THE MARKER INDEX: A NEW METHOD OF SELECTION OF MARKER VARIABLES IN FACTOR ANALYSIS

MARCELLO GALLUCCI
MARCO PERUGINI[1]
UNIVERSITY OF MILANO-BICOCCA

A new method for selecting marker variables in factor analysis is introduced, called the Marker Index. The Marker Index evaluates the usefulness of a variable in representing a factor by weighting both primary and secondary loadings, and by maximizing the similarity between the marker and the underlying factor. The method is compared with other four methods of selections, based on factor loadings or factor weights. Theoretical and empirical comparisons of the performance of the different methods are performed and the advantages of the Marker Index are illustrated. The comparisons demonstrate that the use of the Marker Index as a selection method can improve factorial simplicity, scale validity, and reliability of the composite scales composed by variables selected as markers.

Key words: Factor analysis; Variable selection; Simplicity; Representativeness; Marker index.

Correspondence concerning this article should be addressed to Marcello Gallucci, Facoltà di Psicologia, Università degli Studi di Milano-Bicocca, Viale dell'Innovazione 10 (U9), 20126 MILANO (MI), Italy. E-mail: marcello.gallucci@unimib.it

INTRODUCTION

Factor analysis is one of the most-commonly used statistical methods in psychology. Users of factor analysis often face the problem of dividing the initial set of variables into subsets. A typical situation is selecting those variables that provide the best interpretation (naming) of the retained factors, or discarding the variables that do not contribute to any of the retained factors. Another very common case is the construction of concise sub-scales, by reducing a large number of variables to the best pool of variables available. In both cases the fundamental question is quantifying the usefulness of variables in a given factorial solution. The search for the best variables in a factor solution is the problem of the selection of the *marker variables*.

Most textbooks on factor analysis and test construction give only vague suggestions (e.g., Comrey & Lee, 1992, pp. 241-244; Dillon & Goldstein, 1984, pp. 69-70; Gorsuch, 1974, p. 238; Guilford, 1954, pp. 522-523; Horst, 1965, pp. 554-558; Kline, 1993, p. 140; Rummel, 1970, pp. 472-489; Tabachnick & Fidell, 1996, pp. 639-640), or ignore the issue (e.g., Crocker & Algina, 1986; Fabrigar, Wegener, MacCallum, & Strahan, 1999; Harman, 1976; Loehlin, 1987; Nunnally, 1978; Stevens, 1996). The most common suggestion is simply to consider the primary loading, perhaps excluding those variables with high secondary loadings. For instance, Horst (1965) suggested to "…select variables with the highest loadings in each of the factors. Presumably, we do not select any variables which have high loadings in more than one factor" (p.

557). As a consequence of this lack of adequate formal methods to assess the usefulness of variables, most practitioners simply rely on more or less idiosyncratic rules-of-thumbs.

In the present contribution we propose a new method for selection of marker variables. The new method, called the Marker Index, is derived from the geometrical properties of the factorial solution, such that the variables that are selected as markers are the ones that more closely resemble the underlying factor. We compare the new method with other methods, both analytically and empirically, and show its advantages. We show that variables selected with the Marker Index simultaneously satisfy the most important criteria that define good factorial solutions, namely simplicity, representativeness, and scale reliability.

MARKER VARIABLES

Marker variables are those variables that better represent the factor, conveying the meaning of the underlying dimension without being contaminated by other unrelated factors (cf. Cattell, 1978). To satisfy those requirements, marker variables should have two distinct fundamental properties: representativeness and simplicity. Representativeness means that the variables convey the same concepts and are valid indicators of the underlying dimensions (cf. Cattell, 1978). Thus, marker variables should be highly related to the factor they are measuring. Simplicity means that marker variables contribute to the achievement of structural simplicity (cf. Harman, 1976). A simple structure is achieved when variables have very low secondary loadings, which implies that each variable is related only to one factor.

These two concepts have natural counterparts in the domain of scale construction. When constructing a scale to measure a latent factor, one should select those variables that mostly resemble the factor. Cattell and Tsujioka (1964) define this property as scale validity, meaning that a scale is highly correlated with the underlying factor. In addition to scale validity, marker variables should not simultaneously measure other constructs. Cattell and Tsujioka (1964) define this property as factor trueness. From a scale construction point of view, factor trueness is the counterpart of factorial simplicity, whereas scale validity is the counterpart of representativeness. The best marker variables should therefore be those variables that produce an optimal compromise between the two criteria. The problem becomes how to select such marker variables.

MARKER INDEX

We propose a method based on an index of factorial simplicity, called Marker Index. For an intuitive understanding of the rationale behind the Marker Index, consider the two-dimensional factorial space in Figure 1.

A variable i can be plotted in the factorial space as a point with coordinates represented by its loadings on Factor 1 and on Factor 2. It is clear that a perfect marker variable is a variable with co-ordinates (1,0). Lying perfectly on the factor, in fact, this variable is the simplest variable of the solution, represents perfectly the factor, has the highest possible validity, and its vector is the closest possible vector to the factor axis. Thus, the perfect marker lies on the vertex of the factor (point f in Figure 1). As a variable departs from this location, it loses either in simplicity or in representativeness, or both. Therefore, the distance between a variable and the vertex of the factor captures the discrepancy between the actual position of the variable and the ideal position: the smaller is the distance, the better is the variable as a marker. In other words, the usefulness of a variable *i* can be uniquely identified by the distance between the variable and the point f, with coordinates (1,0), which is the point representing the best possible variable for the factor (Figure 1).

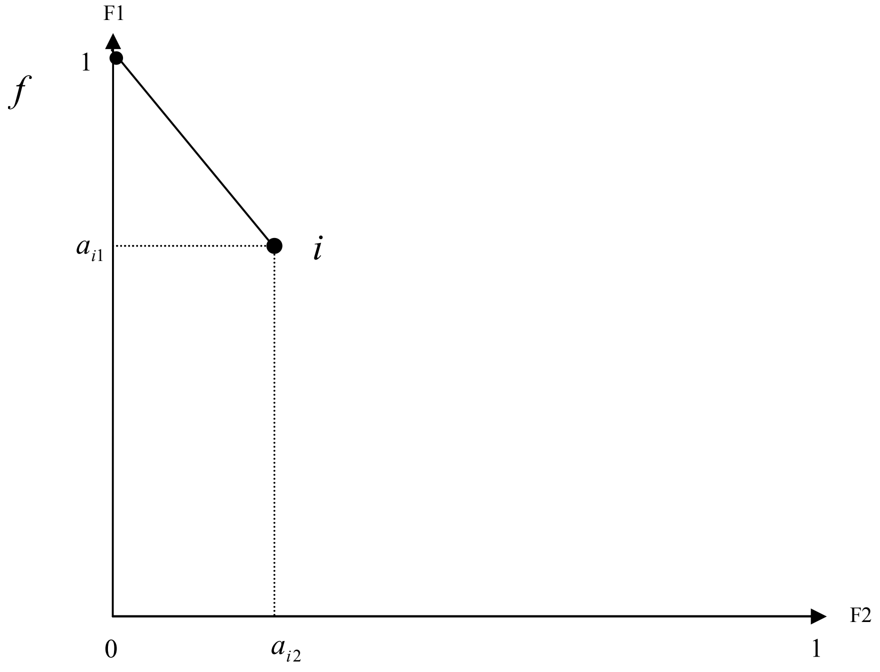

FIGURE 1
Geometrical representation of the Marker Index.

