Behavior Research Methods, Instruments, & Computers
2003, 35 (2), 318-321

IMINCE: An unrestricted factor-analysis-based program for assessing measurement invariance

URBANO LORENZO-SEVA and PERE J. FERRANDO
Rovira i Virgili University, Tarragona, Spain

In this article, a Windows program for analyzing measurement invariance in two different populations is described. Factor analysis is a common way of assessing measurement invariance, and restricted factor analysis is now the most popular method. However, applied researchers have usually found that the theoretical advantages of restricted factor analysis do not always apply in practical situations. For example, when the participant sample is large, as is the case in Internet-based questionnaires, the available software for restricted factor analysis might fail to converge on a solution. Our program is based on unrestricted factor analysis and considers the three parameters that define factor invariance: difficulties, discriminations, and residual variances. The statistical significance of the tests for evaluating invariance is obtained using Bootstrap resampling procedures. A real-life example demonstrates the usefulness of the program.

When we compare members of identifiable groups of individuals in terms of their trait levels, we must assume that the item and test scores that measure the traits have the same meaning in each group. To put it more formally, the scores earned by members of different groups are assumed to be on the same measurement scale (Drasgow, 1984). If this assumption is valid, the item and test scores are comparable, and the test has *measurement invariance* across groups. According to the *Standards for Educational and Psychological Testing* (American Educational Research Association, 1999), the assessment of measurement invariance is critical to sound testing practice, and so, much discussion and research has been devoted to this topic (see, e.g., Reise, Widaman, & Pugh, 1993).

Factor analysis (FA) is one of the most common ways of assessing measurement invariance. The conventional FA approach to examining this issue involves comparing the matrices of item-factor or test-factor regression weights of the different groups (see, e.g., Jöreskog, 1971). However, this procedure addresses only one aspect of invariance. The general FA model assumes that the regression of an item or test score on the factor depends on three parameters: intercept (i.e., difficulty), regression weight or factor loading (discrimination), and residual variance. Strictly speaking, therefore, in order for two item or test scores from two different groups to be comparable, the intercepts, factor loadings, and residual variances of this item or test must be invariant in both groups. Meredith (1993) called this condition *strict factorial invariance*. Following Meredith's terminology, invariance of the factor loadings would be *partial factorial invariance*, whereas invariance in both the intercepts and the factor loadings would be *strong factorial invariance*.

Historically, the FA assessment of measurement invariance has been addressed from the unrestricted (exploratory) FA model. However, according to Reise et al. (1993), the restricted (confirmatory) FA model is more often used nowadays. Restricted FA has important theoretical advantages over unrestricted FA, mainly because (1) it specifies a structural model that can be rigorously tested and (2) by choosing a suitable baseline model, we can assess different forms of measurement invariance (partial, strong, and strict) by means of hierarchical tests.

Applied researchers, however, have found that the theoretical advantages of restricted FA do not always apply in practical cases. For example, the formal tests of fit used in this model rely on assumptions that are difficult or impossible to fulfill when the variables to be analyzed are item scores (e.g., the assumption that the variables are continuous and unbounded). Furthermore, the standard restricted model assumes that most of the variables are factorially pure (i.e., they load on only one factor and have zero loading values on the remaining factors). In real applications, however, the items tend to have nontrivial secondary loadings on other factors. As some authors noted (Church & Burke, 1994; McCrae, Zonderman, Costa, Bond, & Paunonen, 1996), unrestricted FA-based procedures might be more appropriate than the restricted FA approach in most real applications, especially in large, multidimensional solutions that do not approach very simple structures. In addition, when the studied sample is large, the available soft-

This research was supported by the Dirección General de Investigación Científica y Técnica and the European Fund for Regional Economic Development (SEC2001-3821-C05-02). Correspondence concerning this article should be addressed to U. Lorenzo-Seva, Universitat Rovira i Virgili, Facultat de Psicologia, Carretera Valls s/n 43007 Tarragona, Spain (e-mail: uls@fcep.urv.es).

