjects: Transposing the Data Matrix and Calculating Subject Homogeneities

One way of determining which subjects give rise to most model violations is simply to count the number of errors against the Guttman scale in each response pattern. But another way takes advantage of the fact that in a perfect Guttman scale the roles of items and subjects are entirely symmetrical, so they can be reversed. Thus, we can simply transpose the matrix, that is, interchange the rows and the columns, as shown in Table 3.7a and 3.7b. In the case of a perfect Guttman scale (Table 3.7a), we get perfect response patterns regardless of whether we show our data matrix with subjects as rows and items as columns, as is usual, or with subjects as columns and items as rows, In both cases we see a lower triangle of 1s and an upper triangle of 0s.

The symmetry between items and subjects underscores that we can evaluate subjects by their individual subject homogeneity as well. A homogeneity coefficient $\mathbf{H}^{\mathrm{T}}$ can be calculated for the whole dataset, for individual subjects $\mathbf{H}^{\mathrm{T}}(s)$, and for each pair of subjects $\mathbf{H}^{\mathrm{T}}(st)$ for subjects *s* and *t*, in the same way as the original homogeneity coefficients. The major difference is that the calculation of item homogeneities generally is done over hundreds of subjects, but the calculation of subject homogeneities only over some 5 to 20 items. This means that the estimates of these subject homogeneities have to be taken with a grain of salt.

Meijer (1994) has shown that the number of Guttman errors is a simple and powerful person-fit statistic and that it compares well with other alternatives, such as the subject homogeneity. But establishing subject homogeneity is also useful for the evaluation of probabilistic models, to which we will turn later in this monograph.

| *Data matrix* | | | | | |
|---|---|---|---|---|---|
| | A | B | C | D | E |
| 1 | 0 | 0 | 0 | 0 | 0 |
| 2 | 1 | 0 | 0 | 0 | 0 |
| 3 | 1 | 1 | 0 | 0 | 0 |
| 4 | 1 | 1 | 1 | 0 | 0 |
| 5 | 1 | 1 | 1 | 1 | 0 |
| 6 | 1 | 1 | 1 | 1 | 1 |

**Table 3.7a** Perfect Guttman scale with subjects 1–6 as rows and items A–E as columns.

| *Transposed data matrix* | | | | | | |
|---|---|---|---|---|---|---|
| | 6 | 5 | 4 | 3 | 2 | 1 |
| E | 1 | 0 | 0 | 0 | 0 | 0 |
| D | 1 | 1 | 0 | 0 | 0 | 0 |
| C | 1 | 1 | 1 | 0 | 0 | 0 |
| B | 1 | 1 | 1 | 1 | 0 | 0 |
| A | 1 | 1 | 1 | 1 | 1 | 0 |

**Table 3.7b** Perfect Guttman scale with items E–A as rows and subjects 6–1 as columns.

## 3.9 Using Imperfect Patterns to Measure Subject Scale Values

Can we use response patterns that contain errors for the measurement of a subject? This question can be answered with “yes” if we can still assume that the set of items forms a cumulative scale and that the subject understood the questions in the same way as the other people we have tried to measure. If we can use each response pattern to measure a subject, regardless of the number of errors he makes, how do we obtain this measurement?

Since we do *not* define the number of errors in a response pattern on the basis of the number of changes needed to make the pattern perfect, we cannot assign to him the scale value of the nearest perfect pattern. We have also seen that, for the response pattern ABCD,1101, for example, the nearest perfect pattern cannot be assigned unambiguously; it could be either 2 or 4. We therefore opt for defining *the scale value of a subject* who gives a response pattern that contains errors simply as *the number of items to which he has given the positive response.*[3] So subjects with the response pattern ABCD,1101 get 3 as their scale value.

## 3.10 Conclusion

In this chapter we have discussed how to evaluate a dataset as a cumulative scale, using Loevinger's coefficients of homogeneity. We have also discussed how to measure subjects and items: subjects by the rank order of their scale scores and items by their order of popularity in the dataset. The next chapter is devoted to the question of how to find subsets of items that form a cumulative scale, if the whole set of all items does not, and whether to discard items or subjects in that case.

[3] As Mokken (1971, pp. 140–141) demonstrated, this definition of the manifest scale score of subjects correlates highly with their true latent scale score.