Generalizing to a K-dimensional space and reversing the sign for simplicity, we define the Marker Index for a variable i on a factor k as the complement of the Euclidean distance between the variable represented by point a_i with coordinates $(a_{i1}, a_{i2}, .., a_{iK})$ and the vertex of the *k*th factor, represented by the point f_k with the *k*th element equal to 1 and all the other coordinates equal to zero.

Formally (see Appendix A for alternative formulations of the Marker Index):

$$MI_{ik} = 1 - \|f_k - a_i\| = 1 - \sqrt{\sum_{j=1}^{K} \left(f_{kj} - a_{ij}\right)^2} \tag{1}$$

or equivalently:

$$MI_{ik} = 1 - \sqrt{\left(1 - a_{ik}\right)^2 + \sum_{j=1}^{k} a_{ij}^2}, \ \text{j} \neq k \tag{1b}$$

The Marker index can be used both on standardized or raw variables (i.e., factor analysis on a correlation or covariance matrix). However, in this latter case, the standardized solution should be used to compute it.

Properties of the Marker Index as a Criterion

The index shows several properties that are described analytically in Appendix B, whereas in Appendix C a SAS routine is reported for ease of implementation. We can highlight intuitively these properties with the aid of Figure 2, where different variables are represented in a two-factor solution for simplicity.

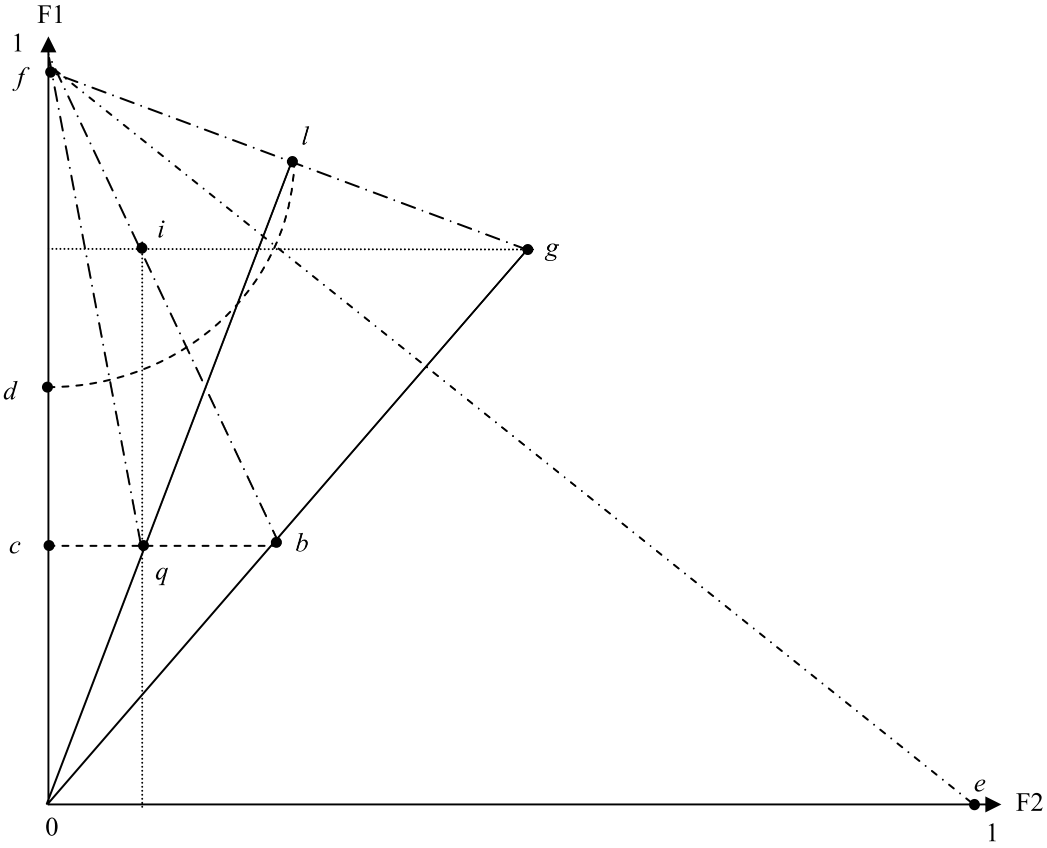

FIGURE 2
Geometrical representation of the Marker Index for variables with different properties.

First, the Marker Index is perfect ($MI_i = 1$) if and only if the variable coincides with the factor (point *f* in Figure 2) and has the minimum ($MI_i = 1 - \sqrt{2}$) when the variable coincides with a different factor (point *e* in Figure 2). Second, the index captures the simplicity of the variables. When two variables have equal primary factor loadings, the one with the lower secondary loading is preferred. In Figure 2, for instance, points *q* and *b* have the same primary loading, but q has a higher Marker Index since it has a lower secondary loading. Third, the index depends on the representativeness of the variables. When two variables have the same secondary loadings, the variable with the highest primary loading is preferred. In Figure 2, points *q* and *i* have the same secondary loading, but *i* has a higher Marker Index since it has a higher primary loading. Fourth, the index is a trade-off between simplicity and representativeness. Different combinations of primary and secondary loadings can produce the same Marker Index. These values are all points of the semi-circle whose ray-vector is a given Marker Index value (points *d* and *l* in Figure 2; see also below). Finally, for the sake of comparison with other methods, we note that the Marker Index distinguishes between two variables that are maximally simple by preferring the variable with the highest primary loading (points *d* and *c* in Figure 2). Moreover, the index distinguishes between variables with equal angular distances to the factor axis (variables *b* and *g* in Figure 2), achieving higher values for variables with higher communalities (point *l* over *q* and point *g* over *b* in Figure 2).

Summing up, the Marker Index weights and achieves an optimal compromise between the two main properties of variables (and of the resulting factorial structure): simplicity and representativeness. Since the Marker Index accounts for the absolute value of the primary loadings of the variables as well as for the ratio between the primary and the secondary loadings, the selected variables are both representative of the factors and factorially simple and, consequently, generate valid and reliable scales.

Interpretation of the Marker Index

We briefly give an interpretation of the feasible values of the Marker Index. First, note that the index ranges from 1 to $1-\sqrt{2}$. When it is positive, it has a straightforward interpretation: given a variable i with $MI_{ik}=v$, this variable is as good as a variable i' with a primary factor loading equal to v and the secondary factor loadings all equal to zero. This means that if *i* shows a higher Marker Index than i', *i* is either more representative or simpler. This property can be seen in Figure 2, where variable *l* and *d* have the same Marker Index. Variable *l* is less simple than *d*, but it is more representative of the factor. Finally, when $MI_{ik}<0$, the variable is less representative and less simple than a variable with $a_i=0$. Second, note that for each variable there are as many Marker Indices as factors. However, only the highest of them should be considered, since an interpretation of its values is meaningful only by reference to the primary loading of each variable.

To illustrate representative values of the Marker Index for different combinations of primary and secondary loadings, we reported a three-dimensional solution in Table 1 (for simplicity, we report only the Marker Index for the first factor).

Variable V_1 is a good marker for Factor 1. MI_{11} is .68. This value is obtained, according to Equation 1, as $1-\sqrt{(1-.70)^2+.10^2+.06^2}$. Variable V_2 shows the same primary loading but it is related also to the other two factors. This lack of simplicity decreases MI_{12}, which goes to .42 ($1-\sqrt{(1-.70)^2+.35^2+.35^2}$). In terms of the Marker Index, V_2 is considered as good as V_3, a very simple variable with a medium size loading. Compared with V_3, variable V_4 shows a lower MI_{14} since the primary loading has a medium size and the secondary factor loadings are higher. V_3 would be ranked even better than V_6, which shows a higher factor loading but also higher secondary loadings. Variable V_5 has a negative secondary loading that does not affect the Marker Index ($MI_{51} = 1 - \sqrt{(1-.60)^2 + (.10)^2 + (-.10)^2}$). Variable V_7 has high primary loading but the variable loads both on the first and the second factor, lacking in factorial simplicity, and therefore the Marker Index is low ($MI_{71} = 1 - \sqrt{(1-.70)^2 + (.70)^2 + (.01)^2}$). Finally, variable V_8 shows a negative Marker Index because it does not belong to the first factor $MI_{81} = 1 - \sqrt{(1-.21)^2 + (.70)^2 + (.10)^2} = 1 - 1.06$. Indeed, V_8 is a good marker of the second factor $MI_{82} = 1 - \sqrt{(.21)^2 + (1-.70)^2 + (.10))^2} = 1 - .38 = .62$.

TABLE 1
Illustrative values of Marker Index for a three-dimensional solution

	a_1	a_2	a_3	h^2	MI_{i1}
V_1	.70	.10	.06	.50	**.68**
V_2	.70	.35	.35	.74	**.42**
V_3	.42	.01	.01	.18	**.42**
V_4	.40	.31	.22	.30	**.29**
V_5	.60	.10	–.10	.38	**.58**
V_6	.60	.50	.40	.77	**.25**
V_7	.70	.70	.01	.98	**.24**
V_8	.21	.70	.10	.54	–**.06**

Note. For simplicity the Marker Index is computed only for the first factor.

To give a practical cut-off, we suggest the value .40 as a minimal value for the Marker Index. This guarantees keeping variables as useful as a perfectly simple variable with a primary loading of .40, one of the commonly used cut-off for the primary factor loadings. Applying this criterion to Table 1, one would select V_1 V_2 V_3 and V_5, that indeed are variables showing good factorial properties.

Practical Use of the Marker Index

Given its geometrical simplicity, the Marker Index can be used to select marker variables even by practitioners who feel uncomfortable with statistical computations. In two-factor solutions, in fact, there is not even the need to undertake any calculation to select a set of maker variables. It suffices to obtain the factor loadings plot (as in Figure 3) and draw two circles of the preferred length centered on the factor vertexes. All the variables that lie within the circles are marker variables (given the length chosen as a threshold), and all the variables that are outside the circles are not. For solutions with more than two factors, the same procedure can be repeated on all bidimensional plots, considering a variable as marker of one factor when it lies in all the circles centered on that factor's vertexes. If the aim is to select a given number (e.g., the best 10) of markers for each factor rather than whatever variable is above a certain threshold value, one can proceed by drawing larger circles until the required number of markers are selected.

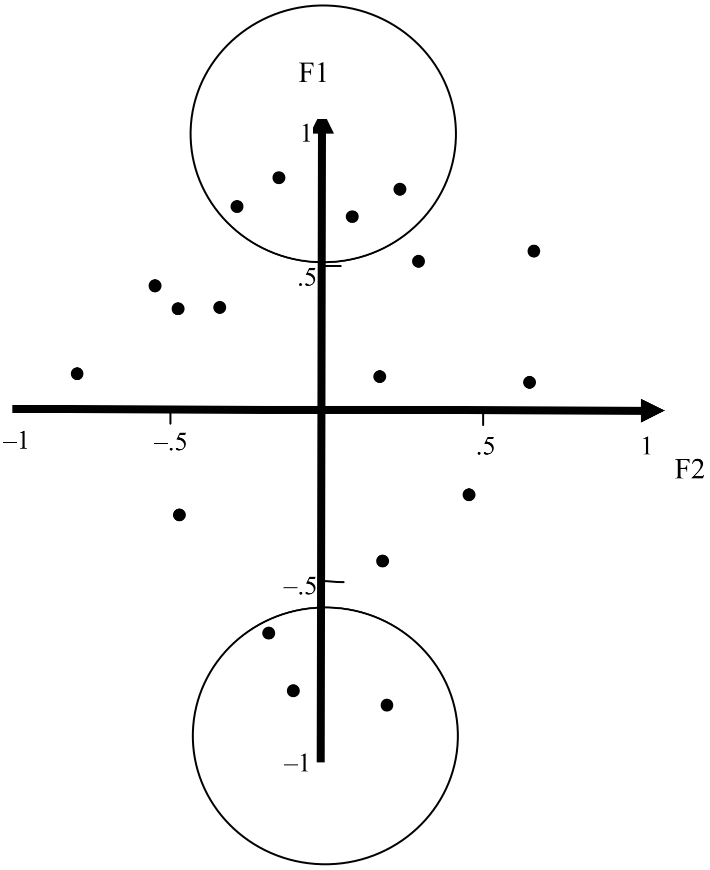

FIGURE 3
Practical selection of marker variables without calculations.

Generalization to Oblique Solutions

The Marker Index can be easily adopted also when the extracted factors are obliquely rotated with a promax or oblimin rotation. Oblique rotations do not change any of the properties of the Marker Index, because the definition of a marker variable does not depend on the correlation between factors. When factors are correlated, in fact, the point with co-ordinates (1,0) still represents the ideal marker, and therefore the distance between the actual and the ideal position of the variable is still the best index of representativeness and simplicity. When factors are correlated, however, the distance between the variable and the vertex of the factor should be computed using the generalized Euclidean distance, which takes into the account the correlation among factors (Appendix A describes in detail the computation of the Marker Index for oblique solutions). The Marker Index can be computed using both the *primary factor pattern* matrix or the *reference factor structure*, depending on the specific applications and on theoretical considerations. In general, researchers should compute the Marker Index in oblique solutions using the matrix they would normally use for selecting the variables (typically the pattern matrix).

ALTERNATIVE SELECTION METHODS

In the following sections we discuss existing alternative methods of marker selections, based either on factor loadings or on factor weights, in order to compare them to the Marker Index. The methods are: 1) a method based on the primary factor loadings (PL); 2) a method based on the ratio between the primary and the secondary loadings, that is the angular distance between the variables and the factors (AD); 3) a method based on the primary factor weights (PW); 4) a method based on the Varimax rotated factor weights (RW; Ten Berge & Knol, 1985).

Each method of selection has the following structure. We start from a pool of variables, which in most applications represent the items of a questionnaire or the set of all the measures available to the researcher. A factor analysis is carried out on these variables, and a (usually rotated) factor solution with *K* factors is retained. A method of marker selection identifies, for each factor k, a small number of variables according to some criterion of usefulness. Each method thus implies two decisions, one about the usefulness and the other about the number of variables to be selected.

The two main sources of information regarding the usefulness of the variables with respect to the factors are factor loadings and factor weights. Therefore, the four methods that we discuss are either a function of the factor loadings or the factor weights. Because a criterion of usefulness indicates how good a variable is in measuring a factor, the criterion is also used to assign variables to the factor.