Table 1
Item Difficulties, Univariate *t* Tests, and Cohen's *d′* Statistic

Item Number	Target Sample	Replication Sample	Student's *t*	Effect Size (Cohen's *d′*)
1	3.42	3.13	4.51*	0.30
2	3.42	3.04	5.25*	0.35
3	3.84	3.82	0.28	0.02
4	2.45	2.42	0.40	0.03
5	2.06	2.12	−0.85	−0.06
6	2.55	2.68	−1.87	−0.12
7	3.07	2.78	3.09*	0.20
8	2.71	2.86	−1.94	−0.13
9	2.88	2.94	−0.72	−0.05
10	2.78	2.73	0.66	0.04

*Significant difference.

ware for restricted FA might fail to converge on a solution. Large participant samples are usually obtained, for example, in Internet-based questionnaires (see, for example, Buchanan & Smith, 1999; Joinson, 1999; Pasveer & Ellrad, 1998).

Because the conventional unrestricted FA approach is mainly descriptive, an important drawback of this model is that decisions are based on arbitrary rules of thumb. To overcome this, several more rigorous procedures have been proposed for assessing item or test invariance when an unrestricted FA approach is used. Some of these procedures are inferential and provide standard errors and test statistics, which give more information and eliminate arbitrariness. However, the relevant procedures are scattered among several journals and, in general, there is no commercial software that implements such procedures (the authors of these procedures usually used ad hoc routines). Furthermore, all the procedures that we reviewed were concerned only with partial invariance. For these reasons, we thought that applied researchers might find useful an unrestricted FA-based general program that would allow them to assess the different forms of invariance (partial, strong, and strict) and to incorporate a variety of inferential procedures that are not available in commercial programs.

Procedures Implemented in IMINCE

IMINCE (Item Measurement INvarianCE) is a program written in Visual C 6.0 and designed to analyze measurement invariance in two populations. Although the program is particularly suitable for analyses of (either binary or Likert) item scores, it can also analyze sums of item scores (parcels) and sets of test scores. In addition, IMINCE is a general-purpose program that can be used with any two-group comparison using a Cattell/Cliff-type Procrustes rotation for analysis of whole scales. Specifically, the following forms of invariance can be assessed by IMINCE.

Invariance of difficulties. The program tests the general hypothesis that the vector of variable means is the same in the two populations to be compared. This is done using Hotelling's T-square and the corresponding *F* ratio. IMINCE also tests the mean differences, variable by variable, using the univariate *t* test. Because the comparisons usually involve large samples, the sizes of the univariate effect (Cohen's *d′*) are also reported.

Invariance of discriminations (partial invariance). The discrimination indices (factor loadings) are computed from the covariance (or correlation) matrix using three optional methods: principal component analysis, unweighted least squares factor analysis, and unrestricted maximum likelihood factor analysis. When the model considers more than one factor or component, the solution is rotated to show simple structure using normalized varimax (Kaiser, 1958) to help the substantive interpretation of the factor solution, and Procrustes (Cliff, 1966) to allow congruence among samples. To test invariance of discrimination

Table 2
Overall Fit Congruence and Discrepancy Indices per Item

	Congruence Values		Discrepancy Values	
Item Number	Observed	Critical Value at $\alpha = .05$	Observed	Critical Value at $\alpha = .05$
1	.840*	.872	.049	.055
2	.607	.582	.070	.092
3	.993	.979	.005	.026
4	.991	.987	.012	.025
5	.974	.959	.036	.058
6	.992	.989	.014	.020
7	.998	.995	.017	.035
8	.992	.986	.035*	.027
9	.994	.987	.009	.031
10	.999	.957	.018	.060

*Significant difference.

indices, three kinds of tests are implemented in IMINCE: factor congruence, factor discrepancy, and approximate confidence intervals for factor loadings. To estimate the discrimination indices in categorical data, the program allows the so-called *heuristic approach*. This approach consists of (1) computing the matrix of polychoric correlations between categorical items (tetrachoric correlations in the binary case) and (2) analyzing this matrix by unweighted least squares factor analysis. This approach is simple, deals with large numbers of items, and yields results similar to those of the more theoretically correct approach.

Chan, Ho, Leung, Chan, and Yung (1999) proposed a Bootstrap method to evaluate factor invariance in terms of congruence of variables, factors, and the overall loading matrix. The method consists of five steps: (1) One sample is taken as the target and one as the replication; (2) the factor solution from the replication sample is rotated against the target using orthogonal Procrustes rotation (Cliff, 1966); (3) empirical congruence indices between samples are calculated; (4) critical values at α are obtained by Bootstrap resampling; and (5) the observed congruence indices are compared to the critical values at α and are considered as statistically nonsignificant if they are larger than the critical value.

Discrepancy of variables, factors, and overall loading matrices are evaluated using a similar method. However, the index is based on least squares measures of fit. In our program, we generalized the overall index proposed by Raykov and Little (1999), so that it is also used for the variables and the factors (as Chan et al., 1999, did for the congruence index). The discrepancy indices are compared with the critical values at α and are considered statistically nonsignificant if they are smaller than the critical value.

At the variable level, IMINCE also computes approximate confidence intervals for factor loadings. These are bias-corrected percentile intervals obtained from a Bootstrap resampling process (for details, see Lambert, Wildt, & Durand, 1991). Nonoverlapping confidence intervals suggest that a particular variable, as a measure of a given factor, is not invariant over the two populations of interest.

To compute all the indices, the user must determine the number of Bootstrap replications from the 500–5,000 range and can decide between a 90% and a 95% critical value. It must be noted that 1,000 samples are usually recommended in Bootstrap methods (see, e.g., Efron & Tibshiriani, 1993).

Invariance of residual variances. This form of invariance is assessed variable by variable, using bias-corrected percentile intervals obtained from a Bootstrap resampling process. Bootstrap resamples are also drawn from the 500–5,000 range, and either a 90% or a 95% approximate confidence interval is computed. Nonoverlapping intervals suggest that the residual variances of a particular variable are not invariant over the populations that are compared.

Input and Output

The input and output of IMINCE is illustrated using an empirical example, a 10-item Spanish anxiety questionnaire developed by Aguilar and Ferrando (1991) that uses a five-point Likert format. The questionnaire was administered to a sample of 707 women and a sample of 335 men. We aimed to assess item-measurement invariance in the corresponding populations. A model of two factors was expected, and the larger sample was taken as the target sample.

Table 3
Bias-Corrected Percentile Intervals of Residual Variances per Item

Item Number	Target Sample	Replication Sample
1	0.621–0.789	0.764–1.085
2	0.943–1.179	0.948–1.246
3	0.478–0.624	0.476–0.679
4	0.395–0.518	0.462–0.687
5	0.768–0.994	0.693–1.113
6	0.379–0.496	0.360–0.554
7	0.087–0.695	0.175–0.680
8	0.460–0.707	0.376–0.680
9	0.458–0.911	0.467–0.976
10	0.863–1.156	0.740–1.070

The input consists of two ASCII format files containing the participants' scores, the number of participants in each sample, and the number of factors expected in the population. IMINCE default configuration consists of an unweighted least squares factor analysis of the covariance matrices, 1,000 Bootstrap samples, and 95% approximate confidence intervals. We used principal component analysis of the covariance matrices and 5,000 Bootstrap samples.

The output consists of (1) item difficulties, item discriminations, and item residual variances for each sample, and (2) the overall, factor, and item-fit indices described above. Even if the default configuration defines a detailed output, the user can configure the statistics and indices to be reported—that is, those stored in the ASCII format file OUPUT.TXT. The main results are shown in Tables 1, 2, and 3.

Invariance of item difficulties. Hotelling's T-square and univariate t tests suggest significant differences (see Table 1). However, Cohen's d' statistic, which is perhaps more appropriate because the comparisons involved large samples, suggests that there are no substantial differences between the populations.

Invariance of item discriminations. The approximate confidence intervals for factor loadings show overlapping for all the loadings between the populations. However, at the item level there are significant differences in the congruence coefficient of Item 1 and in the discrepancy coefficient of Item 8 (see Table 2). Because of these significant differences at the item level, there are also significant differences in the overall congruence and discrepancy indices.

Invariance of residual variances. The overlapping intervals of all items suggest that the residual variances of items are invariant over the populations compared (see Table 3).

In a second analysis, we omitted Items 1 and 8. Without these two items, IMINCE reported perfect invariance

of item difficulties, discriminations, and variances. The conclusion of our study was that there was strict factor invariance for Items 2, 3, 4, 5, 6, 7, 9, and 10, and no factor invariance for Items 1 and 8.

Limitations

We implemented IMINCE for a PC computer using the WINDOWS 95/98/NT operative system. The program uses all the extended RAM memory available in the computer, and the matrices are defined during the execution of the program. This means that there is no clear limit to the maximum number of items that can be analyzed—it depends on the characteristics of the computer that carries out the analyses. The main limitation of IMINCE is the time needed for computing, especially when a large number of Bootstrap samples are defined. The example in this article, which in fact involved large samples, was performed on a Pentium III computer at 866 MHz and 64 MB RAM. For 5,000 Bootstrap samples, IMINCE needed 6 min and 15 sec. However, the corresponding time was 39 sec for 500 Bootstrap samples. When polychoric correlation matrices are computed with the standard computers available, the analysis can take a very long time. For 5,000 Bootstrap samples and polychoric correlations, IMINCE needed 2 h and 23 min. In the not-too-distant future, most computers will be able to deal easily with this type of analysis.

Program Availability

A copy of the software, a demo, and a short manual can be obtained at no charge at uls@fcep.urv.es.

REFERENCES

Aguilar, A., & Ferrando, P. J. (1991). *CAR: Cuestionario de ansiedad y rendimiento académico* [*AAPQ: Anxiety and Academic Performance Questionnaire*]. Unpublished manuscript.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for educational and psychological testing* (Part II). Washington, DC: American Educational Research Association.

Buchanan, T., & Smith, J. L. (1999). Using Internet for psychological research: Personality testing on the World-Wide Web. *British Journal of Psychology*, **90**, 125-144.

Chan, W., Ho, R. M., Leung, K., Chan, D. K.-S., & Yung, Y.-F. (1999). An alternative method for evaluating congruence coefficients with Procrustes rotation: A Bootstrap procedure. *Psychological Methods*, **4**, 378-402.

Church, A. T., & Burke, P. J. (1994). Exploratory and confirmatory tests of the Big Five and Tellegen's three- and four-dimensional models. *Journal of Personality & Social Psychology*, **66**, 93-114.

Cliff, N. (1966). Orthogonal rotation to congruence. *Psychometrika*, **31**, 33-42.

Drasgow, F. (1984). Scrutinizing psychological tests: Measurement equivalence and equivalent relations with external variables are the central issues. *Psychological Bulletin*, **95**, 134-135.

Efron, B., & Tibshiriani, R. J. (1993). *An introduction to the bootstrap.* New York: Chapman & Hall.

Joinson, A. (1999). Social desirability, anonymity, and Internet-based questionnaires. *Behavior Research Methods, Instruments, & Computers*, **31**, 433-438.

Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika*, **36**, 409-426.

Kaiser, H. F. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, **23**, 187-200.

Lambert, Z. V., Wildt, A. R., & Durand, R. M. (1991). Approximating confidence intervals for factor loadings. *Multivariate Behavioral Research*, **26**, 421-434.

McCrae, R. R., Zonderman, A. B., Costa, P. T., Bond, M. H., & Pauonen, S. V. (1996). Evaluating replicability of factors in the revised NEO personality inventory: Confirmatory factor analysis versus Procrustes rotation. *Journal of Personality & Social Psychology*, **70**, 552-566.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, **58**, 525-543.

Pasveer, K. A., & Ellard, J. H. (1998). The making of a personality inventory: Help from the WWW. *Behavior Research Methods, Instruments, & Computers*, **30**, 309-313.

Raykov, T., & Little, T. D. (1999). A note on Procrustes rotation in exploratory factor analysis: A computer intensive approach to goodness-of-fit evaluation. *Educational & Psychological Measurement*, **59**, 47-57.

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin*, **114**, 552-566.

(Manuscript received November 15, 2001;
revision accepted for publication September 15, 2002.)