# Chapter 4
# Reliability and Validity in Classical Test Theory


Chapters 1 and 2 referred to tests and questionnaires used in assessments. The concept of a trait was elaborated, and it was stressed that a trait was latent and not observed directly. It was also stressed that items of an instrument were intended to manifest the trait to be assessed.

The items of an instrument are said to *operationalize* a latent trait. In some cases, it is said that they provide an *operational definition* of the trait. The items make explicit the trait that is engaged with the persons and, through the responses, provide evidence of the degree of the trait through this engagement. A test of proficiency, for example, elicits behaviours that are supposed to count as evidence of the degree of proficiency of a person on the trait. In this case, the items provide the *instances* of the kinds of things that a person taking the test is expected to know, understand, interpret, be able to perform and the like.

In addition to obtaining information about the test-taker, because the items provide an operational definition of a trait, they also contribute to an empirical check of the understanding of the trait. They do this because the performances of the test-takers provide empirical evidence as to whether or not the items have worked together as expected. In summary, the test results provide empirical evidence of the theory of the construct in its context, including the administration of the instrument, the format of the items, and so on. This chapter elaborates the idea of understanding the trait through the empirical evidence from an instrument.

There are two major aspects of the evidence that need to be considered for an instrument; one is *reliability*, already broached in the previous chapter, the other is *validity* which we introduce in this chapter. Evidence on both of these is provided in part by its internal consistency, and by its consistency with expectation. Early texts on educational and psychological measurement that emphasize CTT include material on reliability and validity. In some recent writing on social measurement, the ideas of reliability and validity, and these terms, are not emphasized as such. However, whatever the terminology and fashions, the ideas of reliability and validity are central to any instrument. Messick (1989), Frisbie (1988), and Traub and Rowley (1991) help appreciate the way reliability and validity are presented in the literature. Although they are not recent papers, the points they raise are enduring.



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_4

## Validity

As in Chap. 1 where we described Stevens' terminology in the kinds of applications of numbers in social measurement, we consider some traditional terms in the elaboration of validity. In both cases, the terminology is introduced in part for historical reasons and in part because it has become embedded in the literature in social measurement. We wish to relate these terms and concepts in the literature, with which the reader will inevitably become engaged, to the perspective taken in this book. Traditionally, validity is articulated in terms of the following four ideas: *content validity, concurrent validity, predictive validity* and *construct validity*.

*Content validity* is established by experts judging whether the content was relevant, that is, by considering the operational definition of the trait. For example, in an examination for medical practitioners in some aspect of biology, experts in medicine would attest to the relevance of the content.

*Concurrent validity* is established by showing that the results on a particular instrument were related in an expected way with results on other relevant instruments. For example, a test for aptitude in mathematics would be expected to be related to performance in mathematics, but not to performance in sports. This does not mean that some sportsmen or women are not also excellent in mathematics, but that in general, the two are not related.

*Predictive validity* is established by relating the results of an instrument with performances in the future on the same trait. For example, performances on an entrance examination to university would be expected to be related to the performances of students during their studies at university.

*Construct validity* is established by demonstrating that the results on the instrument are consistent with expectations from a theoretical understanding of the trait. These expectations can be demonstrated in a variety of ways.

Messick (1989) argued for *construct validity* to be the overarching concept and that the other three so-called forms of validity are kinds of evidence for construct validity. Thus, *validity* is taken to be identical to *construct* validity and one uses whatever evidence one can to establish this kind of validity. The paper by Messick (1989) follows many papers on various aspects of validity in the traditional literature. We take Messick's position in this book and indeed expand upon it. As you will see, this position is consistent with Rasch measurement theory (RMT), where every piece of evidence examined that relates responses to an instrument with analysis according to the Rasch model is taken to provide evidence for or against construct validity of an instrument.

