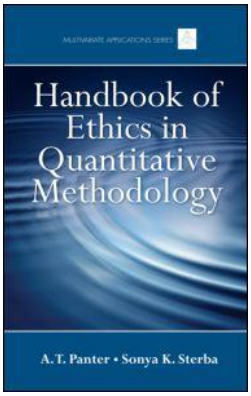

Handbook of Ethics in Quantitative Methodology

A.T. Panter, Sonya K. Sterba

Some Ethical Issues in Factor Analysis

Publication details
https://www.routledgehandbooks.com/doi/10.4324/9780203840023.ch12
John J. McArdle
Published online on: 20 Jan 2011

12

Some Ethical Issues in Factor Analysis

John J. McArdle
University of Southern California

Methodological Issues

This is a book about ethics in data analysis. So we might begin by asking, "Why worry about ethics in data analysis? Isn't this already taken care of by good science training?" The answer of course is, "Yes." Very early in our careers, as early as in elementary school, we are taught to follow and respect the so-called "scientific method" as a guide to obtaining such useful and replicable results. We can all agree that sturdy scientific results require sturdy scientific principles. A key reason we worry about this topic is because we have to trust one another in the creation of sturdy scientific results. But because we are all so trustworthy, what could possibly be the problem?

Unfortunately, we are also well aware of publicized violations of this trust: We know we should not simply graft cancer-free tails onto otherwise sickly rats, and we should not claim to have created a device that creates useful energy from nothing, and we know we should not publish algebraic proofs developed by others as if they were our own invention. We usually assume these are charades posing as good science and believe that we would never knowingly create such problems ourselves. Never! But then we run our key hypotheses about group differences using a one-way *analysis of variance* (ANOVA) and find probability values that are just larger than the arbitrary $p < .05$. We consider using multiple t tests instead of the one-way ANOVA, or using one-tailed tests, but our early statistics training makes us shudder at this obvious violation of statistical laws (see Scheffe, 1959). So we start to say this result is "approaching significance," essentially creating our own new level of probability that is without bounds. Or we eliminate some offending data (i.e., possibly true outliers), or we try a transformation of the dependent variable (DV), and rerun the ANOVA to see whether we can achieve the seemingly magic numbers required for publication.

The ethical basis of this scenario is really no different than when in a more complex modeling analysis, we see that a model fitted using our a priori logic does not seem to fit by acceptable standards (i.e., root mean square error of approximation, $e_a < .05$; see Browne & Cudeck, 1993). In this case, we try hard to find another model that is very close to our original model and does meet the arbitrary standards of good fit (see Brown, 2006). Because the second one will serve our purposes, we report it and, unfortunately, we use standard statistical tests and treat the model as though it were our starting point. In the thrill of a publishing moment, we may completely forget about our starting point model—and our ethical virtues.

As in many areas of life, the reason we cross these ethical boundaries in data analysis is because we desire novel and reasonable results, and we often simply blame the rules of publication for having absurd criteria. We have also become aware that the search for connections and dynamic influences and causes is fairly complex and not easy to describe, so we conclude that our little deception will do no real harm in the long run. We come to realize a good description of what we are doing is much more like we are making a "principled argument" (Abelson, 1995). So we are encouraged to stretch our ethical boundaries, and, unfortunately, they become less clear.

But almost any self-evaluation will lead us to be rightfully concerned that we may be carrying out science without a firm ethical compass. Since the turn of the 20th century, there has been a collective effort to develop helpful ethical principles in all kinds of empirical research studies. Ethical principles were important in the development of the cooperation between the scientist, producer of results, and the consumers of results—and we hope the newspaper reporters do not criticize the scientists. At the same time, the need for accurate, replicable, and reliable information flow is needed for the accumulation of studies and "replicable results" in the "soft-fact" sciences.

In this chapter we will highlight some ethical dilemmas of one widely used technique—*factor analysis* (FA; see McDonald, 1985; Mulaik, 2009). The history of psychological statistics shows a great respect for a priori testing of formal hypothesis (Fisher, 1925) and has led to many organized and successful research programs. Unfortunately, this also led to skepticism and disdain for exploratory data analysis procedures (e.g., Tukey, 1962, 1977), although not all of this criticism is warranted. The previous divisions between confirmation and exploration are apparent in FA as well, but some confusion has led researchers to state that they used confirmatory methods when, in fact, their work was largely exploratory in nature.

To resolve these problems, we try to show how a *structural factor analysis* (SFA; Albert, Blacker, Moss, Tanzi, & McArdle, 2007; Bowles, Grimm, &

McArdle, 2005; Cattell, 1966; McArdle, 1996; McArdle & Cattell, 1994) approach to FA allows us to use the full continuum and avoid the artificial semantic differences of confirmation and exploration. This approach to FA relies on both "confirmation" and "exploration" and is consistent with a "functionalist" view of psychological research (as in McArdle, 1994a, 1994b; McArdle & Lehman, 1992). Further definitions are listed in Table 12.1, and we return to this table at various points in this discussion.

In this chapter, some technical issues of SFA are presented first, but not in great detail, and these are quickly followed by a case study example using real cognitive data. This leads to a discussion of what others have done about ethical problems in FA, as well as five suggestions for future work. My hope is that this approach will lead us to think that ethical principles can always be followed in data analysis. This also leads us to see that the main ethical problem we face in SFA is what and how much should we tell others about what we have done. The ethical answer is clear—we should document all our SFA work and tell others whatever we can. In practice, this is not so easy.

Statistical Background

The Statistical Basis of Factor Analysis

The statistical and psychometric history of FA is long and contains many specialized techniques and colorful concepts (see McDonald, 1985, 1999; Mulaik, 2009). Most of the older techniques will not be used here, and we will only discuss techniques based on the contemporary principles of *maximum likelihood estimation* (MLE; see Lawley & Maxwell, 1971). This approach allows us to carry out both exploratory factor analysis (EFA) and confirmatory factor analysis (CFA) using *structural equation modeling*

TABLE 12.1

A Continuum of Factor Analysis Techniques

Confirmatory	.....	Exploratory
More theory	.....	Less theory
More restrictions	.....	Few restrictions
Overidentified	.....	Exactly identified
More *df*s	.....	Less *df*s
Less absolute fit	.....	Greater absolute fit
Ample stat tests	.....	Few stat tests
Seemingly strong	.....	Seemingly weak

(SEM) computer algorithms (e.g., AMOS, LISREL, semR, OpenMx, M+). This approach used here also allows us to distinguish the most important features of EFA–CFA differences—how many parameter "restrictions" are placed on the data on an a priori basis. Repeatedly, we note that the a priori nature of this selection is still needed for appropriate statistical tests based on the *chi-square* (χ^2) distribution. A *degree of freedom* (*df*) is a model expectation that can be incorrect (i.e., a way to go wrong), so the number of *df*s is used in many indices of model parsimony (to be described).

The techniques of *exploratory factor analysis* are used in most classical FA. A set of unobserved common factors is thought to be responsible for the observed correlations among observed variables (V). In the EFA approach, we propose a specific number of common factors k, possibly with a specific hypothesis, but we almost always explore several models, from no common factors ($k = 0$) to as many common factors as possible ($k = v/2$). In contemporary terms, the number of common factors is predetermined, but the specific set of common factor regression coefficients, termed *factor loadings*, is "exactly identified." This means the factors can be "rotated" to a simpler, possibly more meaningful solution, with no change of common variance and no change in overall fit. A statistical test for the number of common factors in EFA is typically conducted as a sequence of nested chi-square tests, formally based on the size of the residual correlations, and various approaches and indices have been suggested to determine an adequate number of factors (Browne & Cudeck, 1993; Cattell, 1978; Lawley & Maxwell, 1971; McDonald, 1985).

The term *confirmatory factor analysis* was popularized by Jöreskog (1966, 1969, 1977) and used by Tucker and Lewis (1973) to describe the new SEM-based approach to FA. Here we follow their lead and fit "overidentified models" with specific restrictions on the factor loadings. It turned out that classical test statistics (e.g., chi-square) could be applied to this kind of a problem, so CFA fit in very well with ANOVA and other pure forms of statistical inquiry. As a result, many data analysts started to search for clear and a priori factor patterns. Unfortunately, the required level of precision seemed to be lacking. In response, many CFA researchers "trimmed" their data sets and/or parameters or relied on exploratory "modification indices" and "correlated errors," so models appeared to be CFA and benefited from the statistical tests (e.g., Brown, 2006). For similar reasons, Cattell (1978) suggested that we substitute the term proofing FA for MLE–CFA, although this insightful terminology never became popular.

Initial Structural Factor Analysis Models

For the purposes of this discussion, let us assume we have measured six different variables ($v = 6$) on a number of different individuals ($N > 10*v$). When we apply the techniques of FA to this kind of data, we are trying

to understand the best way to represent the observed variables in terms of unobserved factors. A series of alternative models is presented in the path diagrams of Figures 12.1 and 12.2. In these diagrams, the observed variables are drawn in squares, and the unobserved variables are drawn as circles. One-headed arrows represent a direction influence, typically

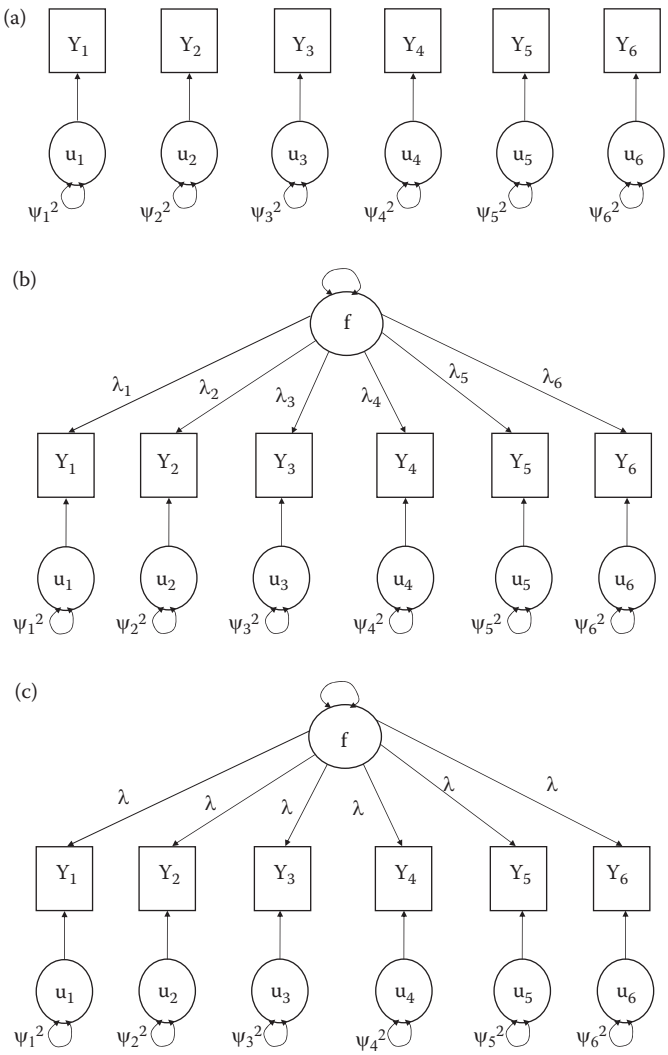

FIGURE 12.1
Alternative common factor models. (a) Six variables—zero common factors but six unique factors ($df = 15$). (b) Spearman-type (1904) one common factor model ($df = 9$). (c) Rasch-type (1961) one common factor model ($df = 14$).

termed *factor loadings*, and two-headed arrows represent nondirectional influences, such as "variance" or "covariance" terms.

The zero-factor model is almost always useful as a starting point or baseline model, and this is presented as a path diagram in Figure 12.1a. Here we assume each of the observed variables is composed of the influence of only one unique factor, labeled u_v, with fixed loadings (unlabeled) but free unique variances, labeled Y_v^2. In this model, the unique latent scores

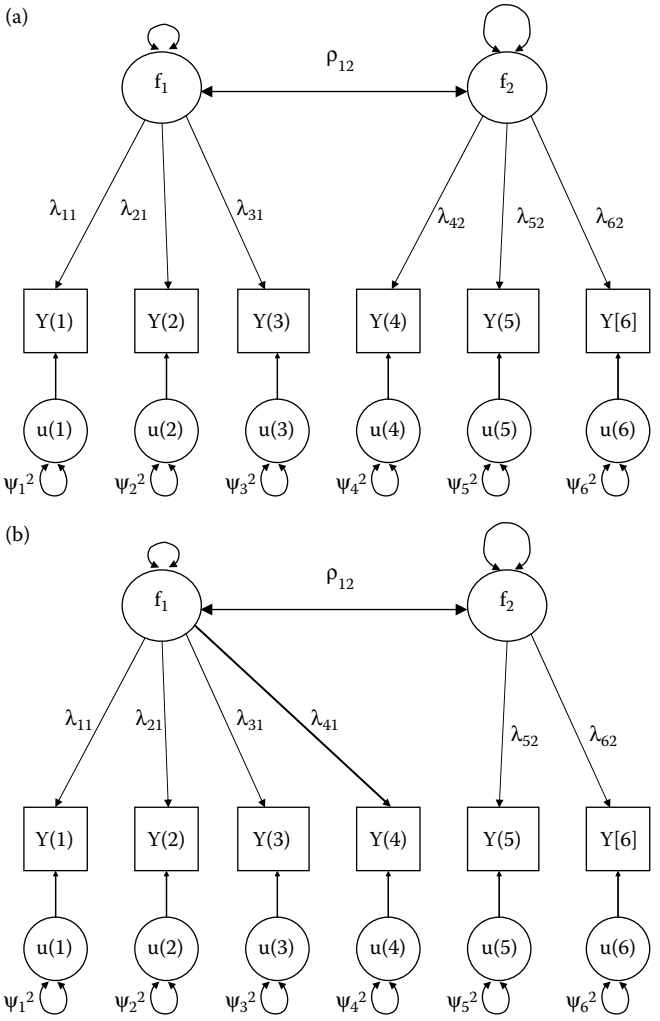

FIGURE 12.2
Alternative two common factor models. (a) "Simple structure" two common factor model (*df* = 8). (b) "Non-nested" two common factor model (*df* = 8). (c) "Exactly identified" two common factor model with oblique constraints (*df* = 4).

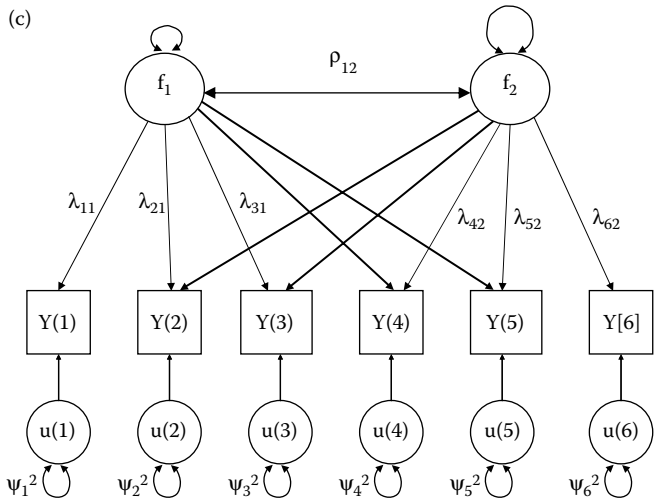

FIGURE 12.2 (Continued)
Alternative two common factor models. (a) "Simple structure" two common factor model (*df* = 8). (b) "Non-nested" two common factor model (*df* = 8). (c) "Exactly identified" two common factor model with oblique constraints (*df* = 4).

are thought to produce the variation we observe. This model restricts the unique variables to have zero correlations, and each restricted correlation counts as *df* = 15, so this is our simplest, most parsimonious model.

If the model of "no correlation" is true, we typically state that "this zero-factor model fits the observed data," and we have no need to go further with data analysis. However, if there are significant correlations among the observed scores, then this simple model does not completely capture the observed correlations, and we then typically say this simple latent variable model does not fit the data. One common statistical test used here is based on the likelihood ratio test (LRT), formed from the *likelihood of the* original data matrix compared with the likelihood of the model estimated matrix (i.e., a diagonal). It is often briefly stated that "under certain regularity conditions" such as "the unique factor scores are normally distributed," the LRT is distributed as a chi-square index that can be used to evaluate the model misfit—that is, with low chi-square relative to the *df*s taken as an indication of good fit (for details, see Browne & Cudeck, 1993; Lawley & Maxwell, 1971; McDonald, 1985).

The next theoretical model is based on the one common factor model, and this can be seen in the path diagram of Figure 12.1b. In this diagram, the observed variables are drawn in squares, and the unobserved variables from our theory are drawn as circles. In Figure 12.1b, we assume each observed variable is composed of the influence of both its own unique factor as before but also that there is one common factor (labeled *f*), each

with its own loadings (labeled λ_v). In this model, the two latent scores are thought to produce the variation we observed, but only the latent common factor is thought to produce the covariance of the observed scores by a simple pattern of expectations (i.e., without details, $\sigma_{ij} = \lambda_i \times \lambda_j$). This model also restricts the unique variables to have zero correlation but requires six additional factor loadings to do so (so $df = 9$). The test of this one-factor LRT hypothesis is that when all model expectations are removed, there is no remaining correlation among the observed scores.

Variations on the One-Factor Concept

Although it is clear that the one-factor model is a strong hypothesis, it is rarely used in this way (see Horn & McArdle, 2007). Part of the reason for this hesitation may be because of the typical problems faced by factor analytic researchers. For example, to obtain the precision required by the LRT, we must have an a priori hypothesis of one common factor for a specific set of variables. Of course, it is not uncommon for researchers to drop participants who either do not meet some sampling requirements (i.e., required language, age > 50, Mini-Mental Status Examination > 24) or whose behavior seems aberrant (i.e., outliers). Although these can be reasonable criteria on practical grounds, any nonrandom sample selection may violate the assumptions of the statistical tests. But perhaps more critically, researchers routinely drop some of variables that are "not working," or rescale some of the variables to remove "odd distributions," and almost any nonrandom variable selection has an impact on the statistical tests. It is not that these are always horrific practices, but it is clear that the standard statistical tests are no longer appropriate after these kinds of changes in the data are made (cf. Brown, 2006).

Let us consider one other CFA extension—the Rasch-type model (see Embretson & Reise, 2000; McDonald, 1999; Wilson, 2005). From a one-factor starting point, if we fix all model loadings to be identical (λ), we end up with properties that mimic a Rasch scale—that is, the summation of the scores is parallel to the Rasch-type factor score estimates. This model has a pattern that is even more restrictive ($df = 14$; $\sigma_{ij} = \lambda^2$), so it may not fit the data very well, but it is needed to establish the adequacy of a simple Rasch-type summation scale. From this viewpoint, the Rasch model is a highly restricted test of a formal CFA hypothesis. It makes little difference that this Rasch model is more typically used with items than with scales.

The knowledgeable researcher will notice that no effort is made here to evaluate the utility of what are often termed *correlated errors* using the statistical techniques of "modification indices" (MI; e.g., Brown, 2006). There are several reasons why these parameters and this approach to model

fitting are completely ignored from this point here. The first reason is that this approach allows, and even embraces, the estimation of *correlated specifics* (CS). The problem is that almost any CS approach does not match the basic goals of common FA at all (e.g., see Meredith & Horn, 2001; cf. McArdle & Nesselroade, 1994). That is, the estimation of any CS in FA typically attempts to isolate part of the data that cannot be fitted by a specific model. A second reason is that the MI approach, which attempts to recursively locate the single parameter that, if estimated in the model, can alter the model fit the most, does not account for dependencies that are multivariate in nature. It is not surprising that this combined CS–MI approach simply does not even work well as an exploratory tool (see MacCallum, Roznowski, & Necowitz, 1992). From a traditional perspective, FA modeling based on this CS–MI approach is viewed as a misunderstanding of the analytic goals of FA.

Expanding Structural Factor Analysis Models

Continuing with the example at hand, Figure 12.2 extends these FA concepts a bit further by proposing a less restrictive two-factor hypothesis for the data. In this model, the first three variables are thought to load on the first common factor (f_1), and the last three variables are thought to load on a second factor (f_2). This is a classic example of a "confirmatory factor" model. The model expectations within each set are the same as before ($\sigma_{ij} = \lambda_i \times \lambda_j$), but across sets of variables we now add a parameter ($\sigma_{ij} = \lambda_i \times \rho_{12} \times \lambda_j$), so this model should fit better than the one-factor version. Given all other assumptions, the difference between model of Figures 12.1b and 12.2a is a testable hypothesis (of $\rho_{12} = 1$).

The two-factor model of Figure 12.2b uses the same number of model parameters but places these in a different location so the first factor has four loadings and the second factor has only two loadings. Unfortunately, the number of parameters in each model of Figures 12.2a and 12.2b is the same, so no formal test of the difference is possible. To create such a test, we often create a composite model where both loadings are allowed. Of course, such a model is no longer "simple" in the sense that a variable such as Y_4 can load on both common factors. However, we can form reasonable LRTs to try to determine which model is best for our data.

Following a similar logic, we can create a model where we allow as much room to fit as is possible, and the result is Figure 12.2c, where only two of the variables are used as "reference variables" (Y_1 and Y_6), and the other four are allowed to load on both common factors. Perhaps it is obvious, but all other models here (Figures 12.1a to 12.2b) are formally nested as proper subsets of the model of Figure 12.2c, so all can be fairly compared for fit. Perhaps it is also obvious that our initial choice of the two reference variables was arbitrary. This means that for a model where one or two other

variables (Y_2 and Y_5) are chosen as reference variables, the misfit would be exactly the same but the parameter values would be different. To wit, there is more than one "exactly identified" two-factor solution that can be fit to solve the same problem. This is a simple example of the problem of "factor rotation"—given a specific number of identifiable parameters (i.e., 10 loadings here), there are many positions that yield the same pattern of expectations and hence the same $df = 4$ and the same misfit.

Using this form of SEM, we can be clear about the exact model that is fit to the data—we can easily present all the summary statistics to be fitted (i.e., correlations) and the exact model (i.e., as a path diagram). This allows others to replicate our model analyses with their own data. Under the assumptions that (a) the data were selected in advance of the analysis and (b) this model was chosen on an a priori basis, then SEM yields statistical tests of the (c) overall fit of the model to the data (χ^2, ε_a, etc.) and (d) individual standard errors for each model parameter ($z = MLE_p/SE_p$). These statistical indices can be used to judge the adequacy of the fit using consensus rules of agreement, but we must be careful about comparing them with a priori distributions if they are not a priori tests.

Case Study Example

Cognition Measurement in the Health and Retirement Study

The example presented next comes from our recent work on an FA of cognition measures in the *Health and Retirement Study* (HRS; see Juster & Suzman, 1995). At the start of this analysis, we recognize it is uncommon to ask cognitive questions in large-scale survey research, even though the cognitive status of the respondent is of obvious importance to providing genuine answers (see Schwarz et al., 1999). Indeed, the HRS has a long and reasonable history of using cognitive items for this purpose (see McArdle, Fisher, & Kadlec, 2007).

Following this logic, the specific application presented here uses publicly available data on a small set of cognitive variables ($v = 7$) measured on a large sample of adults (age > 50, $N > 17{,}000$). Table 12.2 is a list of available cognitive variables in the current HRS data. Some incomplete data have been created by the HRS because not all persons were administered all $v = 7$ tests at any sitting, but overall coverage of all variables is reasonable (>80%). Respondent sampling weights will be used here to approximate a sample that is representative of the U.S. population older than age 50 (see Stapleton, 2002). When using the sampling weights, the model must be fitted using alternative estimators (i.e., a variation of MLE allowing weights, termed *MLR*), and the fit can be altered by a constant of kurtosis (*w4*). Although we do report these values here, the appropriate use of weighted

TABLE 12.2

A Listing of the Health and Retirement Study Cognitive Measures

1. Immediate word recall (IR; 10 items)
2. Delayed word recall (DR; 10 items)
3. Serial 7s (S7; to assess working memory)
4. Backward counting (BC; starting with 20 and 86)
5. Dates (DA; today's date and day of the week)
6. Names (NA; object naming, president/vice president names)
7. Incapacity (IN; to complete one or more of the basic tests)

And on some occasions …

8. Vocabulary (VO; adapted from WAIS-R for $T > 95$)
9. Similarities (SI; adapted from WAIS-R for $T = 92, 94$)
10. Newly created "adaptive" measures from the WJ-III

WAIS-R, Wechsler Adult Intelligence Scale–Revised; WJ-III, Woodcock-Johnson III.

chi-square tests is not a key issue of this chapter. Weighted summary statistics about these HRS cognitive variables are presented in Table 12.3.

Considering One Common Factor

The SFA approach used here starts with a sequence of CFAs but ends on a more relaxed set of EFAs. To initiate the CFAs, the models that were first fitted include the zero-factor model (Figure 12.1a) and the one-factor model (Figure 12.1b). The zero-factor model was fitted mainly as a

TABLE 12.3

Health and Retirement Study Summary Statistics From Respondent Interviews ($N = 17{,}351$)

(a) Means and Standard Deviations

	IR[1]	**DR[1]**	**S7[1]**	**BC[1]**	**NA[1]**	**DA[1]**	**VO[1]**
	55.7	43.9	70.5	95.2	94.2	91.3	55.4
	18.5	22.3	34.1	21.1	14.3	16.7	21.2

(b) Correlations

	IR[1]	**DR[1]**	**S7[1]**	**BC[1]**	**NA[1]**	**DA[1]**	**VO[1]**
IR[1]	1.000						
DR[1]	.773	1.000					
S7[1]	.371	.359	1.000				
BC[1]	.189	.170	.227	1.000			
NA[1]	.283	.280	.255	.201	1.000		
DA[1]	.362	.345	.381	.221	.308	1.000	
VO[1]	.385	.352	.393	.185	.202	.403	1.00

Variable abbreviations appear in Table 12.2. Measured at occasion with most cognitive variables; respondent weights used, 36 patterns of incomplete data, coverage >81%; MLE(MAR) using M+; χ^2(diagonal) = 18,521 on $df = 21$; eigenvalues (%) = [42.4, 14.1, 11.9, 11.2, 8.8, 8.2, 6.6, 3.2].

baseline model for comparison, but the second could be defended based on prior cognitive theory going as far back as Spearman (1904; see Horn & McArdle, 1980, 1992, 2007; McArdle, 2007). The goodness of fit of the zero-factor model is very poor (χ^2 = 18,520, df = 21, k = 1.61, ε_a = .225), and the one-factor model seems much better (χ^2 = 2,530, df = 14, $w4$ = 1.58, ε_a = .102). For illustration, the ML parameter estimates of the one-factor model are presented in Figure 12.3a. Of course, the one-factor results seem to suggest the one common factor model is only adequate for the first two variables, and the other four variables are largely unique.

We next add a test of the Rasch model of one factor with equal loadings, and it seems to fit even worse (χ^2 = 7,033, df = 20, $w4$ = 1.76, ε_a = .142). From these initial analyses, we conclude that more than one common factor is likely to be needed to capture all the variation in these cognitive data. The lack of fit of the Rasch model also provides evidence that the HRS cognitive scores should not simply be added together to form an overall score (i.e., see McArdle et al., 2007). It did not matter what variation of the one-factor model is fitted; it is apparent that one factor of the HRS cognitive variables leaves a lot to be desired.

Considering More Than One Common Factor

The models we have just fit are common using these kinds of cognitive data. The second set of models was decidedly CFA in origin. I (person JJM) asked a more knowledgeable colleague (Dr. John L. Horn, University of Southern California, person JLH) to create a two-factor hypothesis from these data. The conversation follows (from audio tape, 08/24/2002):

JJM: Can you please take a look at this new HRS data set I am now using?

JLH: OK, but I think this is a seriously impoverished data set for any cognitive research.

JJM: Yes, but the sample is very large and representative, over 17,000 people, and the HRS is now using a one-factor model.

JLH: OK, I will show you how bad this is—how about we do an exploratory factor analysis first?

JJM: We could, but that would distort the a priori basis of the chi-square and other statistical tests.

JLH: I agree, but who actually uses those tests anyway? Do I need to remind you that factor analysis is not a statistical problem anyway?

JJM: Are you saying you just can't do it? After 40 years of cognitive research, you don't have any formal a priori hypotheses at all?

JLH: No, I didn't mean that. I can do it. I suggest what you have here is a little factor of short-term acquisition retrieval, and I do mean little, and probably a second common factor based on the rest of them, whatever they are supposed to be.

The results from fitting this kind of what might be said to be a *semi-formal* a priori CFA two-factor model are presented in Figure 12.3b. The fit of the model shows much improvement ($\chi^2 = 222$, $df = 13$, $w3 = 1.50$, $\varepsilon_a = .030$), and a formal test of whether the interfactor correlation is unity ($\rho_{12} = 1$) is indexed by the LRT difference ($\chi^2 = 2{,}308$, $df = 1$). This initially reminds us that when we have $N > 17{,}000$ people we have great power

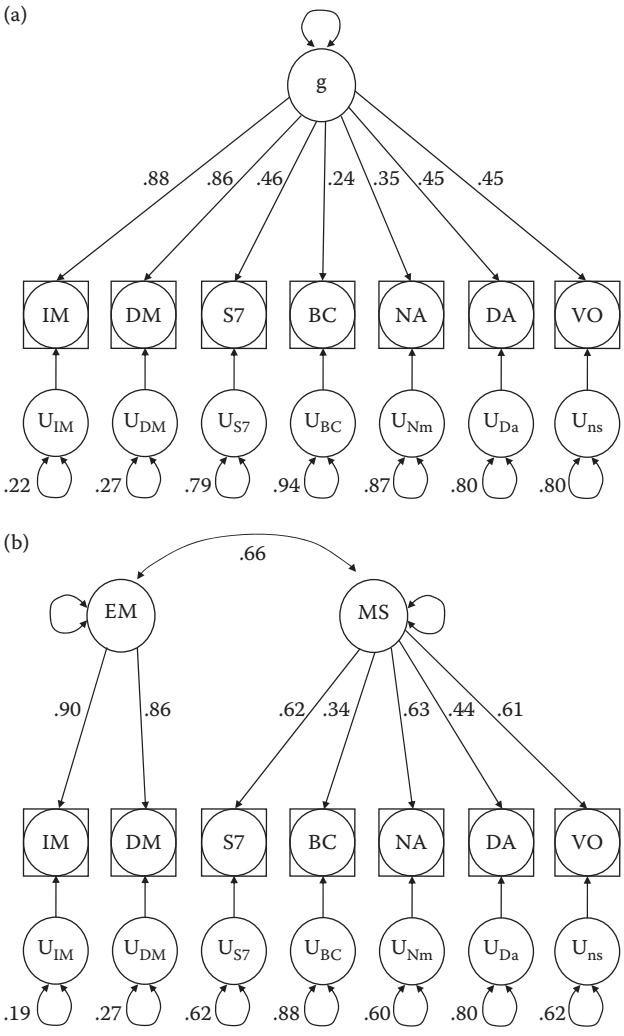

FIGURE 12.3
Alternative factor models for the seven HRS cognitive abilities ($N > 17{,}000$). (a) One-factor model results ($\chi^2 = 2{,}530$, $df = 14$, $k = 1.58$, $\varepsilon_a = .102$; standardized MLE listed). (b) Two-factor CFA model results ($\chi^2 = 222$, $df = 13$, $w4 = 1.50$, $\varepsilon_a = .030$). (c) Three-factor CFA model results ($\chi^2 = 214$, $df = 12$, $k = 1.50$, $\varepsilon_a = .030$).

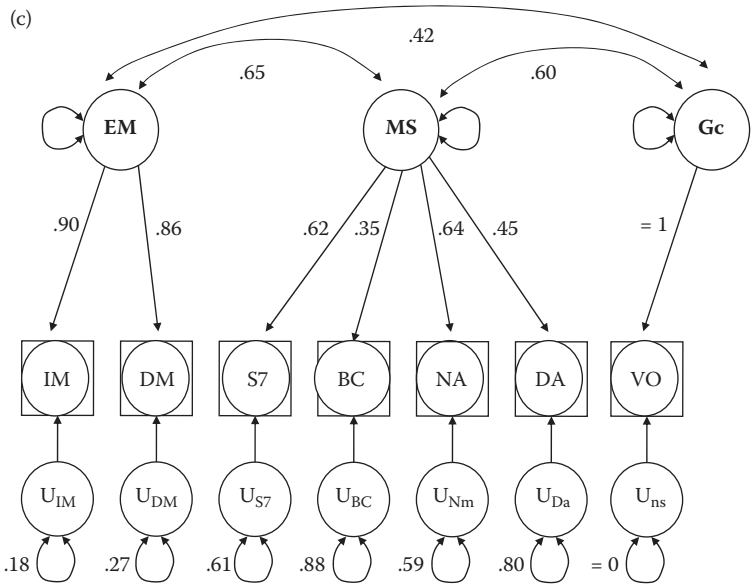

FIGURE 12.3 (Continued)
Alternative factor models for the seven HRS cognitive abilities ($N > 17{,}000$). (a) One-factor model results ($\chi^2 = 2{,}530$, $df = 14$, $k = 1.58$, $\varepsilon_a = .102$; standardized MLE listed). (b) Two-factor CFA model results ($\chi^2 = 222$, $df = 13$, $w4 = 1.50$, $\varepsilon_a = .030$). (c) Three-factor CFA model results ($\chi^2 = 214$, $df = 12$, $w4 = 1.50$, $\varepsilon_a = .030$).

to state that $\rho_{12} = 0.66$ is statistically different than $\rho_{12} = 1$. But the other parameter estimates are more revealing. The isolation of the first two variables load onto a first factor we have labeled *episodic memory* (EM). The second factor has highest loadings for S7 and VO, so it may be a general crystallized (Gc) intelligence factor, but because of all the relatively low level of information required (NA, DA, and BC), we have labeled this as *mental status* (MS). Incidentally, a Rasch version of this two-factor hypothesis does not fit the data very well ($\chi^2 = 1{,}962$, $df = 18$, $w4 = 1.62$, $\varepsilon_a = .079$) and does not seem to fit this model very well. Of course, the model fit could probably be improved further by considering the categorical nature of these three variables (i.e., most people get them all correct). However, it is very clear that the model fits well, and the hypothesis of JLH was clearly confirmed. But this seeming success made us go even further (from audio tape, 08/24/2002):

JJM: So is this enough for now? Are we done? Can we fit it any better?

JLH: Yes. It seems to me that the SAR factor based on the first two variables is reasonable; the next four are simply the mental status of the person, and likely to go together. But the vocabulary is

really a better indicator of crystallized intelligence. Too bad, but the lack of other measures makes vocabulary collapse into the second factor. Can we isolate this one variable in any way?

JJM: Maybe. I will try.

JLH: Incidentally, I think the real problem you have here is that there are no measures of fluid intelligence at all.

The results from fitting this semiformal a priori CFA three-factor model are presented in Figure 12.3c. The fit of the model shows much improvement ($\chi^2 = 214$, $df = 12$, $w3 = 1.50$, $\varepsilon_a = .030$), and a formal test of whether the VO is isolated is indexed by the LRT difference ($\chi^2 = 6$, $df = 1$). Note that no estimate of the uniqueness of VO is estimated because this variable is isolated. This mainly reminds us that there is not much difference between the model of Figure 12.3b and 12.3c in this context. From the parameter estimates, we can see the isolation of the first two factors labeled EM and MS, whereas the third is labeled Gc. Another way to achieve a similar goal was to drop the VO from the data set completely and refit the two-factor model. When this was done, the model fit was excellent ($\chi^2 = 76$, $df = 8$, $w4 = 1.50$, $\varepsilon_a = .022$). Nevertheless, we know that model fitting itself does not seem to be a good way to isolate the Gc factor—a far better way would be to add variables that are indicative of the broader Gc concept (i.e., knowledge tests) and then to test this isolation with these multiple outcomes.

An Exploratory Factor Analysis

To see what happens when an exploratory approach is taken, the same matrices were input into an EFA algorithm, where a succession of common factors are extracted, and where multiple factors were defined by factor rotation procedures (Browne, 2001). The EFA results presented in Table 12.4 include the misfit indices, including the error of approximation and its confidence interval (see Browne & Cudeck, 1993). The results listed here clearly show the progression from zero to three common factors improves the fit at every step—one factor is far better than zero; two factors seem far better than one; and three factors seem even better than two. The index of misfit that first achieves one of the standard criteria of "good fit" (where $\varepsilon_a < .05$) is the two-factor model.

The two-factor model fitted as an EFA does not explicitly state where the salient loadings are located. To understand this model, we need to apply some techniques of factor rotation (see Browne, 2001). Of course, this is not a standard solution, so we may not be interested in "simple structure"–based rotations. One useful possibility here was defined by Yates (1987) in terms of minimizing the geometric mean of the squared loadings—the so-called Geomin criterion. Additional research on this Geomin criterion

TABLE 12.4

Results for a Consecutive Sequence of Four Exactly Identified Factor Models

Statistic	$k = 0$	$k = 1$	$k = 2$	$k = 3$
χ^2	18,520	2,530	188	24
df	21	14	8	3
$\Delta\chi^2$	—	15,990	2,162	414
Δdf	—	7	6	5
ε_a	.225	.102	.036	.020
$-95\%(\varepsilon_a)$	.223	.098	.032	.013
$+95\%(\varepsilon_a)$	.228	.105	.041	.028
w4	1.61	1.58	1.37	1.02

has added standard errors for the rotated loadings (Jennrich, 2007). When we carry out these calculations we obtain the results listed in Table 12.5—the first factor is indicated by the IR and DR and can be termed EM, and the second factor is indicated by the last five variables and can be labeled MS. In other words, the EFA gave nearly identical results to our previous CFA of the model of Figure 12.3b. However, this two-factor CFA was not the only possible EFA rotation of the two factors, and this EFA result shows a remarkable consistency with the CFA model of Figure 12.3b. In this way, this EFA approach gives more credibility to the CFA model of Figure 12.3b.

Beyond the Initial Structural Factor Model

One of the main reasons we want to isolate a reasonable common factor structure is that we can use this model in further forms of data analyses, such as in models of external validity (McArdle & Prescott, 1992). Two examples from our work on HRS cognition measures are presented here.

TABLE 12.5

Results for the Two Common Factor Model

Measure	**Factor λ_1**	**Factor λ_2**	**Unique ψ^2**
IR[1]	**.83**	.07	.24
DR[1]	**.90**	**−.01**	.21
S7[1]	.10	**.60**	.63
BC[1]	−.06	**.39**	.87
NA[1]	−.06	**.69**	.57
DA[1]	.05	**.40**	.81
VO[1]	−.02	**.60**	.63

Maximum likelihood estimation (MLE) with Geomin; $\rho = .65$, $\varepsilon_a = .036$; parameters with MLE/SE = $t > 4$ are listed in bold.

In the *latent variable path analysis* (LVP) approach, we can bring additional variables into the same SEM (McArdle & Prescott, 1992). One benefit of this approach is that we can evaluate the regression model with variables that are purified from measurement error. For example, the LVP of Figure 12.4 (for a full description, see McArdle et al., 2007) shows the three-factor CFA, where the three latent variables of EM, MS, and VO are predicted from only six demographic variables (age, education, gender, cohort, dyad status, and mode of testing). For example, the results show strong negative effects of age on EM (–0.7), and this is a larger effect than the impact of age on any observed variable. The independent impacts of education are positive on all factors (+0.5, +0.6, +0.5). The effects of gender are seen only on EM (females greater by +0.5). The independent effects of cohorts are negative on EM and VO, even though the scores are increasing over successive cohorts (i.e., possibly education effects are responsible). Being in a dyad is somewhat positive, and the mode of testing (telephone or face to face) makes only a little difference in latent test scores.

Another kind of SEM analysis that is now possible is based on longitudinal SEM (see McArdle, 2007, 2009). The longitudinal nature of the HRS data collection is very practical—at the initial testing all persons are measured in a face-to-face setting, but at the second testing about 2 years later, the same people are interviewed over the telephone. Presumably, because the same cognitive questions are asked, the tests used measure the same constructs. Figure 12.5 is a display of this concept about measurement of the same latent variables over time. It is now fairly well known that the

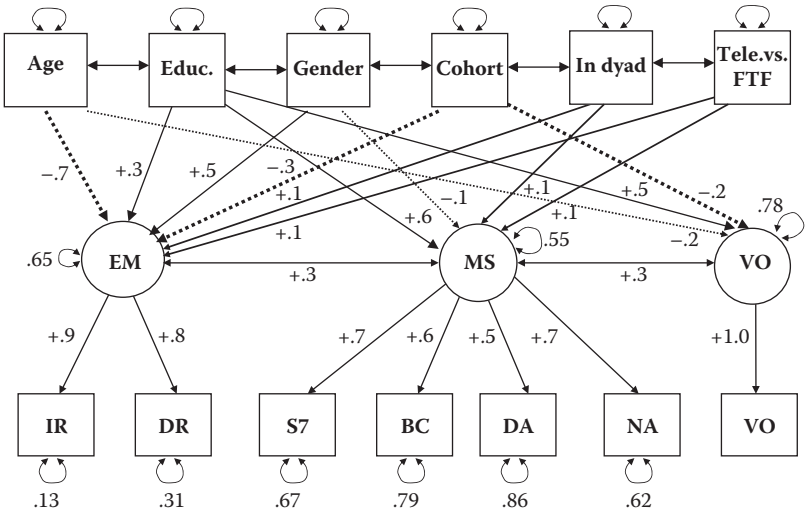

FIGURE 12.4
Three common factors related to other HRS demographic indices ($N > 17{,}000$).

general idea of measurement invariance is a testable SEM hypothesis—we force the factor loadings to be identical (or invariant) at both occasions so we can evaluate the loss of fit. If such a model with this kind of "metric invariance" can be said to fit, then we can easily examine other features about the latent variables—means, deviations, cross-regressions, etc. In fact, the need for some form of measurement invariance is so compelling it is hard not to make this the object of the analysis—that is, why not simply use these SEM techniques to isolate the measured variables that seem to have this useful LV property (McArdle, 2007, 2009). This approach, of course, uses CFA software to carry out an EFA analysis (also see Albert et al., 2007; Bowles et al., 2005; McArdle & Cattell, 1994).

To pursue these longitudinal analyses, a new data set based on cognitive variables was constructed from the available archives of the HRS consisting of the first face-to-face (FTF) and first telephone (TEL) testing. To retain the large and representative sample size ($N > 17{,}000$), the VO variable was no longer considered (i.e., it was not measured twice in most cases). The analytic results for the remaining ($v = 6$) variables are presented in Table 12.6. The first three rows (6a) are a list of the model fits for the one-factor model, first as metrically invariant over time, then as configurally invariant (i.e., same nonzero loadings, but not exact values), and then with one-to-one specific longitudinal covariances (as in McArdle & Nesselroade, 1994). The first model does not fit well; the second fits better; and the third is best so far.

The second set of rows presents the fit of the same three models using a two-factor CFA (much like Figure 12.2b), and the fits are uniformly better. The first model is much better; the second is not much different; and the third model, with metric invariance and longitudinal specific factor covariances, is nearly perfect ($\chi^2 = 423$, $df = 57$, $\varepsilon_a = .023$). Thus,

TABLE 12.6

Fit Indices for One and Two Common Factors Based on Six Measures at Two Longitudinal Occasions

6a: $k = 1$ Models	χ^2	df	$\Delta\chi^2/\Delta df$	ε_a
Invariant Λ, Ψ^2	8,600	69	—	.087
Configural Λ	8,579	64	21/5	.090
MI + specifics covariance	4,534	63	4,056/6	.066
6b: $k = 2$SS Models	**χ^2**	***df***	**$\Delta\chi^2/\Delta df$**	**ε_a**
Invariant Λ, Ψ^2	2,579	63	—	.050
Configural Λ	2,578	59	1/4	.051
MI + specifics covariance	423	57	2,156/6	.023

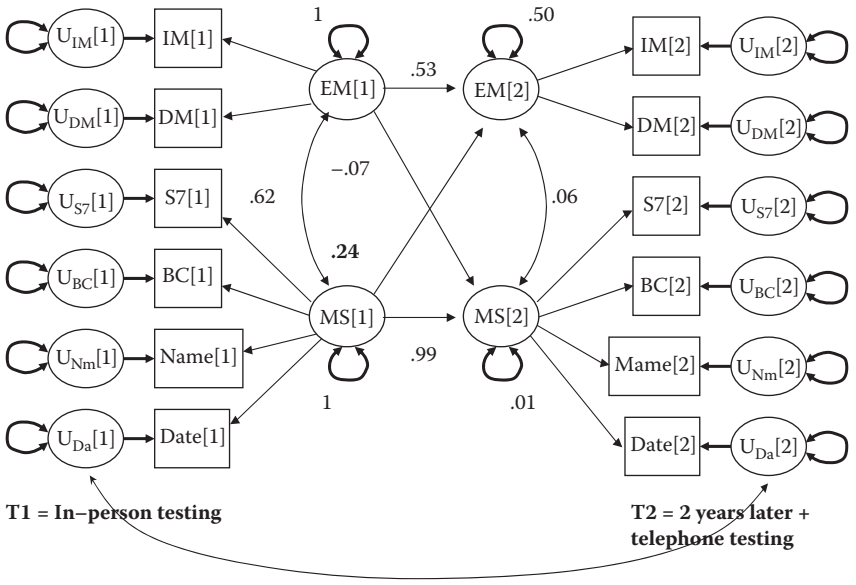

FIGURE 12.5
The HRS cognitive measures with factorial invariance over time and mode of testing ($N > 17{,}000$).

whereas we were unsure about one factor, these results suggest the two factors, EM and MS, can be measured using the same six tests in either FTF or TEL modalities without any measurement biases. The results for the latent variable cross-lagged regressions are given in Figure 12.5, and these suggest that the MS[t] is highly stable and most predictive of EM[t + 1]. More analytic work is now being done on these dynamic relationships.

Prior Work

Let us return to the ethical issues in FA. Ethical issues about the practices in FA have been raised by many others, and the same messages are found in the history of other statistical procedures, such as ANOVA and item response theory. For example, clear recognition of these issues can be found in the classic debates about the "significance of the significance test" (e.g., Harlow, Mulaik, & Steiger, 1997; Lecoutre, Lecoutre, & Poitevineau, 2001). In one compelling resolution, Cattell (1966) rejected the use of the principles of experiment-wise error and suggested the use of what he termed the

inductive-hypothetico-deductive spiral (also see Tatsuoka & Tiedeman, 1954). Basically Cattell, among others, was suggesting we consider a continuum with CFA at one end point and EFA at the other (see Table 12.1).

In using CFA, we assume there is more theory, more restrictions (hence more *df*s), overidentified parameter estimates, and ample statistical tests with corresponding good fits. For these reasons, the topographic presentations of CFA seem very strong and useful in research where there has been a lot of reliable work. On the other hand, the EFA end of the continuum is based on less theory, fewer restrictions (lower *df*s), exactly identified parameter estimates, and fewer statistical tests with less good fit. Thus, the EFA seems weak compared with the CFA, can take advantage of chance occurrences in data, and possibly can be misleading. But perhaps the most important aspect of this continuum is that there is a lot of room for many types of FA between the CFA and EFA extremes. There are many FA approaches that are not extremely simple but are not extremely complex either. There are FA models that have some overidentified parameters but also some exactly identified parameters (most, in fact; see McArdle, 1991; McArdle & Cattell, 1994).

Given the favorable advances of CFA, it was somewhat instructive that the more experienced researcher among us (JLH) wanted to first look at the EFA to form a reasonable hypothesis about the data. This was partly indicative of the meta-theory that a specific factor structure fits the data no matter what models are tried (see Horn, 1972). There was also no intention to obscure the fact that statistical tests were not part of the original psychometric history of FA, and there was some resistance in using statistical tests at all (see Kaiser, 1976). These are due partly to what to many seem like absurd assumptions that need to be made for the resulting probabilities to be accurate (i.e., normality of uniquenesses, etc.). But partly this preference for EFA must also be due to years of training on EFA without the new flexibility of CFA.

For reasons defined by the sequence of Figures 12.1, 12.2, and 12.3, the explicit contrast between CFA and EFA is never really clear. In contrast to the newly developed approaches of CFA, the traditions of EFA are much older and were developed at a time when it was difficult to pose a rigorous pattern on factor loadings (Figure 12.2c), even if one was actually known. This is an obvious and clear benefit of CFA. In the past, the EFA was carried out in a sequence to (a) search for the most reasonable number of common factors, and (b) assuming more than one common factor, rotate the factor loadings to a position that seems most interpretable. The first step can use a generic LRT based on a limited number of degrees of freedom, but the second step usually relies on more substantive information—often when we use factor rotation we say we are trying to find a set of loadings that are "most reasonable" for these variables. For these reasons, many scientists now seem to find factor rotation as more artwork than science.

On the other hand, some researchers tend to think the one-factor model is identical in the CFA and EFA framework, and this ignores several key model possibilities. In a CFA, we have control over all the parameters; thus, as a prime example, we can fix the factor loadings at some known a priori values from another study. Indeed, fixed loadings would be an excellent example of a true confirmatory analysis, but a fixed loading is hardly ever part of any contemporary CFA application. The previous description of the Rasch model makes it seem like an ultrastrong CFA, but this also is a naive way to think the Rasch model is typically used. Instead of a strong CFA approach, a good fit of the Rasch model is simply the required goal of the analysis. That is, because a one-factor model with equal loadings is needed for the purposes of further measurement, this strategy implies that items should be eliminated until this goal is reached, and any statistical tests are merely an indicator of when to stop eliminating variables (Embretson & Reise, 2000). Obviously, any difference between CFA and EFA is muddled again.

Thus, the ethical problems with this newer CFA approach are at least twofold. First, as stated above, the use of the term *confirmatory* is a bit odd when we use this only to refer to the pattern hypothesis and we do not place an a priori value on the parameters. People reading about CFA for the first time may view this as a truly confirmatory procedure when, in fact, confirmation is used in a limited way. Second, the test of this CFA model is only exact when we specify the exact pattern in advance of the data—a priori. Unfortunately, the probabilistic basis of the LRT does not normally hold when there are attempts at a "refinement" or a "trimming" of the model using standard data analysis procedures. That is, it is hard to defend an approach where we simply drop variables and/or add arbitrary parameters until our model fits and then claim we can use the chi-square distribution to defend this model fit (cf. Brown, 2006). When we do not have an a priori hypothesis, we do not know whether the resulting probability is an index of any a priori sampling distribution.

A true CFA requires lots of effort at good measurement design and is not typical at all in the current SEM literature. It follows that a true CFA is rarely the case, and we much more typically need to make serious reorganizations and refinements of the model loadings using the data at hand. This standard "model fitting" approach to FA seems to make all CFAs move toward the EFAs, and there is nothing wrong with this. The main ethical problems emerge when we try to hide behind the CFA approach when in fact we are closer to doing EFA. If we do this, in essence, we are *lying with statistics* so we can tell a good story and get our work published. If this minor deception works once, it will probably work again and again; others will follow our lead, and inappropriate practices will become simply the way we do business.

Conclusion

To its great credit, the American Psychological Association (APA) is a leader in the recognition of ethics problems. Consider the book-length treatments of the *Ethical Principles of Psychologists and Code of Conduct* (APA, 2002) and the earlier book-length commentary of Canter, Bennett, Jones, & Nagy (1994). It is hard to find another group more interested and active in ethical practices than the APA. Unfortunately, when it comes to data analysis, arguably the only component common to all areas of behavioral science, APA guidelines seems to demand little. The APA guidelines present the outdated inclusion of extremely odd practices in presenting probability (multiple asterisks for different *p* levels), but these guidelines focus more on making tables for APA publications. The sensible suggestions of Wilkinson and the Task Force on Statistical Inference (1999) need to be taken more seriously. But, in reality, we must take the lead on this ourselves and express rules of good behavior using statistics. This chapter concludes with five suggested rules that are designed to lead to good practice in factor analyses.

1. *When reporting results, be honest*. The first principle of ethical FA is that we do not need PURITY of statistical rules and assumptions, but we do need HONESTY. Try to tell us exactly (as briefly as possible) how you selected the people, variables, and occasions, even if it is complicated. Consider missing data, outliers, and transformations, but please report their impacts on the results. Try to tell us exactly how you found the models used, especially if they were not a priori and if they emerged as part of the analysis. Tell us ALL relevant results, not just the BEST ones.
2. *The FA goal is replication*. Clarity is essential in any FA, and the key criterion in any experiment or analysis is *replication* (Lykken, 1968). Remember that confusion can be created by brevity, so we should not simply blame the reviewers. Reviewers want to make sure the work is fact, not fiction. What you are doing might not be clear enough to be replicated, and in this case you must clarify it. If the reviewers suggest you have broken the rules of "purity" (i.e., overall experimentwise error rate $\alpha > .05$), then you need to fight against this illogic directly and with vigor. Possibly you will need to change your favorite journal or funding agency, but at least you will be doing the right thing.
3. *Change the FA terminology*. The statistical terminology is often initially defined for one situation but found to be useful in another. