



Modern Psychometrics

The Science of Psychological Assessment

Fourth Edition

John Rust, Michal Kosinski, and David Stillwell

Modern Psychometrics

This popular text introduces the reader to all aspects of psychometric assessment, including its history, the construction and administration of traditional tests, and the latest techniques for psychometric assessment online.

Rust, Kosinski, and Stillwell begin with a comprehensive introduction to the increased sophistication in psychometric methods and regulation that took place during the 20th century, including the many benefits to governments, businesses, and customers. In this new edition, the authors explore the increasing influence of the internet, wherein everything we do on the internet is available for psychometric analysis, often by AI systems operating at scale and in real time. The intended and unintended consequences of this paradigm shift are examined in detail, and key controversies, such as privacy and the psychographic microtargeting of online messages, are addressed. Furthermore, this new edition includes brand-new chapters on item response theory, computer adaptive testing, and the psychometric analysis of the digital traces we all leave online.

Modern Psychometrics combines an up-to-date scientific approach with full consideration of the political and ethical issues involved in the implementation of psychometric testing in today's society. It will be invaluable to both undergraduate and postgraduate students, as well as practitioners who are seeking an introduction to modern psychometric methods.

John Rust is the founder of The Psychometrics Centre at the University of Cambridge, UK. He is a Senior Member of Darwin College, UK, and an Associate Fellow of the Leverhulme Centre for the Future of Intelligence, University of Cambridge, UK.

Michal Kosinski is an Associate Professor of organizational behavior at the Stanford Graduate School of Business, USA.

David Stillwell is the Academic Director of the Psychometrics Centre at the University of Cambridge, UK. He is also a reader in computational social science at the Cambridge Judge Business School, UK.



Taylor & Francis

Taylor & Francis Group

<http://taylorandfrancis.com>

4 Psychometric measurement

In the early 19th century, the physicist William Thomson, also known as Lord Kelvin—arch-measurer and originator of the Kelvin scale of absolute temperature—opined, “Whatever exists at all exists in some amount, and can therefore be measured.” (Thomson, 1891) Toward the end of that century and somewhat in his dotage, he made the additional claim that “there is nothing new to be discovered in physics now. All that remains is more and more precise measurement.” Of course, after quantum theory, we now know that he was wrong. But no one can deny that more and more precise measurement of, for example, the speed of light or even time itself has played a key part in scientific advancement. Could the same be true for psychometric measurement?

In this chapter, we address several of the mathematical concepts central to psychometrics. These come from several sources:

- 1 True-score theory, originally conceived to make sense of issues underlying disagreement between examiners in the awarding of school grades
- 2 Factor analysis
- 3 Vector algebra, a mathematical technique designed to understand physical systems involving both force and direction, such as gravity
- 4 Functional psychometrics, the black box, and explainability

Despite the huge advances made in the 20th century, psychological measurement was not without its detractors. Originally most, but not all, came from those with an ideological predisposition to oppose testing in schools. Others were more persuaded by a functional “black box” approach, today much favored within machine learning. These proposed alternatives, such as criterion-referenced testing and predictive analytics, also had their impact on psychological assessment. Modern psychometrics is a synergy of all these ideas, interests, and methods.

True-score theory

True-score theory was at first an attempt to introduce more scientific rigor to grading exams. Essay graders often disagreed with each other, and there was confusion about what a grader’s mark on an essay actually meant. What was it a measure of? In 1888, Francis Ysidro Edgeworth thought he had an answer. Each examiner was aspiring to obtain a mark that represented the candidates’ “true score” on the essay, but with varying levels of success. He wrote:

“If we tabulate the marks given by the different examiners they will tend to be disposed after the fashion of a gendarme’s hat. I think it is intelligible to speak of the mean judgment of competent critics as the true judgment, and deviations from that mean as errors. This central figure which is, or may be supposed to be, assigned by the greatest number of equally competent judges, is to be regarded as the true value, just as the true weight of a body is determined by taking the mean of several discrepant measurements.”

This is the first known definition of what is now called true-score theory, which later became known as latent-trait theory. It is fundamental to classical test theory. It states simply that any score on an item or a test by a respondent can be represented by two component parts: the respondent’s true score on the item or test and some error of measurement. This is traditionally stated as

$$X = T + E,$$

where X symbolizes the observed score, T the true score, and E the error. If all one knows about a test is that a particular respondent obtained a score of X , then one knows nothing at all. In these circumstances, the error and true score are inextricably mixed. For example, X may be 5, yet this could be the case if $T = 3$ and $E = 2$ or equally if $T = 110$ and $E = -105$. Thus, an observed score X on its own is of no use whatsoever. It is the true score T that we are interested in, and we need additional data to estimate this; primarily, we need some idea of the expected size of the error term E . To put this another way, we cannot know how accurate a score is unless we have some idea of how inaccurate it is likely to be.

The theory of true scores takes us through various techniques for obtaining an estimate of the size of the error. This is typically done through the process of replication, either by obtaining several scores from the same respondent or by obtaining scores from many different respondents. To successfully estimate true scores from such data, three assumptions must be made.

First, all the errors terms E associated with observed scores X are assumed to be random and normally distributed. In the context of grading an exam, for example, graders are assumed to randomly overestimate or underestimate the true score T when giving their grades (or observed scores). This means that their errors E are random. Moreover, they are more likely to make small errors than large errors—or, in other words, their errors are normally distributed with a mean equal to 0. This is the same assumption made when we argue that when an unweighted coin is tossed a number of times, on each of which there is an equal chance of heads or tails, it is very unlikely that the coin will land the same way every time, more likely that about half the time it will land as a head and half the time as a tail. The normal curve is itself derived from this theory of random error, known as probability theory.

Second, true scores are assumed to be uncorrelated with the errors. That is, the distribution of errors is approximately the same, regardless of whether a high, medium, or low score has been observed. In the context of grading an exam, for example, this can be rather more problematic. There are circumstances under which this assumption fails, for example when too many of those in a sample obtain very high or very low scores, perhaps because the test is too easy or too difficult. But these deviations are all adjustable (in principle at least) by various algebraic transformations of the raw data.

Third, it is assumed that the observed scores X from the same respondent are statistically independent of each other. In the context of grading an exam, for example, if the second examiner had already seen the mark of the first examiner before making their own mark, then the two marks would not be statistically independent.

If the three assumptions of true-score theory are made, then a series of very simple equations falls into our lap. First, we can define the true score T statistically as an average of a very large number of observed scores X from the same person. As the number of observations approaches infinity, then the error terms E , being random, cancel each other out and leave us with a pure measure of the true score. Of course, it is not possible to take an infinite number of measures on the same person—or indeed, even more than a few such measures—without changing the measuring process itself because of the respondent's fatigue, practice effects, and so on. But this is unimportant from the point of view of the statistical definition, which states that the true score is the score that we would obtain *were* this possible. Second, from true-score theory and its assumptions, we are able to derive the amount of error and hence get an idea of a test's accuracy.

Although true-score theory has been widely criticized and many attempts have been made to improve it, the alternatives are generally complicated and usually turn out to have flaws of their own. For most of the last century, true-score theory continued to provide the backbone for psychometrics. While the assumptions of true-score theory can never be perfectly met, they are a good enough approximation in most situations and have stood the test of time.

Identification of latent traits with factor analysis

True-score theory introduced the idea of obtaining greater accuracy by increasing the amount of information available for the estimation of the true score, whether this be from more examiners, more respondents, or more items. In a multi-item test, each item is there because it is believed to be related to the same latent trait, and thus provides some information about the true score on this latent trait. The true score on a classical test is the sum of the true scores on each of its items. But for every item that is endorsed, there will be some degree of error—as well as, perhaps, something specific about that item that differentiates it from all the others. This was the insight that inspired Charles Spearman's discovery of factor analysis in 1905.

Spearman's two-factor theory

Spearman was addressing the problem of how to interpret a uniformity of structure that he observed in correlation matrices, such as those consisting of correlations between various intelligence subtests (verbal, numerical, and so on) that he believed could be combined to generate an overall score on an intelligence test, which he called general intelligence, or “g.” Each subtest will have a correlation with all the other subtests, so there can be a fair number of correlations involved. For example, with only five subtests, we have 10 correlations ($4 + 3 + 2 + 1$); with 20 subtests, it would be 180 ($19 + 18 + 17 + \dots + 1$). These can be tabulated in a matrix (see Table 4.1), where rows and columns represent subtests and the cells represent the correlation between the row subtests and column subtests. Such matrices can contain scores on different tests, grades on different subtests, or even scores on individual items.

Table 4.1 A correlation matrix, representing correlations between five subtests (a, b, c, d, and e) in a psychometric test

	("g")	a	b	c	d	e
("g")	(1.0)	(.9)	(.8)	(.7)	(.6)	(.5)
a	(.9)	(.81)	.72	.63	.54	.45
b	(.8)	.72	(.64)	.56	.48	.40
c	(.7)	.63	.56	(.49)	.42	.35
d	(.6)	.54	.48	.42	(.36)	.30
e	(.5)	.45	.40	.35	.30	(.45)

The example given in Table 4.1 illustrates his approach. Ignore for the moment all the figures in parentheses. First, Spearman arranged all the variables in what he called hierarchical order, with the variable that showed the highest general level of inter-correlation with other variables on the left, and the variable with the least correlation on the right. He then drew attention to an algebraic pattern in the relationship between the variables. He noted that the product of the correlation between subtests a and b and the correlation between subtests c and d tended to be equal to the product of the correlation between subtests a and c and the correlation between subtests b and d:

$$(r_{ab} \times r_{cd}) \approx (r_{ac} \times r_{bd}).$$

Moreover, he observed that this could be extended to all the sets of four subtests (tetrads); thus, for example,

$$\begin{aligned} r_{bc} \times r_{de} &\approx r_{be} \times r_{cd}; \\ r_{ac} \times r_{be} &\approx r_{ae} \times r_{bc}; \end{aligned}$$

and so on. He measured the extent to which this rule held as the "tetrad difference." If the four corners of the tetrad are called A, B, C, and D, then the tetrad difference is

$$AD - BC.$$

Thus, in Table 4.1, the tetrad differences are:

$$\begin{aligned} .72 \times .42 - .63 \times .48 &= 0; \quad .56 \times .30 - .42 \times .40 = 0; \quad \text{and} \\ .63 \times .40 - .56 \times .45 &= 0. \end{aligned}$$

Spearman noted that such a pattern of relationships would be expected if a, b, c, d, and e were subtests of intelligence and each represented a combination of two elements: general intelligence ("g") that contributed to each of the subtests and specific intelligence that was unique to each. Thus, if subtest a were a test of arithmetic, then the components of subtest a would be "g" and a specific ability in arithmetic. If subtest b were verbal ability, this would be composed of "g" and a specific verbal intelligence. He argued that it was the common existence of "g" in all subtests that caused the correlation. He called this his two-factor theory because each subtest was composed of two elements: "g" and something specific to that subtest. Each subtest was composed of scores on just two

factors. The general factor was common to all subtests, but the specific factors were all unique to each.