In Chap. 1, we noted that the terms *property*, *trait*, *construct*, *attribute* and *variable* were used more or less interchangeably, each having its own nuances and often more appropriate than another term in different contexts. At this point, we may elaborate on the term *construct*. It is used in three forms in social measurement:

(i) as a verb, *to construct*;
(ii) as a noun, *a construct*;
(iii) as an adjective to describe validity, *construct* validity.

An essential aspect of the use of *construct* is to emphasize that the trait, property or attribute is assessed in a social context and, at least in part, socially constructed. For example, the trait of neuroticism is conceptualized and formulated as useful by humans in trying to understand certain patterns of behaviour. Thus, *neuroticism* be referred to as a construct. Instruments assessing neuroticism have been constructed using indicators considered to be manifestations of neuroticism, which in turn provide an operational definition of neuroticism. Before application, these instruments need to have been validated in the various ways summarized above, and in other ways to be examined in this book. The same argument for construct validation of instruments holds for all social measurements.

## Reliability

We introduced the concept of reliability in CTT in Chap. 3. Reliability concerns the *consistency* of measurement. Traub and Rowley (1991) use the everyday life example of a car to explain the concept of reliability. Whether a car is a reliable starter or not can only be determined after repetitions of starting the car.

CTT focuses, as we have seen, on the test score as a whole. Therefore, one formalization of reliability is based directly on the idea of two parallel forms of an instrument. The *correlation* between the performances of the same persons on the two forms gives an estimate of the reliability of the test. We keep in mind that instruments are generally composed of multiple items.

In practice, one could actually construct two parallel tests. However, there are different conceptualizations of *parallel* which are relevant operationally. To understand at least two of these, an understanding also relevant when studying RMT, it is helpful to appreciate a distinction between the *identity* of an item and the relevant *property* of an item. This distinction is illustrated easily with assessment of proficiency with dichotomously scored items. The identity of the item is its content, format, marking key and so on. The relevant *property* of an item in this case is its *difficulty*. Two items from the same instrument generally have a different identity. However, they may have the same difficulty.

Thus, one conception of two parallel forms is to consider that all items of the two forms, all of which have different identities, are equally valid for the assessment of the trait, and where the average difficulty and standard deviation of the difficulties of the items are the same in the two forms. Another conception is to have matched pairs of items as similar in identity as possible (content, format, marking key) and similar in difficulty in the two forms. In general, because it is more flexible than the latter, we take the former conception. However, there are instruments that have parallel forms built on the latter principle.

The concept of a set of items which are equally valid in assessing a variable has important implications for the construction of instruments. In principle, it implies that any particular set of items is a sample of items from a whole *class* of items. Sometimes such a class is referred to as a *population* of items or as a *universe* of

items. We refer to it as a class of items. The class of items is hypothetical and infinite in the sense that all items of the class can never be listed. Thus, no matter how many items are constructed, another item (with another identity), can in principle be constructed. In addition, for example, if two different experts were given the specifications for constructing an instrument to assess some trait, they are likely to come up with items of different identities but a similar distribution of difficulties. In this sense the items are exchangeable. In the last section of this chapter we qualify the concept of item exchangeability and see that, as with many aspects of measurement, it is relative to a context.

The idea of a class of items with different identities making up parallel forms of a test, where the different forms, even if administered to the same persons, will give somewhat different results, implies errors of assessment. Errors of assessment, when transformed into measurements, imply *errors of measurement*. The presence of errors results in the observed correlation between the measurements from the two parallel forms to be less than 1.0. The greater the deviation from 1.0, the greater the size of the error.

The conception of error here is that it is random. Being random means that there is no pattern to it, and that it is not correlated with any other feature of the assessment. It also implies that the errors tend to cancel each other out. Systematic deviations which impinge on the reliability and validity of assessment generally are given different names. For example, if human markers are part of the assessment, in assessing writing proficiency or in assessing the functioning of limbs, it is possible and indeed likely that different markers will have some systematic relevant differences, perhaps in their harshness or leniency. The presence of such human errors in assessment, and understanding their presence, was part of the history of the development of statistics. Alder (2002) gives an instructive account of this development. Kane (2011) summarizes conceptions of errors in social measurement in general, and educational measurement in particular.