As regards the number of variables to be selected, there are two different approaches. One can define a fixed number (*m*) of variables to be selected for each factor and keep the best *m* variables. Alternatively, one can define a minimal threshold of the criterion such that only the variables showing a criterion larger than the threshold are retained. In this contribution we do not discuss this issue any further, because both strategies equally benefit from a selection criterion that selects the best variables.

Method 1: Primary Factor Loading

The most-commonly used method to evaluate and to select variables is based on the primary factor loading, that is, on selecting the variables with the highest (absolute) loadings on a factor. In looking for the best markers of a factorial solution, authors often select, for each factor, the *m* variables with the highest factor loadings (cf. Goldberg, 1992). In scale construction, items are often evaluated in terms of their primary factor loadings. In fact, if we select variables with loadings higher than .30 (or sometimes .40), we are dealing with the Little Jiffy method for scale construction (Kaiser, 1974; Nunnally, 1978).

We can now focus on the properties of the set of the selected variables. The method is particularly suited to maximize the representativeness of the variables, but it does not guarantee a simple structure. In fact, the method does not consider the secondary loadings and therefore does not discriminate among complex and simple variables. This drawback also implies a weakness in the domain of scale construction. Since the selected variables for a factor may correlate with other factors (due to high secondary loadings), the resulting scale may not correspond to the original factor. The consequence is that scale validity and factor-trueness of the scale will be low (cf. Cattell, 1978; Ten Berge & Knol, 1985).

Summing up, the method is expected to be efficient in selecting representative variables, but deficient in respect to criteria, such as factorial simplicity, scale validity and independence, that are affected by the presence of high secondary loadings.

Method 2: Loading Ratio

To overcome the problems associated with the previous method, Fürntratt (1969) proposed an approach based on the ratio among primary loading and communality $\left(\frac{a_i^2}{h_i^2}\right)$. The basic idea is that a marker variable should be factorially simple, that is it should ideally load only on one factor. The simplicity of a variable can therefore be indicated by the relative amount of variance in a variable $\left(a_i^2\right)$ exclusive to the factor where the variable has the highest loading as compared to the total variance (or communality) in the reproduced factorial space $\left(h_i{}^2\right)$. The value ranges between 0 (the variance of a variable does not belong to a given factor) and 1 (all the variance belongs to a given factor). This index is equivalent to the square of the angular distance. Namely, the distance, in radians or degrees, between a variable and a factor. This method is also (empirically) equivalent to Kaiser's index of factorial simplicity, which also considers the ratio among the primary loadings and the communality of a variable. Since all those methods are almost equivalent and are all meant to maximize the simplicity of the selected variables, in the following sections we refer only to the angular distance.

This method relies on the notion that, representing each variable as a vector in the factorial space, the closer the variable vector is to the factor axis, the simplest, and thus the better, is the variable. In Figure 2, for instance, variables *q* and *l* show the same angular distance. Considering the angular distance, variable *q* should be considered as good as variable *l*, even though it is intuitively evident that *l* is a better marker than *q*. The same reasoning applies for variables *b* and *g*. (cf. Figure 2).

The angular distance is computed as follows:

$$AD_{ik} = \sqrt{\frac{a_{ik}^2}{h_{ik}^2}} \tag{2}$$

This method positively solves the problem outlined for the method of primary factor loading, because when two variables have the same primary loadings, the method prefers the one with lower secondary loadings and, for equal secondary factor loadings, the method prefers the one with higher primary loading. On the other hand, the ratio between the loadings does not consider the absolute value of the primary loading. Thus, this method is inefficient in selecting variables highly representative of the factor. For example, in a two-dimensional factorial solution two variables will have the same angular distance if they show, for instance, loadings as $\mathbf{a}_i = (.70,.10)$ and $\mathbf{a}_j = (.77,.11)$, but also when the loadings are $\mathbf{a}_i^* = (.14,.02)$and $\mathbf{a}_j^* = (.77,.11)$. Obviously, no user of factor analysis would say that the latter two variables are equivalently good. Comparing this method with the method of primary factor loading, it appears that what one gains in factorial simplicity and scale validity is lost in the representation of the factors.[2]

When the variables are selected in order to obtain scales as an estimation of the corresponding factors, all methods based on factor loadings suffer from a general criticism. In fact, it has been pointed out by Ten Berge and Knol (1985) that factor loadings do not represent the contribution of the variables in forming the factor. Hence, their use to evaluate their contribution in forming the factor has no strict theoretical (and mathematical) justification. The proper parameter upon which to base a selection criterion should be the factor weights, representing the contribution of a variable in forming a factor. We can describe two methods of selection based on factor weights. The two methods apply the logic of two procedures proposed in the literature for constructing scales (cf. Gorsuch, 1974; Ten Berge & Knol, 1985).

Method 3: Primary Factor Weight

The first method is simply the method of primary factor loading substituting the factor loadings with the factor weights. The use of the absolute factor weights guarantees that the variables will produce scales with high scale validity (Ten Berge & Knol, 1985). Furthermore, since the factor weights increase with increasing correlations between variables and factors, the method will select variables that are representative of the factors. However, this method does not guarantee factor-homogeneity, since it is insensitive to secondary loadings.

Method 4: Rotated Primary Factor Weight

To overcome the latter problem, the method can be modified with the following rationale (cf. Ten Berge & Knol, 1985). In selecting variables contributing to only one factor, it is desirable to have variables with the simplest possible weights. Therefore, it would be better to consider the weights resulting from a rotation of $\mathbf{W}$ to obtain a simple structure (cf. Harris, 1975). To sum up, the selection criterion should be based on the weights $\overline{\mathbf{W}}$, resulting by the Varimax rotation of the weights $\mathbf{W}$. This modified approach does not suffer of particular deficiencies from a logical point of view.

EMPIRICAL COMPARISON AMONG METHODS

In the previous analyses we have shown the formal properties of the Marker Index as compared with other methods of selection. We have argued that in several conditions, the Marker Index maximizes the accuracy of the selection as it is sensitive to many relevant characteristics of the variables that other methods fail to consider. In this section we test these predictions by comparing the performance of the five methods in selecting marker variables. With this aim, we conducted a series of empirical studies. We tested the performance of the methods on a series of generated data sets, formed by normal random variables generated from a factorial model. The model allows us to control the characteristics of the data sets and to vary systematically the parameters of the model (e.g., number of cases, variables, factors). For the sake of comparability with other methods of selection, we focus on orthogonal solutions as obtained with Varimax rotation, which represents the most-commonly used rotation method in psychology.

Empirical Criteria

The following criteria have been chosen to reflect desirable properties both of the factorial solution and of the scales resulting by summing the selected variables for each factor.

1. *Factor representativeness*. As a measure of factor representativeness we used the percentage of explained variance for each factor. In fact, the more the selected variables are representative of the factor, the higher the variance shared by the variables and the factor.

2. *Factorial simplicity*. As a measure of factorial simplicity, we used the Index of Fit for Factor Scale (IFFS) proposed by Fleming (1985). The index is particularly suited to measure the simplicity of sets of variables conceived of as scales of the factors. The index is computed as the ratio between the mean squared loadings on one factor of the variables of interest and the mean squared loadings of the same variables on the other factors. When the variables are maximally simple the index is equal to 1, whereas when they are maximally complex the index is equal to .5. If the index is zero the scale variables do not measure the factor they are expected to measure (Fleming, 1985).

