COUNCIL OF EUROPE CONSEIL DE L'EUROPE

December 2004 **DGIV/EDU/LANG (2004) 13**

Reference Supplement

to the

Preliminary Pilot version of the Manual for

Relating Language examinations to the Common European Framework of Reference for Languages: learning, teaching, assessment

Section F: Factor Analysis

Language Policy Division, Strasbourg

Section F

Factor Analysis

N.D. Verhelst
National Institute for Educational Measurement (Cito)
Arnhem, The Netherlands

The performance on a test is usually summarized by a single number, the test score. This test score is a composite score, because it is built (by taking a sum) from item scores. In general one might ask whether it is meaningful to put a number of items together in a test and to let the performance be represented by a single number. What if the test consists of a mixture of two kinds of items, each kind measuring a different concept? Is reporting a single score meaningful or should one treat this composite test as two tests and report two test scores?

A model suitable to detecting if there are more dimensions responsible for the performance on the test is Factor Analysis (F.A.). The model originated in psychology, more than a hundred years ago and is still one of the most applied models in the social sciences. Although not defined originally as such, the model fits very well in the family of IRT-models to be discussed in Appendix G. But since the model and the techniques to carry out the analyses are so wide-spread (as well as a lot of misunderstandings about them), a separate, though short appendix is devoted to F.A.

The basic observation from which F.A. originated is the non-zero (but also not perfect) correlation between several measures that belong to some broad domain, like cognitive tests. F.A. is a model which explains the pattern of correlations that issues from observations in testing (or other measurements). Basically it says that since the correlations are not zero, the measurements must have something in common, and, since the correlations are not perfect either, the measurements must have also something unique. This is the general idea, which will be made more concrete next.

The common thing that tests share is called a factor (or, as the case may be, several factors). A factor is conceived as a non-observable (or latent) continuous variable, and every person taking the test can be represented by a value on this variable, called a **factor score**. If there are more factors, every person has a factor score on each factor. The 'unique thing' can also be conceived of as a factor, where the person also has a score. The **observed score** on a test is conceived of as a weighted sum of the factor scores, including the unique factor. In Table F.1 an example is provided with three tests and two common factors (The notion of 'common' factors is explained by means of the table)

Table F.1. The basic model of Factor Analysis

	weights for	
	factor 1	factor 2
test 1	0.4	0.2
test 2	0	0.7
test 3	0.7	-0.3

Suppose John's factor scores on the two factors are +1.2 and 0.8, respectively. Then the model says that John's observed score on test 1 is $0.4\times1.2+0.2\times0.8+$ his score on the unique factor for test 1. But we know from Classical Test Theory that the observed score also contains a measurement error. Therefore we have to conceive the score on the unique factor as a mixture of something systematic (but unique to the test) and the measurement error. But these two are confounded and cannot (with the three tests at hand) be disentangled. The other two factors are called common factors, because for each factor there exist at least two different tests with a non-zero weight for that factor. These weights are called **factor loadings**, and the main purpose of Factor Analysis (as a technique) is to determine these

weights. All one needs to carrying out such an analysis is the table of correlations (or covariances[1]) between the tests.

The discussion in the present section will be restricted to points which are essential in the interpretation of factor analytical results.