By including the parenthetical components in Table 4.1 within the calculation of the tetrad difference, he developed the first-ever technique for factor analysis. For example, if x is the unknown value where column a and row a cross in Table 4.1, it can be obtained from the tetrad difference formula:

$$x \times r_{bc} \approx r_{ab} \times r_{ac};$$

that is,

$$x \times .56 \approx .63 \times .72;$$

thus,

$$x \approx .81.$$

He called this value the saturation value of “ g ” on a , and deduced that the square root of this value would give the correlation of a with “ g ,” general intelligence. Thus, in Table 4.1, the column of figures under “ g ” represents the factor loadings of each of the five subtests on general intelligence. We see that subtest a is very highly saturated, while subtest f is less so.

Of course, the example in Table 4.1 is artificial. In practice, the arithmetic would never be this neat, and the tetrad differences will never come to exactly zero. However, we can calculate such differences, find the average of the estimated values for each saturation, and use this to estimate factor loadings on “ g .” Spearman took the process a step further. For his two-factor theory to be true, he used the loadings to estimate the values for each correlation. By comparing these values with the actual correlations, he could therefore find the goodness of fit of his theory. He was also able to subtract the observed correlations from the expected values and carry out the process again on the residuals, leading to the extraction of a second factor. This might occur if some of the specifics were not in fact unique. Spearman’s insight was brilliant, and it was many decades before statisticians were able to confirm his theory statistically and catch up with his intuition.

Vector algebra and factor rotation

Spearman achieved his factor-analytic technique using numbers alone. However, it was difficult to visualize when more factors were involved. Improved understanding became possible when graphical techniques were used to represent the data. These techniques (see Figure 4.1), which have produced visual representations of the process, have had a major impact on the development of psychometrics.

Visualizing the relationships between variables made the conceptualization of psychometric issues much easier. Graphical representation of ideas have emerged again and again in psychology, from multidimensional scaling in psychophysics to the interpretation of repertory grids in social and clinical psychology. They are fundamentally based on models provided by vector algebra in which two values are ascribed to a variable: force and direction. Within factor-analytic models, variables are represented by the force element of the vector, which is held constant at a value of 1, while the angle between two variables represents the correlation between them, in such a manner that the cosine

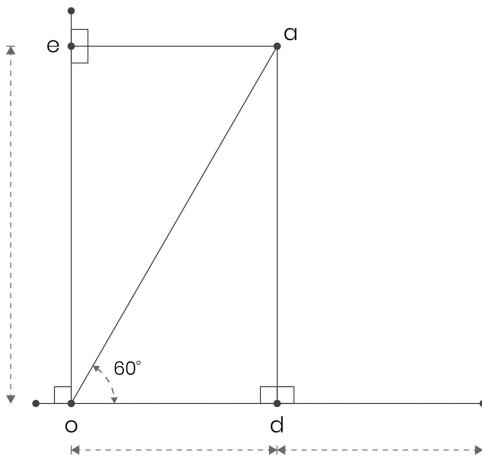


Figure 4.1 Spatial representation of the correlation coefficient. A correlation of 0.50 between two variables a and b can be graphically represented by two lines of the same length that have an angle between them whose cosine is 0.50 (60°).

of this angle is equal to the correlation coefficient. Thus, a correlation of .50 between the variables a and b (as in Figure 4.1) is represented by two lines oa and ob of equal length, with an angle between them whose cosine is .50, that is 60° . A correlation of .71 would be represented by an angle whose cosine is .71 (45°).

There are many useful characteristics that follow from this visual representation of the correlation. In Figure 4.1, we can see that one of the vectors, ob , is drawn horizontally, and the other, oa above it, is drawn at angle aob , equal to 60° . A perpendicular ad is then dropped onto ob from point a to point d . If we assume that ob and oa both have a length of 1, the distance od will then be equal to the cosine of the angle between the vectors, and therefore to the correlation itself. Also in the figure, we see that a vertical oe at a right angle to ob has been drawn and projected onto a horizontal line ae . Through Pythagoras, we know that as oa has a length of 1, then $od^2 + oe^2 = 1$. This gives us a graphical restatement of the statistical formula $r^2 + (1 - r)^2 = 1$, which tells us how we can use the correlation coefficient to partition variance. To give an example, if the correlation between age and a measure of reading ability is .50, then we can say that $.50^2$ —that is, .25 or 25%—of variance of reading ability is accounted for by age. It also follows that 75% of the variance in reading ability is not accounted for by age. This is represented graphically in the figure: the cosine of the angle between oa and ob (60°) is .50, and therefore the distance od is .5. What is the distance oe ? Well, its square is .75 ($1 - .25$ by Pythagoras), hence oe must be the square root of this, i.e., .87. This number represents a correlation, but it is a correlation between reading ability and some hypothetical variable, as no vector oe was originally given by the data. However, we can give a name to this variable, which is itself a latent trait; we could call it “that part of reading ability that is independent of age.” This value could be estimated by partial correlation analysis and used in experimental situations to eliminate age effects. While considering the graphical representation of correlations, think about two special cases. If $r = 0$, the variables are uncorrelated. The angle between oa and ob is 90° ; the cosine of 90° is 0. The variables are thus each represented by their own separate spatial dimensions; they are said to be

orthogonal. If $r = 1$, then the angle between oa and ob is 0° (the cosine of 0° being 1), and they merge into a single vector. Thus, the variables are graphically as well as statistically identical and are measures of the same underlying latent trait.

From this very simple conception, we can demonstrate a fundamental idea of factor analysis: while the two lines oa and ob represent real variables, there is an infinite number of hidden variables that can exist in the same space, represented by lines drawn in any direction from o . The hidden variable oe represents a latent variable: that part of oa that is independent of ob . There are many example applications of these models. If oa were, for example, humans' weight, and ob their height, then oe would be that part of the variation in the weight of human beings independent of height. It is thus a measure of obesity, not measured directly but by measuring weight and height and applying an appropriate algorithm. Another example relevant to psychometrics may have ob as the score on a psychological test and oa as a measure of the criterion against which it is to be validated; thus the angle between oa and ob , representing the correlation between a (the test score) and b (the criterion), becomes a measure of validity, and oe becomes the aspect of the criterion that is not measured by the test.

Moving into more dimensions

A flat piece of paper, being two-dimensional, can accurately represent at most two totally independent sources of variation in any one figure. To extend this model to factor analysis, we need to conceive not of one correlation but of a correlation matrix in which each variable is represented by a line of unit length from a common origin and the correlations between the variables are represented by the angles between the lines. Taking first the simple case of three variables x , y , and z (see Figure 4.2) and the three correlations between them, if the angle between ox and oy is 30° , between oy and oz is 30° , and between ox and oz is 60° , then it is quite easy to draw this situation on our piece of paper. This represents correlations of .87 (the cosine of 30°) between x and y and between y and z , and a correlation of .5 (the cosine of 60°) between x and z .

However, if all these angles between ox , oy , and oz were 30° , it would not be possible to represent this graphically in two dimensions. It would be necessary to conceive of one

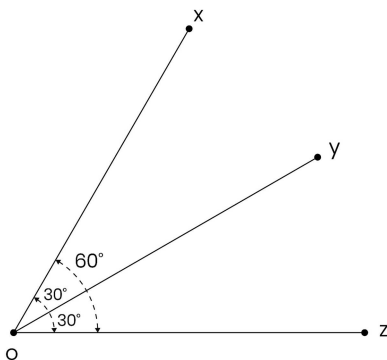


Figure 4.2 Figural representation of the correlations between three variables. The variables represented are x , y , and z , where the correlations between x and y and between y and z are .87 (cosine of 30°) and that between x and z is .50 (cosine of 60°).

of the lines projecting into a third dimension to represent such a matrix. With more than three variables, it may be that as many dimensions as variables are required to “draw” the full matrix, or it may be that some reduction is possible. Factor analysis seeks to find the minimum number of dimensions required to satisfactorily describe all the data from the matrix. Sometimes matrices can be reduced to one dimension—sometimes two, three, four, five, and so on. Of course, there is always a certain amount of error in any measurement, so this reduction will always be a matter of degree. However, the models used will seek solutions that can describe as much variance as possible and will then assume that all else is error.

In 1931, Louis Leon Thurstone developed a technique for carrying out factor analysis within the vector model that has just been described. He extracted the first factor by the process of vector addition, which is in effect the same process as that for finding the center of gravity in another application of vector algebra, and is thus called the centroid technique. The centroid is a latent variable, but it has the property that it describes more variance than any other dimension drawn through the multidimensional space, and we can in fact calculate this amount of variance by summing the squares of all the projections onto this vector from all the observed variables. The square root of this value is called the eigenvalue of the factor, which we will discuss in more detail later in this chapter. The position of the first factor can be described by reporting the correlation between it and each of the observed variables, which are functions of the angles involved. The first factor within the centroid technique describes a basic fixing dimension, and when its position has been found, it can be extracted from the multidimensional space so that further factors are sought only in regions at right angles to it. The cosine of an angle of 90° , that is, a right angle, is 0; and thus factors represented by lines drawn at right angles to each other are independent and uncorrelated. It is for this reason that factors are sometimes called dimensions, as figuratively they behave like the dimensions of space. A unidimensional scale is one that requires only one factor to describe it in this way. If further factors exist, a unidimensional scale will not be adequate to fully describe the data, which will be described by further consecutive dimensions.

Multidimensional scaling

There are many similarities between factor analysis and the process known as multidimensional scaling. Multidimensional scaling originally achieved popularity among psychophysicists and proved particularly useful in practice for defining psychophysical variables. It was by this technique, for example, that the idea of there being only three types of color receptors in the retina of the eye was generated, as it was found that people required only three color dimensions to describe all colors. Multidimensional scaling also provided a useful model for the behavior of the hidden values in parallel distributed processing machines (Rumelhart & McClelland, 1986). These models of parallel-processing computation show similarities with the actual arrangement of the system of connections between neurons in the human brain, and played an important part in the development of machine learning. It is likely that representational analogies of the type used in multidimensional scaling may turn out not just to be a convenient tool but also to tell us something about how the networks of neurons in the brain function.

In much the same way that multidimensional scaling models have provided a conceptual underpinning for psychophysics, factor analysis fulfills a similar role for psychometrics. Its success may be due to more than mere statistical convenience: it could be

that the figural representation of factor analysis is so powerful because it mirrors the cognitive processes whereby human beings make judgments about differences between objects (or persons). In fact, a particular neural architecture found within the human brain has been shown to carry out factor analysis when emulated within a computer. There is therefore the possibility that the brain itself uses factor analysis when trying to make sense of large amounts of data.

Application of factor analysis to test construction

In psychometric test construction, the factor analysis of the intercorrelations between test items has provided an alternative to traditional item analysis and proven particularly useful in examining a test specification's conceptual structure and the bias in its scales.

Eigenvalues

The original factor-analytic transformation generates as many factors as there are variables and calculates a value, called an eigenvalue, for each. The original set of variables defines the total amount of variance in the matrix, with each variable contributing one unit. Therefore, with a factor analysis of data on 10 variables, the total amount of variance present will be 10 units. The factor analysis rearranges this variance and allocates a certain amount to each factor while conserving the total amount. The quantity allocated to each factor is a function of the eigenvalue, such that the sum of the squared eigenvalues of all the original factors adds up to the total number of variables. With 10 variables, for example, there will be 10 original factors. The sum of the squares of the eigenvalues of these factors will be 10. The larger the eigenvalue of a factor, the greater the amount of the total variance it accounts for, and the more important it is. The first factor typically accumulates a fair amount of this common variance, and subsequent factors progressively less. At some point, they start having eigenvalues of less than 1, indicating that they account for less variance than is accounted for by a single variable.

Identifying the number of factors to extract using the Kaiser criterion

As the purpose of factor analysis is to extract information spread across many variables and represent it with fewer dimensions, factors with eigenvalues of less than 1 are typically discarded. This intuitive rule is sometimes called the Kaiser criterion after its original advocate, Henry Kaiser. Sometimes, however, the situation is more complicated. For example, there may be too many unreliable variables or uncorrelated variables, leading to solutions containing many factors with eigenvalues greater than 1. Also, sometimes, there are many factors with eigenvalues around the cutoff threshold of 1. It would make little sense to use an eigenvalue criterion level of 1 where the eigenvalues for the first seven factors were 2.10, 1.80, 1.50, 1.10, .90, .60, and .5. Although the Kaiser criterion would suggest retaining the first four factors, it might make more sense to inspect the three- and the five-factor solutions as well: there seems to be a major drop in variance explained between the third and fourth factors, and there seems to be little difference in variance explained by the fourth and fifth factors.

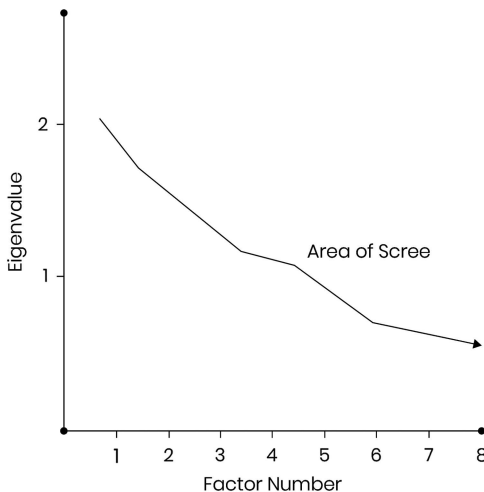


Figure 4.3 Plot of eigenvalue vs. factor number demonstrating a Cattell scree.

Identifying the number of factors to extract using the Cattell scree test

An alternative to the Kaiser criterion is provided by the so-called Cattell scree test, which uses the metaphor of the pattern of pebbles on a seashore for the shape of a plot of eigenvalues vs. factor numbers. Cattell suggested that a scree might be expected just at the point that divides important factors from noise.

Figure 4.3 presents eigenvalues of factors extracted from hypothetical data. The scree is clearly visible here, and the scree test would suggest extracting five factors.

Other techniques for identifying the number of factors to extract

However, some data produce no scree, so alternatives are needed to decide on the number of factors to extract. In fact, the best guide is typically given by examination and interpretation of the meanings of the solutions containing different numbers of factors. Generally, it is best to retain as many factors as can be reasonably interpreted by their correlations with original variables. For example, the first factor could represent a general factor, the second factor age, the third factor bias, the fourth factor a potential subscale factor, and so on. Eventually, factors will be reached that are uninterpretable, and this is a good place to stop. An additional technique, particularly useful when there are large samples, is to break down the sample into subgroups and investigate the extent of similarity between the separate factor analyses. The first factors that look similar across subgroups are good candidates to be retained.

Factor rotation

The idea of rotating factors has been around for some time and was of particular interest to Thurstone in the 1930s. It is useful when multiple factors are required to adequately

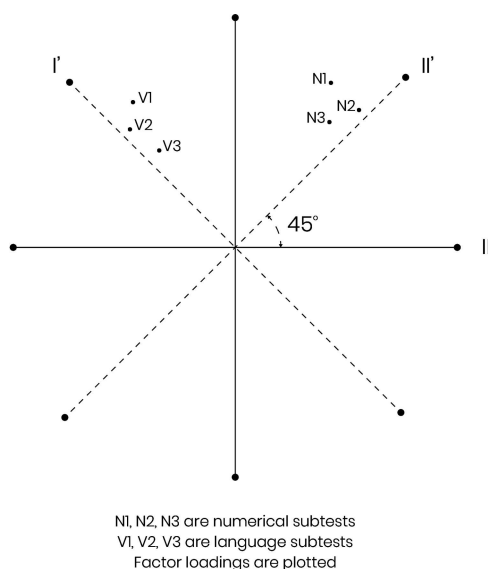


Figure 4.4 Rotation of orthogonal factors.

describe the data. Factor rotation can most easily be explained using a simple two-factor solution. As discussed before, the first factor extracted accounts for most of the variance, and the second represents the remaining variance unexplained by the first factor. However, the actual position of these factors with respect to the underlying variables is rarely easily interpretable. In fact, there are any number of different ways in which we could define such latent factors within the two-factor space. An example two-factor solution is presented in Figure 4.4.