3. *Scale consistency*. As a measure of internal consistency we used Cronbach's alpha coefficient.

4. *Scale validity*. As a measure of scale validity, we used the correlation between the scales obtained adding the variables selected according to the assignment criterion and the factor scores (cf. Ten Berge & Knol, 1985).

5. *Scale independence*. As a measure of factor independence we used the determinant of the correlation matrix among scales. Since the extracted factors are orthogonally rotated, the less the scales correlate, the more they measure the factors properly. The determinant gives an overall compact value of orthogonality. It is equal to 1 if the scales are perfectly orthogonal, and it is equal to zero if the scales are perfectly correlated.

Data Generation

We constructed a series of data sets following the factorial model described in Kiers (1997). Specifically, let the $n \times r$ matrix **Y** be the artificial sample (the data set) of n cases and r variables to be generated. Let k be the number of components decided a priori to underlie the data structure, the $n \times k$ matrix **F** be the matrix with k normally distributed variates representing the factor scores, **P** be the $r \times k$ pattern matrix, and **E** be the $n \times r$ matrix of normally distributed unique terms of the variables. Then, $\mathbf{Y} = \mathbf{FP'} + \mathbf{E}$.

This method allows us to control, besides the other parameters of the data sets, the specific relationships between the variables and the underlying factors. In order to simulate data structures where the underlying dimensions are known, but the assignment and the quality of the variables with respect of the dimensions are not unique, the a priori factor loadings (the pattern matrix **P**) are generated according to a uniform distribution, ranging from –1 to 1. Thus, each column of the pattern matrix is a uniform distribution, independently of the other columns. Because the loadings are uniform, each factor has average loadings of .5 (in absolute value). Because the loadings of one variable on different factors are independent, this method produces factor structures with variables that vary both in simplicity and representativeness.

As far as the parameters of the generated data are concerned, we conducted the study varying the following experimental variables: 1) number of factors *k*, with *k* = 2,4,6,8,10,12; 2) number of variables for each factor (*r/k*): 15,20; 3) number of cases *n*, with *n* = 500, 1000; 4) number of markers to be selected (*m*): 4,8.

Procedure

For each set of variables produced, we obtained an orthogonal Varimax rotated factor solution of *k* factors from the *r* variables. All the solutions have been obtained using principal component analysis. We also replicated the results varying the factor method (i.e., minimal residual and maximum likelihood factor analysis), without significant differences in the results. We selected a set of *m* variables for each method of selection. For each selected set we performed a new factor analysis and evaluated the criteria. To minimize the risk of random effects, we replicated the analysis with 10 repetitions for each combination of experimental variables.

The experiment is therefore based upon 480 generated data sets, obtained as follows: six (number of factors) X 2 (number of variables) X 2 (number of cases) X 2 (number of markers) X 10 (repetitions).

RESULTS

The factorial method and number of cases did not influence remarkably the relative performance of the methods, thus we aggregated the results across these variations. Although our main aim was to investigate the overall performance of the methods in producing good sets of markers, first we comment briefly on the performance of the methods for each criterion across different parameters.

Success Criteria

1. *Factor representativeness*. As shown in Table 2, the set of variables selected using PL and PW explains the highest percentage of variance. This confirms empirically our argument about the efficiency of this method in maximizing the representativeness of the variables.

2. *Factorial simplicity (*IFFS). As expected, the AD performs better than all other methods. Interestingly, the MI's performance is the second best, better than other methods that are not explicitly designed to maximize this criterion. This shows that MI efficiently takes into account the simplicity of the variables.

3. *Scale consistency.* Concerning reliability obtained adding the selected variables, the best method is PL, with MI and the methods based on the factor weight performing well. Whereas AD shows a poor performance.

TABLE 2
Success criteria for the five selection methods

Selection method	# studies	Representat.	Simplicity	Reliability	Validity	Independence
MI	480	37.6	.944	.607	.760	.953
		(5.9)	(.032)	(.068)	(.057)	(.042)
PL	480	37.7	.940	.608	.759	.950
		(6.0)	(.035)	(.068)	(.057)	(.044)
AD	480	35.9	.949	.579	.750	.961
		(5.5)	(.033)	(.079)	(.056)	(.036)
PW	480	37.7	.940	.607	.759	.952
		(5.9)	(.034)	(.068)	(.057)	(.043)
RW	480	37.6	.941	.606	.759	.953
		(5.9)	(.034)	(.068)	(.057)	(.042)

Note. Representativeness = % of explained variance; Simplicity = Index of Fit for Factor Scale; Reliability = Cronbach alpha; Validity = average correlations between factor scores and corresponding scales; Independence = determinant of the correlation matrix among scales; MI = Marker Index; PL= Primary Loadings; AD= Angular Distance; PW = Primary Factor Weight; RW= Rotated Primary Factor Weight. Standard deviations are in parentheses.

4. *Scale validity.* The MI performs better than all the other indexes. As expected, MI guarantees the selection of variables with high representativeness and high simplicity. Since these two properties constitute the basic ingredients of validity, their achievement is mirrored in the empirical results.

5. *Scale independence.* AD performs better than the other methods. This result is consistent with the logic underlying this method, namely the maximization of the simplicity of the solution.

The overall results indicated that different methods lead to the maximization of a specific criterion. However, the average performance tends to hide differences among methods because possible discrepancies in the performances of the methods for specific combinations of the experimental parameters can be compensated by the performance in other combinations. Furthermore, users of factor analysis rarely aim for a unique criterion, as we extensively argued. They wish to have the best compromise between the simple and the representative structure. Moreover, once they have selected a set of variables using a given criterion, they may be more interested in the properties of the factorial structures or in the properties of the resulting scales. We analyzed the average performance of each method considering these two aspects, the number of extracted factors, and the proportion of variables selected as markers, since both appeared to be relevant sources of variation in the success criteria.

With this aim, we standardized each success criterion over the 480 outcomes resulting from the 48 combinations. To minimize the impact of the performance of the methods for some specific combinations of parameters, we standardized the data within each of the major experimental variables. This allows us to evaluate the overall performances unbiased with respect to the overall variability not due to the methods.[3] To simplify the detailed discussion of the empirical performance of the methods, we considered the results from two perspectives, the quality of the factorial structure of the quality of the resulting scales. We will call the first *structural good-*

ness, and the second *scale goodness*. Structural goodness results from adding the standardized score relative to the percentage of explained variance and the IFFS, whereas scale goodness results by adding the standardized alpha, scale validity, and scale independence. We also considered the overall performance, by calculating the average of the two aggregated criteria. Thus, positive values of these indices indicate comparatively high performance, and negative values comparatively poor performance (see Table 3).

TABLE 3
Aggregated performance of the methods (z-scores)

Selection method	# studies	Goodness indices		
		Structure	Scale	Overall
MI	480	.214	.159	.186
PL	480	.027	.112	.069
AD	480	–.285	–.515	–.400
PW	480	.032	.136	.084
RW	480	.013	.109	.061

Note. Structure = Representativeness, Simplicity (sum of z-scores). Scale = Reliability, Validity and Independence (sum of z-scores). Overall = mean of structure index plus scale index. MI = Marker Index; PL = Primary Loadings; AD = Angular Distance; PW = Primary Factor Weight; RW = Rotated Primary Factor Weight.

The results show that the MI performs better than all other methods, both considering the overall performance and considering structural goodness and scale goodness separately. As expected, the MI guarantees a good performance because it considers both the structural properties of the variables (simplicity and representativeness) and the pureness of the variables (and consequently the validity and the trueness of the resulting scales).

Considering the performance as a function of the number of factors (Table 4), the MI performs (relative to the other methods) at best with more factors, being by far the best method according to all criteria. For each number of factors we considered, MI shows the best performance for structural goodness, scale goodness, and consequently for overall goodness (see also Figure 4). As the number of factor increases, the differences in performance do not change markedly, even though we observe a slight increase in the relative performance of the Marker Index as compared with all other methods.

Concerning the performance as a function of the number of markers selected, we examined the relative performance of the methods for varying proportions of selected variables over the initial number of variables. The first experimental variable included two levels, 4 and 8 selected variables, and the second experimental variable included two levels, 10 and 15 initial variables per factor. The combination of these two variables produces a new experimental variable that indicates the proportion of initial variables selected as markers. There were four proportions: .26 .53 .40, and .80. Figure 5 shows the overall performance as a function of the proportion of markers. The Marker Index appears to perform better than any other method across all the proportions we investigated, especially when few variables are selected. As the proportion of markers increases, the relative performance of the methods becomes more similar. This effect is due to the fact that, when 80% of the variables are selected as markers, the overlapping of the selections increases, because few variables are discarded, and all the methods tend to agree upon the exclusion of the worst variables. When only few variables are selected, the overlapping decreases, and thus the quality of the methods are more distinguishable. Additional analyses show that the same effect can be found for structural goodness and scale goodness, with MI performing better than the other methods for both criteria.

TABLE 4
Structural, scale, and overall performance of the methods by number of factors (*z*-scores)

Selection method	# factors	Goodness indices		
		Structure	Scale	Overall
MI	2-4-6	.149	.148	.149
PL	2-4-6	.022	.110	.066
AD	2-4-6	–.234	–.505	–.370
PW	2-4-6	.024	.125	.074
RW	2-4-6	.039	.122	.081
MI	8-10-12	.278	.170	.224
PL	8-10-12	.032	.113	.073
AD	8-10-12	–.337	–.525	–.431
PW	8-10-12	.040	.146	.093
RW	8-10-12	–.013	.096	.041

MI = Marker Index; PL = Primary Loadings; AD = Angular Distance; PW = Primary Factor Weight; RW = Rotated Primary Factor Weight.

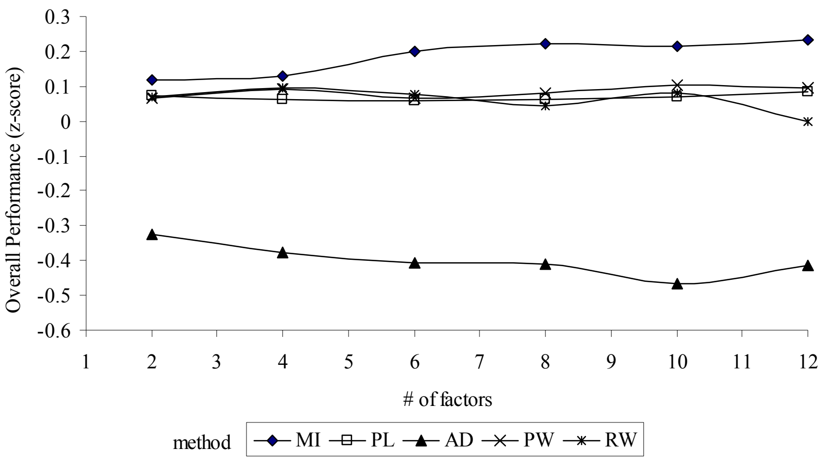

FIGURE 4
Overall performance of the selection methods across different numbers of factors.
Data for each method are standardized within experimental conditions.

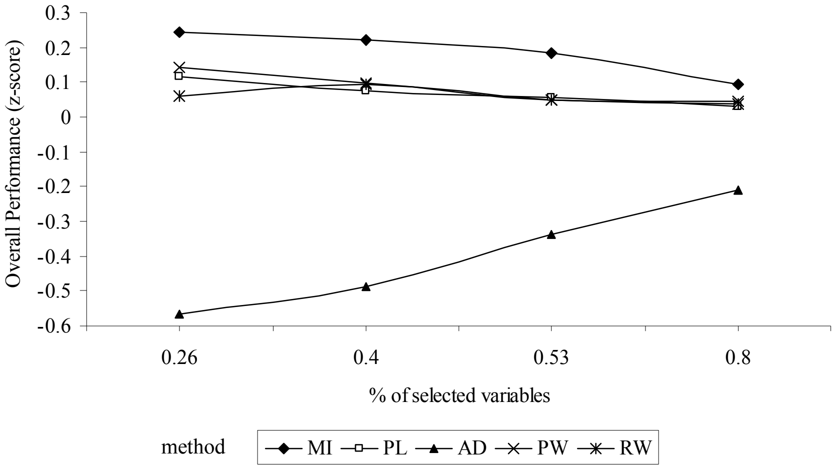

FIGURE 5
Overall performance of the selection methods across different proportions of selected variables over total number of variables. Data for each method are standardized within experimental conditions.

Overall, the Marker Index is the undisputed winner of this empirical comparison with other methods for each combination of parameters we have considered. It is important to note that this overall superiority of the Marker Index appears in our data *despite* the fact that the data have a clear and regular structure. These regularities are likely to hide differences among methods, because, for well-defined structures, performance tends to converge. In other words, because in our data parameters such as the number of factors, correlations among factors, distribution of the variables, and distribution of loadings are precisely defined in advance, the methods do not suffer from weakness due to the uncertainty in these parameters. In real data, however, the underlying structure is rarely known in advance, and greater irregularity is to be expected in the distribution of the variables and of the parameters. The logical and geometrical properties of the Marker Index ensure that its performance would hold also when such irregularity occurs. Furthermore, we have additionally performed some ancillary empirical comparisons among the selection methods using real data in personality research.[4] We have found again that Marker Index is superior to all other selection methods considered in this contribution. Remarkably, in real data the superiority of the Marker Index is even more pronounced that in simulated data.

As a final remark, we wish to note that the findings we obtained are in perfect agreement with and follow from the theoretical analyses that we have conducted. In each data set and across all data sets, all the selection criteria maximize the property that we expected from the theoretical analyses. As expected, the best overall performance is observed for the Marker Index, which provides the best compromise among the criteria. The consistency of the findings and the type of data we have considered suggest that the results we have obtained are not due to the specificity of the data sets.

CONCLUSIONS

The Marker Index appears to be the winner of this contest. It satisfies many criteria both for scale construction and for structural goodness. The Marker Index shows a remarkable performance for different purposes and in different conditions. The empirical strength, furthermore, is well supported by the logic and the rationale behind this method. We maintain that the superiority of the Marker Index is not an empirical property but it is due to analytical and logical properties and therefore it is also reflected in empirical findings. In other words, due to its reliance on the Euclidean theorem, the Marker Index necessarily represents the best measure of the distance between an ideal and an actual point. This distance is the complement of how close a variable is to be a perfect variable for a given factor. The method is simple to implement, and it has a direct and intuitive interpretation. In practice, users of factor analysis are mostly looking for a compromise between different properties, and so far they could rely just on rules-of-thumbs and experience. The Marker Index represents a simple algorithmic solution to a very old problem.

NOTES

1. The authors' order is alphabetical, as both authors have contributed equally. Correspondence concerning this manuscript can be addressed also to Marco Perugini, Facoltà di Psicologia, Università degli Studi di Milano-Bicocca, Viale dell'Innovazione 10 (U9), 20126 MILANO (MI), Italy. E-mail: marco.perugini@unimib.it
2. Perugini and Leone (1996) tried to overcome the logical problems of the methods discussed so far by proposing an index of prototypicality that combined both methods. While that index is superior to both primary factor loading and angular distance, it is inferior to the Marker Index and it is less elegant. For this reason the index of prototypicality will not be considered in this contribution.
3. Since all methods perform better for a small number of factors, with few markers selected and with more initial variables, we operated a standardization within experimental variables such that the results are not affected by those differences that are independent of the differences between methods.
4. The results are not reported here both for the sake of brevity and because the data simulation approach that we have used in the manuscript is more precise and mathematically elegant.

REFERENCES

Cattell, R. B. (1978). *The scientific use of factor analysis*. New York: Plenum Press.

Cattell, R. B., & Tsujioka, B. (1964). The importance of factor-trueness and validity, versus homogeneity and orthogonality, in test scales. *Educational and Psychological Measurement*, *24*, 3-30.

Comrey, A. L., & Lee, H. B. (1992). *A first course in factor analysis*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt, Rinehart and Winston.

Dillon, W. R., & Goldstein, M. (1984). *Multivariate analysis: Methods and applications*. New York: Wiley.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, *4*, 272-299.

Fleming, J. S. (1985). An index of fit for factor scales. *Educational and Psychological Measurement, 45,* 752-728.

Fürntratt, E. (1969). Zur Bestimmung der Anzahl interpretierbarer gemeinsamer Faktoren in Faktorenanalysen empirischer Daten. *Diagnostica, 15*, 62-75.

Goldberg, L. R. (1992). The development of markers of the Big Five factor structure. *Psychological Assessment, 6,* 26-42.

Gorsuch, R. L. (1974). *Factor analysis*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Guilford, J. P. (1954). *Psychometric methods*. New York: McGraw-Hill.

Harman, H. H. (1976). *Modern factor Analysis (3rd edition)*. Chicago: University of Chicago Press.

Harris, R. J. (1975). *A primer of multivariate statistics*. New York: Academy Press.

Horst, P. (1965). *Factor analysis of data matrices*. New York: Holt, Rinehart & Winston.

Kaiser, H. F. (1974). An index of factorial simplicity. *Psychometrika, 39*, 31-35.

Kiers, H. A. (1997). Techniques for rotating two or more loading matrices to optimal agreement and simple structure: A comparison and some technical details. *Psychometrika*, 62, 545-568.

Kline, P. (1993). *The handbook of psychological testing*. London: Routledge.

Loehlin, J. C. (1987). *Latent variable models: An introduction to factor, path, and structural analysis*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Nunnally, J. (1978). *Psychometric theory*. New York: McGraw-Hill.

Perugini, M., & Leone, L. (1996). Construction and validation of a Short Adjectives Checklist to measure Big Five (SACBIF). *European Journal of Psychological Assessment*, *12*, 1-10.

Rummel, R. J. (1970). *Applied factor analysis*. Evanston, IL: Northwestern University Press.

Stevens, J. (1996). *Applied multivariate statistics for the social sciences*. Mahwah, NJ: Lawrence Erlbaum Associates.

Tabachnick B. G, & Fidell, L. S. (1996). *Using multivariate statistics (2nd edition)*. New York: Harper Collins Publishers.

Ten Berge, J. M. F., & Knol, D. L. (1985). Scale construction on the basis of components analysis: A comparison of three strategies. *Multivariate Behavioral Research, 20*, 45-55.

APPENDIX A

Different Formulations of the Marker Index

Orthogonal solution. For computational purposes we give alternative but equivalent formulations of the Marker Index .

Matrix form: $MI_{ik} = 1 - \|\mathbf{f_k} - \mathbf{a_i}\| = \mathbf{1} - \sqrt{(\mathbf{f_k} - \mathbf{a_i})'(\mathbf{f_k} - \mathbf{a_i})}$

Scalar form: $MI_{ik} = 1 - \sqrt{(1 - a_{ik})^2 + \sum_{p=1}^{k} a_{ip}^2}$, $\mathrm{p} \neq k$

Communality known: $MI_{ik} = 1 - \sqrt{1 - 2a_{ik} + h_i^2}$

Polar system:

Let transform the loadings of a variable i in a ray-vector (ρ) and an angular value (θ), following the formula (cf. Harman, 1976, p. 57): $\rho_i = \sqrt{\sum_k a_{ik}^2}$, $\theta_i = arc\tan\left(\frac{a_{ik}}{\rho_i}\right)$

The Marker Index is $MI_{ik} = 1 - \sqrt{\rho_i^2 + 1 - \rho_i \cos\theta_i}$

Oblique solutions. Let $\mathbf{\Phi}$ the **Errore. Non si possono creare oggetti dalla modifica di codici di campo.** correlation matrix among the factors. The Marker Index will be the complement of the generalized Euclidean distance between the variable i and the point f in the oblique axis system (cf. Harman, 1976, p. 61):

$$MI_{ik} = 1 - \sqrt{\sum_{v=1}^{K}\sum_{p=1}^{K}(f_{pk} - a_{pi})(f_{vk} - a_{vi})\phi_{vp}}$$

or, in vector notation $MI_{ik} = 1 - \sqrt{(\mathbf{f_k} - \mathbf{a_i})'\mathbf{\Phi}(\mathbf{f_k} - \mathbf{a_i})}$

APPENDIX B

Marker Index Properties

We now show formally some key properties of the Marker Index sketched in the main text, and we show how the Marker Index solves the cases where the other methods may fail. In the following statements, the loadings are implicitly referred to as always positive and the Marker Index of variable *i* on factor *k* is noted as mi_{ik}.

Properties.

1) Range: $m_{ik} = 1$ iff $a_i = f_k$ and $m_{ik} = 1 - \sqrt{2}$ iff $a_i = f_p$ with $p \neq k$.

This statement is intuitively true.

2) If two variables have the same primary factor loading, the variable with the lower secondary loading is preferred. Namely, if $a_{ik} = a_{jk}$ and $h_i^2 < h_j^2$ then $mi_{ik} > mi_{jk}$.

Proof: to prove this the Marker Index just needs to be written as

$$mi_{ik} = 1 - \sqrt{1 - 2a_{ik} + h_i^2} \tag{1b}$$

and note that

$$mi_{ik} > mi_{jk} \text{ if } h_i^2 - 2a_{ik} < h_j^2 - 2a_{jk} \tag{2b}$$

So, for any $a_{ik} = a_{jk}$ and $h_i^2 < h_j^2$, it follows that $mi_{ik} > mi_{jk}$.

3) Variables with equal secondary loadings: if two variables have the same secondary factor loadings, the variable with the highest primary loading is preferred. That is, if $h_i - a_{ik} = h_j - a_{jk}$ and $a_i > a_j$ then $mi_{ik} > mi_{jk}$.

Proof: note that

$$mi_{ik} > mi_{jk} \text{ if } (1 - a_{ik})^2 + \sum_{v=1}^{k} a_{iv}^2 < (1 - a_{jk})^2 + \sum_{v=1}^{k} a_{jv}^2 \text{ with } v \neq k \tag{3b}$$

Since $\sum_{v=1}^{k} a_{iv}^2 = \sum_{v=1}^{k} a_{jv}^2$, then

$$mi_{ik} > mi_{jk} \text{ if } (1 - a_{ik})^2 < (1 - a_{jk})^2 \tag{4b}$$

Since $0 \leq a_{ik} \leq 1$ hence

$$mi_{ik} > mi_{jk} \text{ if } a_{ik} > a_{jk} \tag{5b}$$

4) Variables maximally simple: if two variables are perfectly simple, the one with the higher primary loading is preferred. That is, if $a_{ik} > a_{jk}$, and $a_{ik} = h_i$, and $a_{jk} = h_j$, then $mi_{ik} > mi_{jk}$. Property 4 is a sub-case of Property 3.

5) Variables with equal angular distance: two variables with the same angular distance (with respect of a given factor) are considered equal only if $\left| \frac{a_{ik}}{h^2{}_i} - 1 \right| = \left| \frac{a_{jk}}{h^2{}_j} - 1 \right|$. That is, if $a_{ik} \neq a_{jk}$, and

$\frac{a_{ik}}{h_i} = \frac{a_{jk}}{h_j}$, we have $mi_{ik} > mi_{jk}$ if $\left|\frac{a_{ik}}{h^2{}_i} - 1\right| < \left|\frac{a_{jk}}{h^2{}_j} - 1\right|$, and we have $mi_{ik} = mi_{jk}$ if $\left|\frac{a_{ik}}{h^2{}_i} - 1\right| = \left|\frac{a_{jk}}{h^2{}_j} - 1\right|$.

Proof: we first show that a variable with $\frac{a_{ik}}{h_i^2} = 1$ is the variable with the highest Marker Index among the variables with the same ratio $\frac{a_{ik}}{h_i}$. Consider the 2-dimensional Cartesian space in Figure A1.

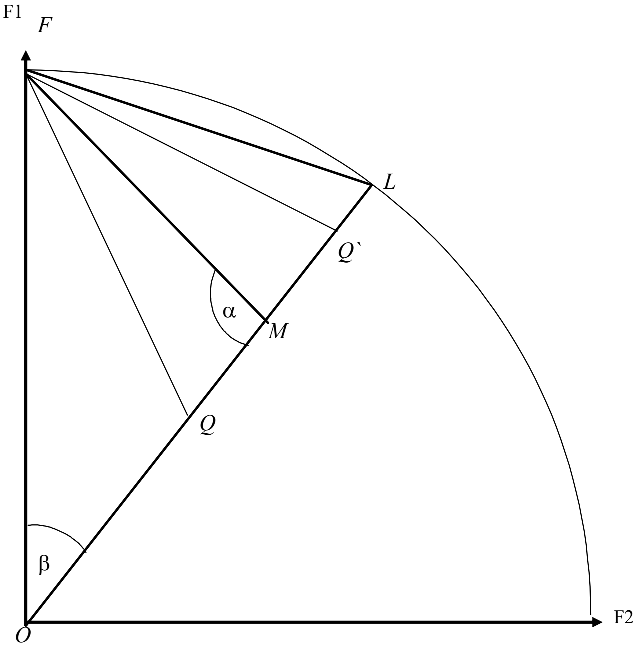

FIGURE A1
Geometrical representation of a two-factor solution.

The segment $\overline{OL}$ has unit length and it is separated to $\overline{OF}$ by the angle β. All the variables lying on $\overline{OL}$ have the same ratio $\frac{a_{ik}}{h_i} = \cos(\beta)$. Consider now the triangle OFL. Applying basic geometry, we know that the shorter line connecting *F* with $\overline{OL}$ is the height of the triangle, the segment $\overline{OM}$ orthogonal to $\overline{OL}$. The point *M* is the point with the shortest distance to *F* among all the points on $\overline{OL}$. Since OMF is a right triangle $\left(\alpha = 90^\circ\right)$, it follows that $\overline{FM}^2 = \overline{OF}^2 - \overline{OM}^2$.

Thus, because $\overline{FO} = 1$ and $\overline{OM} = h_m$, and $\overline{FM}$ is the Euclidean distance to the point *F* of the variable at point *M*, variable *M* has the highest Marker Index when

$$1+h_m^2-2a_{mk}=1-h_m^2 \tag{6b}$$

Which simplifies to:

$$\frac{a_{mk}}{h_m^2}=1 \tag{7b}$$

Consider now a variables q lying on $\overline{OL}$, with $\frac{a_{qk}}{h_q^2}\neq 1$. Its distance from F is $\overline{QF}=\sqrt{\overline{QF}^2=\overline{FM}^2+\overline{QM}^2}$, implying that the shorter $\overline{QM}$, the shorter $\overline{QF}$. Because $\overline{QM}$ is the distance between q and m, namely $\left|\frac{a_{qk}}{h_q^2}-1\right|$, the lower $\left|\frac{a_{qk}}{h_q^2}-1\right|$, the shorter will be $\overline{QM}$. Since $MI_q=1-\overline{QF}$, the lower is $\left|\frac{a_{qk}}{h_q^2}-1\right|$ the higher will be the Marker Index.

Finally, if we take $q`$ such that $\left|\frac{a_{qk}}{h_q^2}-1\right|=\left|\frac{a_{q`k}}{h_{q`}^2}-1\right|$, we have $\overline{QM}=\overline{Q`M}$ and thus $\overline{QF}=\overline{Q`F}$. Therefore, two variables *i* and *j* with $\frac{a_{ik}}{h_i}=\frac{a_{jik}}{h_j}$ have the same Marker Index if and only if $\frac{a_{ik}}{h_i^2}=\frac{a_{jik}}{h_j^2}$. This concludes the proof.

APPENDIX C

SAS Macro for Computing the Marker Index

The following simple SAS macro can be used to compute the Marker Index of each variable *i* for *K* factors, in orthogonal solutions. SPSS code and SAS macros for oblique solutions and for different strategies of selection are available from the authors on request. The macro operates on the raw data dataset. It should be submitted to the SAS system before executing it.

MACRO

```
%macro makerindex(data=_last_,out=mi,var=,nfactor=2,r=v);
/* Execute the factor analysis */
proc factor data=&data n=&nfactor r=&r out-
stat=l(where=(_type_="PATTERN")) ;
var &var ;
run;
proc transpose data=l out=l; id _name_; run;

/* Read the factor loadings and compute the marker indices */
data &out;
set l;
array Factor (&nfactor); /*array of factor loadings */
array mi (&nfactor); /*array of marker indices for each variable */
communality=uss(of factor(*)); /* Communality of the variable */
do i=1 to &nfactor;
mi[i]=1-sqrt(1-2*abs(Factor[i])+communality); /* Marker indices */
end;
drop i;
run;
Title "Marker indices";
proc print; var communality mi1-mi&nfactor; run;
title ;
%mend;
```

USAGE

```
%makerindex(data=_last_,out=mi,var=,nfactor=2,r=v);
```

where:
data = the raw data dataset containing participants’ scores (default = last dataset used);
out = the output dataset where the marker indices are stored (default = mi);
var = variables included in the factor analysis (required parameter);
nfacto r= number of factors to retain (default=2);
r = type of rotation (possibly none). Accept **proc factor** syntax (default = varimax).