1. **Unique factors**. Suppose that in the preceding example, test 1 is a reading test, test 2 is a writing test and test 3 is a listening test. Suppose further that the reading test contains a lot of items (or text passages) on history, while the other two tests have nothing to do with history. Suppose, finally, that John is particularly good at history, such that his score on test 1 is determined to a considerable extent by his knowledge of history, while Mary is not very good at history, such that her knowledge in that domain will not be of much help in answering the questions of the reading test. This makes clear that knowledge of history will account for some variability in the test scores of test 1. But since the other two tests have nothing to do with history, 'knowledge of history' is unique for test 1, and cannot appear as a common factor. If we add a fourth test to the collection (a history test, for example), then there will be two tests which have history as a common factor, and this will show up in the analysis, and we might end up with three common factors, where the third factor has loadings of zero for tests 2 and 3, but non-zero loadings for test 1 and the added history test. More generally this means that unique factors are to be considered relative to the collection of tests included in the analysis.
2. **Origin and unit**. Suppose the factor scores on factor 1 for all people are multiplied by 2, and at the same time the factor loadings in column 1 are divided by 2, then the product of the transformed factor scores and the transformed weights would not change. Multiplying the scores by 2 is choosing another unit of measurement (if one owns 1000 euros, one also owns 2000 'half-euros'). The unit of measurement is in principle free (arbitrary), but to make communication possible, the unit used must be specified. It is common practice to choose the standard deviation of the factor scores as unit, or in other words, the standard deviation (in the population) of the factor scores is one. With a similar reasoning, one can choose the origin of the scale in an arbitrary way. It is common practice to choose the average factor score (in the population) as origin. Therefore, it is a common convention (and not a metaphysical truth) to say that factors have a mean of zero and a standard deviation of one. (Notice that this is **not the same** as saying that the factor scores are normally distributed.)
3. **Correlations and covariances**. Factor analysis can be carried out on tables (matrices) of correlations and on tables of covariances. A covariance (between two variables) is a measure of covariation (meaning literally: varying together). Its value depends on the unit of measurement used for the two variables. A correlation is a kind of standardized measure of covariation and varies between –1 and +1. If the correlation matrix is used for a factor analysis (as we will assume in the sequel), then the factor loadings cannot be larger than one in absolute value.
4. **Orthogonal factors**. The indeterminacy of what are called factors is more complicated than only the freedom in the choice of the unit and the origin. Also the correlational structure of the factors in the population is arbitrary (not completely, but to a large degree). For example, if there are two common factors, they can always be defined in such a way that the correlation between the factor scores (in the population) has an arbitrary value (different from –1 and +1). But changing the correlation will also lead to a change in the factor loadings. In many applications, the factors are chosen in such a way that their correlation is zero. Any pair of factors with zero correlation is called orthogonal. Most software give factor loadings for orthogonal factors as their primary output.
5. **Communality**. The sum of squares of the factor loadings (on the common factors, and with orthogonal factors) of a particular test is called the **communality** of that test. From Table F.1, we see that the communality of test 3 equals $0.7^2 + (-0.3)^2 = 0.58$. The communality is the proportion of the test variance that is explained by the two factors. In this case 58% of the variance

[1] The **covariance** between two variables is the correlation **multiplied** by the product of the two standard deviations. Or, conversely, the correlation is the covariance **divided** by the product of the two standard deviations. If one of the standard deviations equals zero, then the covariance is also zero, but the correlation is not defined, because the division of zero by zero is not defined.

is due to the two factors, and the complement (42%) is explained by the unique factor, part of which (but unknown) is due to measurement error. Thus we see that from F.A. we get another lower bound for the reliability of the test: the reliability is at least as large as (but may be larger than) the communality. As can be deduced from the discussion on unique factors, this lower bound may change as more or other tests are analyzed jointly in a F.A.

6. **Contribution of factors**. One may also take the sum of the squared loadings for a particular factor across the tests. This sum is called the contribution of that factor (to the total variance). In Table F.1 the contribution of the first factor is $0.4^2 + 0^2 + 0.7^2 = 0.65$. The contribution of the second factor is 0.62. Their sum (0.65+0.62=1.27) can be compared to the total variance which is the number of tests, in the present case 3 (Since we use correlations, each variable has been standardized, and thus has a variance equal to one). So, in the example we see that about 42% $(100 \times 1.27/3)$ of the total variance is explained by the two common factors. The remaining part is due to the unique factors. Most techniques of factor analysis determine the factors in such a way that the first factor explains as much variance as possible, the second factor then explains as much variance of the variance not explained by the first factor, etc. The technical term used for the determination of factors is **extraction of factors**. Notice that this way of extracting factors is just a mathematical procedure; it does in no way justify any substantive meaning or interpretation whatsoever to be attached to these factors. We will come back to this point.
7. **Reproduced correlations**. If we have the factor loadings, we can reproduce the correlation matrix from them. The reproduced correlation between two tests is the sum (over factors) of the products of the factor loadings of the two tests. From Table F.1 we can compute that the correlation between test 1 and test 3 is $0.4 \times 0.7 + 0.2 \times (-0.3) = 0.22$. Factor analysis as a technique does the reverse in some sense: from the correlations it has to compute the factor loadings. This reverse operation (which is mathematically not simple), however, is not well defined, because there does not exist a unique solution but infinitely many of them, even if we require that the factors are standardized and mutually orthogonal. This is explained next.
8. **Orthogonal rotation**. The factor loadings of Table F.1 are displayed graphically (as points in a plane) in the left hand panel of Figure F.1: the loading on the first factor corresponds to the x-value of the point, the loading on the second factor to the y-value. The points representing tests 1 and 3 are connected to the origin by a dashed line. Although the reproduced correlation was computed as a formula involving the loadings, it can also be computed from the distances of the points to the origin (the length of the dashed lines) and the angle between the dashed lines. Now imagine that the points representing the tests are fixed on the paper surface, but that the axes of the system lie loosely on the paper surface, fixed at the origin, such that they can rotate. In the middle panel of the figure this is shown by the dashed lines: both axes are rotated 45 degrees clockwise. In the right hand panel then, the old axes are removed, the new (rotated) ones are displayed as solid lines now, and the whole picture is turned such that one axis is horizontal and the other vertical. Notice that the pattern of dashed lines connecting the two test points to the origin has not changed: the dashed lines have the same length as in the first case, and they form the same angle. But the values of the x- and y-coordinates have changed. Their values are given in Table F.2, together with the old ones. It can be checked easily that the reproduced correlation from either solution are identical. Of course, we could have rotated the original axes an arbitrary number of degrees, each rotation giving a different solution, and there is no best solution, because they are all equivalent.