Although the idea of two parallel tests is useful, and in some cases, some instruments have parallel forms, *repeated assessments* from the same instrument on one person are not in general feasible in social measurement. How, then, is the reliability of an instrument in this case established?

The procedure that is most popular and efficient rests on the idea that responses to multiple items within a single instrument are themselves replications. Taking responses to multiple items as replications leads to a calculation of a reliability index under certain assumptions. We proceed with the estimation of the reliability with this conception of items as replications, and then return to consider more closely these assumptions. The assumptions are implied by the formulation in terms of equations.

The general index of reliability calculated this way is known as coefficient $\alpha$. We first express the reliability of a test, the ratio of true score variance to the total variance, in terms of items as replications.

## Reliability in Terms of Items

We now explicate the definition of reliability in CTT in terms of items as replications. This explication helps provide, first consolidation of how the items are viewed in CTT, and second how this contrasts with the modern test theory approach. The results are entirely consistent with the traditional approach to CTT which focuses on the tests and gives the same formula for calculating a reliability index, which we describe later in the chapter.

As indicated above, first we consider that each item of an instrument is a replication of every other item. Thus, the items are considered as a random selection from some class of items that assess the particular trait and that are relevant to administer to some population of persons. Of course, the items would be administered only to a sample of the population. Each item also assesses some unique aspect of the trait and there is some error. Because the focus is on assessing the single trait, the unique aspect is embedded in the error. The implication of this embedding of the unique aspect in the error can be further considered, and we do so in a later chapter.

With the same definitions of variables as in Chap. 3, we let the observed score $x_{ni}$ of person $n$ to item $i$ also be composed of the sum of a true score and an error score. We denote the true score referenced to the item by the Greek letter $\tau$, giving $\tau_n$ as person $n$'s true score, and denote the error at the item level by the Greek letter $\varepsilon$, giving $\varepsilon_{ni}$ as the error when person $n$ responds to item $i$.

Then, again taking that the observed score is the sum of the true score and error score, gives

$$x_{ni} = \tau_n + \varepsilon_{ni}. \tag{4.1}$$

We postulate the following conditions:

(i) Just as with the errors at the test level, the error scores of persons are uncorrelated with their true scores.
(ii) The error scores across persons and items sum to 0.
(iii) The errors across all person item combinations are homogeneous, which is identical to postulating that the error variances are equal. We denote this variance, defined below as $s_\varepsilon^2$.

Now consider the test score $y_n$ in terms of Eq. (4.1): $y_n = \sum_{i=1}^{I} x_{ni} = \sum_{i=1}^{I} (\tau_n + \varepsilon_{ni})$.

Expanding,

$$\begin{aligned} y_n &= \sum_{i=1}^{I} (\tau_n + \varepsilon_{ni}) \\ &= \sum_{i=1}^{I} \tau_n + \sum_{i=1}^{I} \varepsilon_{ni} \end{aligned}$$

$$= I\tau_n + \sum_{i=1}^{I} \varepsilon_{ni}, \tag{4.2}$$

we cannot simplify $\sum_{i=1}^{I} \varepsilon_{ni}$ because although the variances are the same across person–item combinations, each actual person–item combination has a unique error, thus giving a unique sum. However, from Eq. (3.1) in Chap. 3, we have that $y_n = t_n + e_n$. Therefore, we can identify

$$t_n = I\tau_n \text{ and } e_n = \sum_{i=1}^{I} \varepsilon_{ni}.$$

It may seem odd that the true score $t_n$, as it is traditionally written, is a function of the number of items. It may seem more natural to divide the person scores by the number of items, so that the true score is not a function of the number of items. However, the way it is written in CTT means that the true score is on the same kind of scale as the observed scores. For example, if the observed scores range between 0 and 50, then the true score will in principle have the same range, and this has some convenience. This apparent advantage implies that the items of an instrument are fixed. This is not consistent with other conceptions where items are not unique, but are a sample from a class of items, and where in principle, even different numbers of items might be present in an instrument.