Consider as an example a situation where the loadings of six subtests of ability (arithmetic, calculation, science, reading, spelling, and writing) on a factor analysis are, respectively, .76, .68, .62, .64, .59, and .51 on factor I, and .72, .67, .63, -.65, -.67, and -.55 on factor II. If a graph is drawn plotting these loadings on the two factors (I as the y axis and II as the x axis), then they form two clusters: on the top right-hand side we have numerical abilities, while verbal abilities are on the top left-hand side. In this situation, we could interpret factor I as a general ability factor, while factor II would contrast people who are good at mathematics and bad at language with people who are good at language and bad at mathematics. However, if we draw two new lines on this graph, both through the origin and at 45° to I and II, we note that one of these lines passes very close to the mathematics cluster with hardly any loadings on language, while the other passes very close to the language cluster with hardly any loadings on mathematics. This could be interpreted as reflecting the existence of two independent ability factors, one of mathematics and one of language. Both these solutions are compatible with the same set of data! Interpretation in factor analysis is never straightforward, and to fully understand the results, it is necessary to be familiar with its underlying conception.

Rotation to simple structure

At one time, rotations of factor-analytic data were carried out by hand, in the manner just described, and solutions sought by drawing lines on graphs that gave easily interpretable solutions. However, there was some criticism of this approach (it is valid, merely open to abuse), claiming that it was too subjective. Thurstone therefore introduced a set of rules that specified standard procedures for rotation. The main one of these was rotation to simple structure. The rotation carried out in the previous example on the mathematics and language data is such a rotation, and it involves attempting to draw the rotated factors in such a position that they pass through the major dot clusters. In practice, the easiest way to do this algebraically is to require that as many of the variables as possible have loadings on the factors that are close to 0 (a loading is like a correlation between the variable and the factor). Rotation is then defined in terms of the minimization of loadings on other factors, rather than the maximization of loadings on the factor in question. This is the process that is carried out by the varimax procedure offered by most factor-analysis software, and is by far the most popular factor rotation technique.

In practice, data rarely behave quite as nicely as in our example, and the software may find it difficult to decide which fit of many poor fits is best. There are other rotation solutions that can be tried in these situations, selecting solutions based on other criteria. For example, if it is impossible to find a solution where the lines pass through one set of variables while other sets of variables have low loadings, priority can be given to one or another.

Orthogonal rotation

Generally, in classical factor analysis, the derived factors are independent of each other—that is, they are drawn at right angles (orthogonal) to each other. There are good reasons for this. The factor structure—because it lies in “possible” space, rather than the real space of the original correlations between the variables—needs to be constrained, as there would otherwise be far too many possible solutions. This was after all one of the reasons why the rotation to simple structure was introduced as an algorithmic alternative to the subjective drawing of rotations. There is a further advantage of orthogonal factors in that they are relatively easy to interpret.

Oblique rotation

However, there are situations where the data do not easily fit an orthogonal solution, and further situations where such a solution is artificial. It might be felt that there are good reasons why two particular factors would be expected to relate to each other. An example here might be anger and hostility as personality variables. In these situations, an oblique solution may be more appropriate. These are more difficult to interpret, as one of the main constraints has disappeared, and the factors found are not independent. The extent to which the orthogonality criteria can be relaxed can vary, and it is often found that different degrees of relaxation produce quite different solutions, so that a great deal of experience is required to interpret rotations of this type. They are best avoided by people without experience of the technique.

Limitations of the classical factor-analytic approach

Factor analysis is a confusing technique that can easily produce contradictory results when used by the unwary. Generally, the analysis is particularly unsuited to testing hypotheses within the hypothetico-deductive model, as it can so easily be used to support many different hypotheses from the same set of data. A common error is the assumption that if two variables have loadings on the same factor, then they must be related. This is nonsense, as can be demonstrated by drawing two lines at right angles to each other, representing two uncorrelated variables, and then drawing a line between them at 45° to each, to represent a factor with a .71 loading (the cosine of 45° is .71) from each variable! In early research, major theoretical battles were often carried out based on the results of different factor-analytic rotations. For example, was there one factor of intelligence or many? These disputes were in the end seen as pure speculations; either position could be supported depending on how the data were interpreted. An important debate between Eysenck and Cattell about whether personality could be explained best in terms of two or 16 factors turned out to be dependent on whether orthogonal or oblique rotations were used on the same data. Thus, there could be two personality factors or there could be 16—depending on how the situation was viewed.

A general dissatisfaction with factor analysis was widespread in the second half of the 20th century because of the apparent ability of the technique to fit almost any solution, and at this time, stringent criteria were recommended concerning its use. It was felt that sample sizes had to be very large before the analysis was contemplated. In fact, the recommended samples were often so large as to render the use of factor analysis impractical in these early days. There were further constraints introduced in terms of the assumptions of the model, and the requirement that the variables in the correlation matrix have equivalent variance. This again produces a considerable practical problem, as binary data in particular often fall short of this requirement, and item scores on psychometric tests are frequently binary. It was not until modern psychometric methods—such as logistic and confirmatory factor analysis (see Chapter 5)—were introduced that these issues were resolved. Looking back, it seems amazing that so much was achieved using the approximate solutions of earlier days.

Criticisms of psychometric measurement theory

The last century saw many debates about whether psychometric traits, and intelligences in particular, were the sort of thing that could be measured at all, and the facetious definition of intelligence as being merely that which is measured by intelligence tests received widespread acclaim (Boring, 1957). This came particularly from those who were opposed to psychological testing more generally, but these critics have all made contributions to modern psychometrics, each in their own way. Their arguments are worthy of consideration.

Psychological and educational tests carry out a form of measurement, but unlike physical measures such as length or weight, there is considerable confusion over what they measure and how they can do so. But to ask “Do latent traits actually exist?” begs a number of questions. One problem is that what is measured is not a physical object but an intervening construct or a hypothetical entity. For example, in assessing whether a test of creativity is really measuring creativity, we cannot compare a person’s score on the test directly with their actual creativity. We are restricted to seeing how the test scores

differentiate between creative and noncreative individuals according to some other ideas about how creative people should behave. The measurement of concepts like creativity, extraversion, and intelligence is limited by the clarity with which we can define the meaning of these constructs. The aim is to identify these latent traits and to obtain as good a measure as possible on each of them for each individual: the true score on each latent trait. It is limited by the inability of not just the psychometric community but also of society more widely to agree on what constitutes the essential character of each trait.

Major criticisms of true-score theory were directed against the concept of the true score itself by those who felt that the statistical definition was misleading. It was argued that we cannot deduce from a score on a test that anything whatsoever “exists” in the brain, as intelligence is merely a construct arising from the use of the test. The true score is seen as being an abstraction and therefore of no theoretical importance. The essence of this view is that psychological measurement is fundamentally different from the way in which measurement is used in science more generally. To address this issue, we can have recourse to an alternative definition of a true score, based on Plato’s theory of reality and truth in his *Metaphysics*.

The Platonic true score

The Platonic concept of a true score is based on Plato’s theory of truth. He believed that if anything can be thought about, then even if it does not exist in the physical world, it must exist somewhere, perhaps in some sort of Platonic heaven where imaginary things exist. The unicorn is often given as an example of such an object. Nonexistence is reserved for objects about which we cannot even think—perhaps Donald Rumsfeld’s “unknown unknowns.”

Many argue that the Platonic idea of the true score is a mistake, and this comes from both those for and those against psychometric measurement in principle. Those who are for criticize the Platonic approach as unnecessary and misleading, feeling perfectly satisfied with the statistical definition. Those who are against believe that those who argue for the existence of a construct from the existence of reliable test scores make a category error. Just as behaviorists argue that there is no mind (only behavior), so it is said that there is no true score, only a series of observed scores and deductions. However, this is an oversimplification. There are many abstract nouns in use that, although not attached directly to objects, certainly exist: for example, justice. Certainly we might agree that in one sense, justice does not physically exist, but we would probably not see this as being equivalent to agreeing with the statement “There is no justice in the world” or “There is no such thing as justice.” Just because an abstract object has no physical existence, this does not mean that it cannot “exist in some quantity and therefore be measured.” For example, some have suggested that love cannot be measured. But are they really trying to say that expressions such as “Do you still love me?” or “I love you more than ever” are meaningless? No, love can be rated, and it can therefore be measured—probably even by questionnaire.

Psychological vs. physical true scores

Is there something special about physical measurements that sets them apart from psychological ones, making physical measurement immune from the stratagems of true-score theory? Not necessarily. While normally most people are quite happy with the idea of the

length of common objects being fixed, this is in itself a somewhat Platonic approach. If we take, for example, the length of a physical object such as a table, we can never be wholly accurate in its measurement—no two people taking measurements down to a fine degree are going to agree exactly. And even if they did, unfortunately by the next day, both would be proved wrong. A little damp may cause the table to expand, or heat may cause it to either contract or expand depending on its material. It might perhaps be said that the table did indeed have a length, but these were different at different times. But even then, a table can never be completely rectangular (this is itself a Platonic concept), so which of several measurements counts as the “true” length? The length of the table to which they aspire is a true score. It sets a standard to which no real-world table can ever conform. And what is true for the measurement of tables is even more so for medical measurements such as blood pressure, which varies not only with every measurement but also second by second and across different parts of the body. Again, blood pressure is more than an observed measurement; it is a true score to which we aspire. In practice, it is as accurate as it needs to be. It might be hoped that some entities at least may be measured perfectly, but this is a forlorn hope. Even the speed of light is not known with complete accuracy, and mathematical fundamentals such as π (the ratio of the circumference of a circle to its radius) or e (the growth constant) can only be known up to a certain number of decimal places. Many Platonic entities have indeed turned out to be unicorns, but without imagination there is no science. True-score theory may have started its existence as a bit of a rhinoceros, but it has not ended as a unicorn. Rather, it is today a model for deep learning in a wildlife park for AI.

Functional assessment and competency testing

During the second half of the 20th century, an alternative approach to psychological assessment was promulgated. This was functional, and focused on the assessment of competencies rather than latent traits. In the trait approach, the test is there to measure an underlying psychological construct, such as intelligence or extraversion. In the functional approach, on the other hand, the test is there simply to achieve a purpose: that of successfully separating respondents in terms of some target application, such as likely success at a job.

Within the functionalist approach, the design of a test is completely determined by its use, and “what it measures” has no meaning other than this application. Two examples are the work of David McClelland, an advocate of competency testing in the workplace, and W. James Popham (1999), an enthusiast for criterion-referenced testing in education. McClelland argued that trait-based assessment was completely ineffective as a tool for selection in employment settings, and that neither ability tests nor grades in school predict occupational success. He concluded that criterion-referenced competencies are better able to predict important behaviors than more traditional norm-referenced tests. The influence of his early work remains today in the popularity of the competency-based approach for the assessment of vocational qualifications. Popham argued that there had been too much emphasis on normative factors in testing. He pointed out that if, for example, we were interested in whether someone could ride a bicycle, then the performance of other people on their bicycles was irrelevant. Indeed, we should be particularly delighted if we found out that all were able to do so, and not in the least concerned that we did not have a wider spread of abilities. For him, it is only performance on the criterion that matters, even if all individuals obtain the same score.

More recently, the functional model has been the basis for most psychometric or psychographic systems built by machine-learning algorithms that simply learn to discriminate between predefined groups. There is no consideration of how or why these particular groups were chosen in the first place.

The functional approach can produce tests for many practical circumstances, but it has several weaknesses. First, we cannot assume that a test developed with one particular purpose in mind will necessarily be of any use for another. In many areas of application, however, this has been a strength of the model rather than a weakness. In education, for example, the separation of the function of formative assessment—where tests are used to identify areas of the curriculum that need to be developed by both the teacher and student during the remainder of the educational session—and summative assessment—where a final indication of the student's attainment is given—has been generally well received. The way in which summative examinations control the curriculum has been widely criticized, and the formative assessment process has been welcomed as an approach that not only limits this control but also introduces feedback at a time when something can be done about it rather than when it is too late. However, it should be recognized that the actual content of both types of examination will be broadly similar, and in practice there will be considerable overlap between the content of each.

Second, the functional model insists, almost as a point of principle, that no psychological intervening variables or traits can be relevant. As with early 20th-century behaviorism, the only interesting aspects of traits are the behavior to which they lead, and as this is measured and defined directly and functionally, the traits are redundant. Within functionalism, there is no such thing as, for example, an ability in mathematics; there is only the performance of individuals on various mathematics items. The pursuit of such an approach is, however, somewhat idealistic and certainly does not reflect existing practice in the real world. People do tend to use concepts such as “an ability in mathematics” and frequently apply them. Indeed, it is normally on the basis of such concepts that generalization from a test score to an actual decision is made, whether justified or not. How else could a GCSE in mathematics, for example, be used by an employer in selecting a person for a job? Certainly, it is unlikely that the mathematics syllabus was constructed with any knowledge of this employer's particular job in mind. Neither is it likely that solving simultaneous equations will be a skill called for in the job in question. Indeed, how many people who have a GCSE in mathematics have ever “found x ” since leaving school? No, the criteria used in practice here are not functional ones but involve the use of commonsense notions about the nature of individual differences in human ability.

Thus, we see that in spite of the superficial advantages and objectives of the functionalist approach, trait psychology remains essential because it so closely represents the way in which people actually make decisions in the real world. While some have argued that all such trait-related processes are wrong and must be replaced by functionalism, this represents an unreasonable and unwarranted idealism. It is really no good trying to prescribe human thought processes. To an extent, much of psychology is no more than an attempt to be objective and consistent in predicting how people behave. If this can be achieved by assuming the existence of traits, then so be it. Examples of the success of the approach abound, particularly in clinical psychology. A test of depression such as the Beck Depression Inventory (BDI)—although originally constructed around a framework defined by the functional model that identifies a

blueprint of depressive behaviors and thoughts—would be of little use if it had to be reconstructed with each application of the concept of depression in different circumstances. Solely functional tests on their own can only be specific to a situation; they cannot easily be generalized. If we wish to generalize, then, we need a concept, a trait of depression, to provide justification for saying that the depression scale might be applicable in changed situations—for example, with children, or with reactive as well as endogenous depression. To function in this way, the BDI needs to have construct validity, and this cannot exist without presupposing the construct and trait of depression itself. The BDI relates to a wide range of mood changes, behaviors, thoughts, and bodily symptoms that psychologists, psychiatrists, and therapists consider to be part of depression.

Machine learning and the black box

In spite of this, the functional approach has seen a recent resurgence within AI, where it is argued that so long as the machine is able to learn how to identify the key elements that differentiate groups, then how this is achieved is irrelevant. Generally, such systems are designed to maximize particular outcomes—usually profit. Attempts by insurance companies to base premiums on post codes were made illegal by the EU in 2013. Such premiums would necessarily discriminate not just against the poor but also against any group that was more likely to experience poverty, and hence they would potentially be in breach of equal-opportunity legislation.

AI algorithms are trained using vast amounts of data collected over years; if the data include past racial, gender, or other biases, the predictions of these AI algorithms will reflect these biases. With no requirement to explain how decisions are reached, the internal workings of the algorithm become a black box. This is a serious issue in the use of AI by courts and correction departments to assist in bail, sentencing, and parole decisions, as well as in areas like predictive policing. In a review of the “techlash,” Atkinson et al. (2019) concluded that in order to reduce the potential for algorithmic bias to cause harm, regulators should “ensure that companies using AI comply with existing laws in areas that are already regulated to prevent bias.” However, the regulation of this presents difficulties, some of which may be insurmountable. While the EU is in the process of passing legislation that will insist that all recommendations made by AI be explainable, the sophistication of these systems is often far beyond anything that can be explained in human terms.

Summary

Accurate measurement is the key to success in many sciences, and this is as true in psychology as it is in physics or chemistry. And all sciences have their unicorns, whether they be the ether, phlogiston, or animal magnetism. But in spite of these dead ends, scientists have pursued their dreams and achieved what would once have been seen as miracles. The early application of factor-analytic techniques for the identification of true scores has evolved through path diagrams and latent-variable analysis to the hidden layers within deep-learning neural nets that are so essential to modern AI. If these hidden layers remain hidden, they will continue to be black boxes. If AI one day becomes truly intelligent, it too may want to know what lies within these black boxes. Maybe it will do so before humans actually discover how their own brains work.

We can see that from an ethical perspective, both the trait and functionalist approaches have their advantages and disadvantages. Neither can be said to be wholly right or wholly wrong. What is important is that psychometricians and data scientists realize which set of assumptions they are using in a particular situation and be prepared to justify this use. Regarding the use of AI for any psychometric purposes, we are still awaiting regulation, which is becoming increasingly necessary.