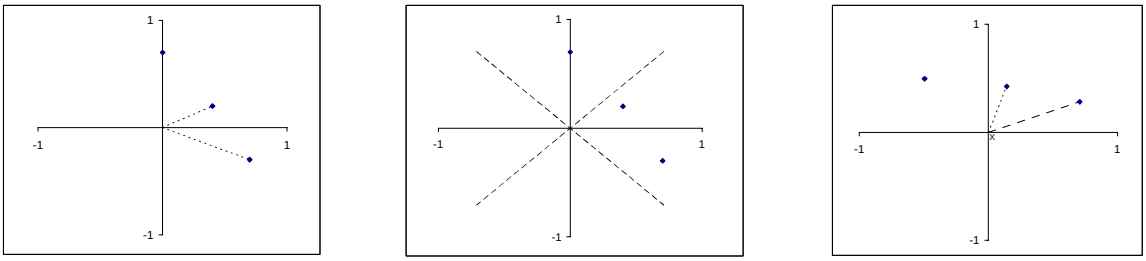

Figure F.1. Orthogonal rotation

Table F.2. Factor loadings before and after rotation

	before rotation		after rotation	
	factor 1	factor 2	factor 1	factor 2
test 1	0.4	0.2	0.141	0.424
test 2	0	0.7	-0.495	0.495
test 3	0.7	-0.3	0.707	0.283

9. **Interpretation**. Suppose the tests used in a single factor analysis consist of four reading tests and four listening tests. Suppose further that we can find a rotation such that the four reading tests have a positive loading on the first factor and a zero loading on the second factor, while the reverse holds for the listening tests. Then we could say (but this is a summary of our finding) that the first factor is a 'reading factor' and the second a 'listening' factor. This means that we can describe the covariation between eight original variables by a more parsimonious conceptualization which involves only two concepts. It does not follow, however, that there 'must' exist 'something real' (like a brain center) which is responsible for reading and another something which is responsible for listening. Conclusions like these are called reification, and they are not logically allowed: maybe there do exist such brain centers, but their existence does not follow from a factor analysis.
10. **Statistics and the number of factors**. All that has been said up to now is related to an analysis of a correlation matrix as it exists in the population. But the only thing one can analyze in practice is a correlation matrix computed on the data of a sample (usually the calibration sample). Therefore the correlations in the matrix are estimates of the population correlations, and the factor loadings will also be estimates of the population factor loadings. This all may sound quite familiar by now, but there is an extra (and quite difficult problem) associated with F.A. Suppose the population correlation matrix for 10 variables can be reproduced completely (i.e., without any error) with two factors. Then the matrix of estimated correlations will very likely not be reproduced with two factors. In general more factors will be needed, and in many cases the number of factors will be equal to the number of variables. This is caused by the estimation errors in the sample correlations. Usually one will not use as many factors as there are variables, but if we do not know the exact number of factors required for the reproduction of the population matrix (and usually we do not know), we have to guess it. There exist some mathematical criteria to help in this guessing but none is foolproof.
11. **Exploratory and confirmatory F.A.** Originally, F.A. was developed as an exploratory technique. A collection of tests is factor analyzed 'to see' the factorial structure. Much effort has been devoted to develop special rotation techniques which might be helpful in the interpretation of the factors. The best known, and still frequently used method of rotation is the varimax rotation. It is available in most statistical packages. The big problem with exploratory factor analysis is that it is quite difficult to determine the 'real' number of factors. (This number must be specified by the user in carrying out the analysis.) In the 1970's statistical theories were developed where one can impose a prespecified structure on the factor loadings as a hypothesis. Here is an example: suppose the test constructor wants to factor analyze jointly four reading tests and four listening tests, and he has the hypothesis that reading and listening should be conceived of as two distinct proficiencies. This hypothesis can be translated in a partial fixing of the table of factor loadings, by requiring that the reading tests have a loading of zero on the first factor (so this factor represents the 'listening factor'), while the listening tests have zero loadings on the second factor. So, eight of the sixteen cells of the table of factor loadings are filled already with numbers issuing from the hypothesis. With the software for confirmatory F.A. the non-specified loadings are estimated, but things are a little bit more complicated now: the researcher also has to specify if he thinks that these two factors are orthogonal (i.e., uncorrelated) or not. In the latter case, the software also estimates the correlation (in the population) between the two factors. But it does more: it performs a statistical test that can be used to decide whether the hypothesis put forward is tenable or not. In general the use of such models is not a simple matter, and special training is strongly advised.
12. **When tests are items**. There is no objection in principle to use one-item tests to carry out a F.A. So, one can use the items of a test under construction as one-item tests, compute the correlations

between the items on the calibration sample and submit it to a computer program for factor analysis. There are, however, a number of problems associated with this approach. Three of them are discussed briefly.

a. Since factors are conceived of as continuous variables, any weighted sum of factor scores (and the observed score is such a weighted sum) is also continuous. If the tests are items, and their score can assume only the values 0 and 1, this leads to an inconsistency which usually shows up in the following way. If the correlations between the items are computed using the usual Pearson correlation coefficient (also called φ-coefficient), F.A. will usually find (too) many factors which are hard to interpret. Therefore it is strongly advised to use tetrachoric correlations, which are based on the assumption that a binary variable is the result of a dichotomisation of an underlying continuous variable. There is no simple formula to compute these correlations but they can be computed with many software packages.

b. Tetrachoric correlations have relatively large standard errors. If the sample size is small, this may lead to a difficult decision as how to choose the correct number of factors and to large standard errors of the factor loadings, complicating the interpretation of the extracted factors.

c. There exist many mathematical methods to do a F.A. Most of them require that the correlation matrix to be analyzed has a special mathematical characteristic called 'positive semi-definiteness'. A matrix of tetrachoric correlations often does not possess this characteristic, so that the operation of extracting factors will fail. There are two methods that do not require this characteristic, the so-called MINRES method and Principal Axes method. One should choose one of these in carrying out an exploratory analysis, because other methods will fail if the matrix is not positive semi-definite. Confirmatory analyses will fail in such a case.

13. **The case of a single common factor**. If there is only one common factor (in the population), one might conclude that this is a 'proof' of unidimensionality, which makes the operation of summarizing the test performance by a single number meaningful. One should be very careful with such reasoning: a one-common-factor case is better interpreted as a necessary, and not a sufficient requirement. This is illustrated with a small example. Suppose a F.A. is carried out on three reading tests, where questions are asked on text passages. In the first test, the passages are on art, in the second on technology and in the third one on sports. The loadings on the common factor are 0.72, 0.70 and 0.40 respectively. Here are some comments:

a. Sometimes comments are heard like this one: "The performance on (my) reading tests are governed by a single proficiency irrespective of the content of the text passages; the fact that there is only one factor 'proves' that the tests measure reading ability and nothing else." Such reasoning, however, is a fallacy: it may be the case that the scores on the three tests are (partly) determined by specific knowledge of arts, technique and sports. If the amounts of knowledge in these three domains are not correlated in the population, their effect will be absorbed into the unique factors and cannot be distinguished from measurement error. So the only way to know is to add another three tests in the same domains. In that case, the systematic effect of the specific domain knowledge will show up as three common factors. This is an example of performing a thorough validation of a test, even without the technical tool of confirmatory F.A.

b. The example also gives a nice opportunity to help in the interpretation of the factor loadings. In principle, factor loadings have nothing to do with the difficulty of the tests that are analyzed, but they are indices of discrimination. It can be shown mathematically that a factor loading is the correlation between the test score and the common factor. So in the example considered, the tests on arts and technology correlate substantially higher with the common factor than the test on sports. If the tests used in the F.A. are single items, the same principle applies: the factor loadings express the correlations between the items and the underlying factor, and can thus be used instead of the correlation between items and test score as a measure of discrimination.

c. The problems associated with factor analysis on items are hard and in the literature no completely satisfactory solution to handle them is available in the framework of factor

analysis, i.e., in the approach which takes the correlation matrix between the items as the basic data to be analyzed. In a sense, students of factor analysis tend to consider F.A. on binary variables as a kind of nuisance. There is, however, a different approach possible which puts the binary character of the variables to be analyzed at the center of the approach. This approach is known as Item Response Theory (which developed historically quite independently from factor analysis). It is discussed in Section G.