We now proceed to express the reliability using Eq. (4.1) and see that it illuminates the relationship between reliability and the number of items.

The variance of the observed scores from Eq. (4.2) is then given by

$$s_y^2 = I^2 s_\tau^2 + I s_\varepsilon^2. \tag{4.3}$$

Therefore, from Eq. (3.1) in Chap. 3, we can identify

$$s_t^2 = I^2 s_\tau^2 \text{ and } s_e^2 = I s_\varepsilon^2.$$

From Eq. (3.3) in Chap. 3, we have the definition of reliability as

$$r_{yy} = \frac{s_t^2}{s_y^2}. \tag{4.4}$$

Substituting the variances from Eq. (4.3) into Eq. (4.4) gives

$$\begin{aligned} r_{yy} = \frac{s_t^2}{s_y^2} &= \frac{I^2 s_\tau^2}{I^2 s_\tau^2 + I s_\varepsilon^2} \\ &= \frac{s_\tau^2}{s_\tau^2 + s_\varepsilon^2 / I}. \end{aligned} \tag{4.5}$$

Thus, the reliability is also a ratio of the true variance relative to the total variance at the item level, but we can now see the relationship of reliability to the number of items. As the number of items $I$ increases, so the error variance term $s_\varepsilon^2/I$ decreases and the reliability increases. Because the reliability is constrained between 0 and 1, this relationship is not linear.

## Coefficient Alpha ($\alpha$): Estimating Reliability in CTT

As indicated above, the estimate of reliability we consider is provided by coefficient $\alpha$. In proceeding with items directly, rather than from total scores, the way we derive the equation for coefficient $\alpha$ is slightly different from its usual development. This coefficient was developed by Guttman (1945) and elaborated substantially by Cronbach (1951). This elaboration was so well received that the index is also known simply as Cronbach's $\alpha$. Coefficient $\alpha$ can be applied to tests composed of items with different maximum scores. The equation can be specialized to the case where all items are scored dichotomously. This specialization was derived earlier by Kuder and Richardson (1973) and so coefficient alpha can be seen as a generalization of the Kuder–Richardson formula.

Both Kuder and Richardson, and Guttman, had various equations in their papers, and their formulae now have the name of the equation in their original papers. Kuder and Richardson's most common formula is their formula 21 often referred to simply as KR-21. Guttman's equation was the first and he used Greek letters to name them, hence coefficient $\alpha$.

Our development of the equation for coefficient $\alpha$ is provided in *Part IV* of this book. It takes the form

$$\alpha = \frac{I}{I-1}\left(\frac{s_y^2 - \sum_{i=1}^{I} s_i^2}{s_y^2}\right), \tag{4.6}$$

where the variances of the total scores and the items, $s_y^2$ and $s_i^2$, are calculated simply as $s_y^2 = \frac{1}{N-1}\sum_{n=1}^{N}(y_n - \bar{y})^2$ and $s_i^2 = \frac{1}{N-1}\sum_{n=1}^{N}(x_n - \bar{x})^2$ and where $N$ is the number of persons in the sample.

In coefficient $\alpha$, the idea that the items are replications of each other is explicit in the sense that any subset of the class of items is parallel to any other, with different numbers of items simply affecting the reliability. Because the reliability is calculated from responses of a single administration of an instrument and is based on the items within the instrument, this form of reliability is known as *internal consistency*.

We now show the application of the general formula of Eq. (4.6) above. It can be used for both dichotomous and polytomous items with variable maximum scores.

**Table 4.1** Calculating the reliability for the data in Table 3.1 of Chap. 3

| Item | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | Total score |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Variance | 0.08 | 0.04 | 0.11 | 0.12 | 0.08 | 0.75 | 0.21 | 0.22 | 0.58 | 1.51 | 6.42 |

## Example

To see how the coefficient $\alpha$ equation is applied, consider the data from Table 3.1 in Chap. 3. In computing the variance in each case we continue dividing by $N-1$. Table 4.1 shows the variance of each of the 10 items as well as the variance of the total score.

Substituting the values of each $s_i^2$ and $s_y^2$ gives$= 6.42\ (0.53) = 3.40$.

$$\begin{aligned}\alpha &= \frac{I}{I-1}\left(\frac{s_y^2 - \sum_{i=1}^{I} s_i^2}{s_y^2}\right)\\ &= \frac{10}{9}\left(\frac{6.42 - (0.08 + 0.04 + 0.11 + 0.12 + 0.08 + 0.75 + 0.21 + 0.22 + 0.58 + 1.51)}{6.42}\right)\\ &= \frac{10}{9}\left(\frac{6.42 - 3.69}{6.42}\right)\\ &= (1.1)(0.43) = 0.47.\end{aligned}$$

This is quite a low value; however, the test is not very long. We can calculate the error variance from Eq. (3.5) of Chap. 3 as $s_e = s_y\sqrt{1 - r_{yy}} = 6.42(0.53) = 3.40$.

There is no absolute standard in interpreting reliability coefficients. When decisions about individuals are made, the reliability needs to be greater than when decisions about groups are made. When decisions about *groups* are made a coefficient of at least 0.65 is recommended (Traub & Rowley, 1991).

## Factors Affecting the Reliability Index

Having expressed the equation for calculating reliability and having used it in an example, makes it opportune to consider some of the factors that affect the value of reliability.

The factors can be considered *internal* and *external* to the instrument, though because the responses arise from the engagement of the person to the items of the instrument, in principle all factors are always related to each other in some sense. The explication as internal or external is made for purposes of exposition.

## *Internal Factors*

### Number of Items

As indicated in Eq. (4.5), all other factors being equal, the reliability is clearly affected by the number of items. The relationship is not linear, but it is possible to express the reliability of a greater or smaller number, assuming all other factors are the same, given the reliability of a particular number of items. For example, it can be derived readily from Eq. (4.5) that if the number of items is doubled, then the new reliability, denoted say $r^*_{yy}$ is given by

$$r^*_{yy} = \frac{2r_{yy}}{1 + r_{yy}}, \tag{4.7}$$

where $r_{yy}$ is the original reliability. Equation (4.7) is known as the Spearman–Brown formula. It can be generalized for any factor of the number of original items.

### Discrimination of Items

We have already considered the discrimination of the items. The greater the discrimination of the items, as defined in Chap. 3, the greater the reliability. Note that we have used the plural of items here. The reason is that the basic equation of CTT expressed in terms of items, Eq. (4.1), together with the three conditions, implies that all the items have the same discrimination. They imply the same discrimination because the error variance is assumed to be the same magnitude for all items. In observed data, of course, they will not have the same discrimination. In the method of assessing discrimination that we have used above, and in using the item–total correlation, the check is whether each item is discriminating very much like the average discrimination of all of the items.

### Independence of Items

Another implication of Eq. (4.1), which we have broached, is that each of the items is a replication of each other item. We have, in deriving the expression for the variances, also assumed independence of the responses. This implies, for example, that one item does not artificially relate to any other item. An example of the violation of independence in tests of proficiency is when the answer to one item implies, or gives a clue to, the answer to another item (Mehrens & Lehman, 1991).

### Unidimensionality Among Items

The use of the total score as a summary of a person's value on the variable implies that persons can be compared by their total scores, or the estimates of their true scores, and this implies a unidimensional construct. In addition, because each person has a single true score, which can vary in value from the true scores of other persons, Eq. (4.1) also implies a unidimensional construct. Unidimensionality can be violated if some items assess, in some systematic way, a different construct. An example in tests of proficiency is when the majority of items assess proficiency in mathematics using very little verbal description, and some items involve a large amount of complicated written expression. It may be a surprise that within such a data set, the actual value of $\alpha$ is inflated relative to what it would be had there been the same number of items and they were all assessing mathematics proficiency in the same way as the other items with the written complexity.

## *External Factors*

### Variance of the True Scores in the Sample

It is evident from Eq. (4.5) that the reliability will be a function of the true score variance in the population. It also depends on the sample being a representative sample from the population for the index to be referenced to that population. Thus, the reliability of an instrument is not simply a property of the instrument, but is related to the population of persons to which it is referenced. Thus, if all factors are equal, but for some reason, one sample of persons has a smaller true variance than another one, which could occur by chance, then the sample with the smaller variance will provide a smaller reliability.

### Alignment of the Persons to the Items

A critical implication of Eq. (4.1) is that the persons are well aligned to the items. This means that the persons are not likely to all obtain the same score on the items. If persons are not well aligned to the items, for example, in a proficiency test all students find the test very easy and many have the maximum score, then the variance of the observed scores will be truncated artificially and the instrument will have a lower reliability than otherwise. The same effect would arise if many persons obtained the minimum score of zero. The former effect is known as a *ceiling* effect and the latter as a *floor* effect. We consider further each of these features affecting the reliability, and therefore the precision, throughout the book as the opportunity arises.

## Common Factors Affecting Reliability and Validity

Crucial to measurement is the quality of the engagement of the persons to the items. In proficiency assessment, it is important that students are in a position to answer the questions that they can answer, that difficult questions are not at the beginning of the test and easy ones at the end, that students are prepared and know the format of the assessment, and so on. In the case where assessors are involved, as they are in clinical psychological assessment and health assessment, as well as in some achievement assessment, the quality of the assessor is critical. Poor instructions, poorly understood and applied instructions, confusing marking keys and weak training of the assessors will add to random error, lower discrimination of items and lower reliability.

In considerations of both reliability and validity, the concern is with potential inferences about future data, or future observations, given the available data or available observations. If we have a reliable instrument we would expect that a replicated assessment with a similar population would give similar results. If we have a valid instrument we would expect that in the relevant circumstances we could predict performances of the persons measured with it. It is relevant to note that persons can and do change on a trait as a result of natural growth, teaching, rehabilitation and so on. In the index of reliability above, the evidence provided is its reliability at a single administration assuming that during the administration the person's true score is constant.

It is generally emphasized that a high reliability is necessary for validity of an instrument. However, it is possible for high reliability to be contrived artificially and that the high reliability is obtained at the expense of validity. Such a situation can be envisaged readily in the assessment of attitude. For example, suppose that the different questions in appearance are in fact the same substantive question but with different wording. Then, unless the persons get bored and do not engage with the items as intended, a very high value for coefficient $\alpha$ might be obtained. However, this would be at the expense of validity. Within CTT, such a situation is known as the *attenuation paradox*.

## Causal and Index Variables

In the above discussion, we indicated that the idea of parallel forms of items meant that in principle the items were exchangeable. They would be exchangeable in the sense that many thermometers are exchangeable to measure a temperature in some situation, say the temperature of a cellar. However, even thermometers are not all exchangeable in all circumstances. The thermometer for measuring the temperature of a cellar might range from $-10$ to 50 °C. Clearly, such a thermometer would not be useful to measure temperatures to $-20$ °C or to 60 °C. This case is analogous to not having items that are far too easy or far too difficult for students in proficiency

assessment. In principle, other than practical factors, thermometers are exchangeable because they measure the same variable.

However, CTT and RMT are not only applied where in principle the items are not exchangeable. The summary discussion on this topic concerns *causal* or *reflective* versus *index* or *formative* constructs. This distinction is discussed in greater detail in Andrich (2014), Stenner, Stone, & Burdick (2009), and Tesio (2014).

Briefly, in a causal construct, where the items assessing proficiency are in principle exchangeable, a student's measure on the construct governs their response to all items. For example in a test of the construct of light in physics for a specified curriculum and class level, many different items which are in principle exchangeable might have been written. In the case of an *index* construct, the items help define the variable and so are not in principle exchangeable. An example is given by Stenner et al. (2009) in which socio-economic status (SES) is defined by the items 'level of education', 'occupational prestige', 'level of income' and the 'desirability of the neighbourhood in which people live'. The score on these items will be correlated positively in most populations and it might be justified to sum them to provide a single number to characterize SES. However, there is a sense in which if one of these items were removed from the set, then the definition of the variable of SES is changed.

Most assessments are some combination of the construct being causal and index. For example, in educational assessment, Andrich (2014) gives the example of a test in physics which assesses not only the topic of light, but those of heat, sound, electricity and magnetism, and mechanics. In that case, the items testing the knowledge of the topic of sound are not exchangeable for the items assessing the knowledge of the topic of heat, for example. Tesio (2014) gives examples in health outcomes assessment which from some perspectives are causal and from others are index. The perspective from which an item is considered contributes to its selection in instruments and how it is dealt with if it happens not to work as well as desired.

## Exercises

In the Exercises of Chap. 3, you were given a table of person–item responses.

1. Calculate the variance of each of the eight items in the test and the total score and summarize them as below:

| $s_1^2$ | $s_2^2$ | $s_3^2$ | $s_4^2$ | $s_5^2$ | $s_6^2$ | $s_7^2$ | $s_8^2$ | $s_y^2$ |
|---|---|---|---|---|---|---|---|---|
| | | | | | | | | |

2. Calculate the reliability of this test according to coefficient $\alpha$. Show your working. Use the variances of the eight items and the variance of the total score that you calculated in question 1.
3. Comment on the size of the reliability.

4. Consider a test or examination with which you are familiar with. Describe the test and its purposes first, then comment on the reliability of the examination and the validity in terms of the various functions the examination is supposed to serve. How might these be investigated?

For further exercises, see *Exercise 1: Interpretation of RUMM2030 printout* in Appendix C.

## References

Alder, K. (2002). *The measure of all things: The seven-year odyssey and hidden error that transformed the world*. New York: Free Press.

Andrich, D. (2014). A structure of index and causal variables. *Rasch Measurement Transactions, 28*(3), 1475–1477.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16,* 297–334.

Frisbie, D. A. (1988). Reliability of scores from teacher-made tests. *Educational Measurement: Issues and Practices. National Council on Measurement in Education, 7*(1), 25–35.

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika, 10,* 255–282.

Kane, M. (2011). The error of our ways. *Journal of Educational Measurement, 48*(1), 12–30.

Kuder, G. F., & Richardson, M. W. (1973). The theory of the estimation of test reliability. *Psychometrika, 2,* 151–160.

Mehrens, W. A., & Lehman, I. J. (1991). *Measurement and evaluation in education and psychology* (4th ed.). New York: Harcourt Brace.

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher, 18*(2), 5–11.

Stenner, A. J., Stone, M. H., & Burdick, D. S. (2009). Indexing versus measuring. *Rasch Measurement Transactions, 22*(4), 1176–1177.

Tesio, L. (2014). Causing and being caused: Items in a questionnaire may play a different role, depending on the complexity of the variable. *Rasch Measurement Transactions, 28*(1), 1454–1456.

Traub, R. E., & Rowley, G. L. (1991). Understanding reliability. *Educational Measurement: Issues and Practices. National Council on Measurement Education, 10*(1), 37–45.

## Further Reading

Andrich, D. (1988). *Rasch models for measurement* (pp. 84–86). Newbury Park, CA: Sage.

Andrich, D. (2016). Components of variance of scales with a bi-factor structure from two calculations of coefficient alpha. *Educational Measurement: Issues and Practice, 35*(4), 25–30.

Roscoe, J. T. (1975). *Fundamental research statistics for the behavioral sciences* (2nd ed.). New York: Holt, Reinhart and Winston.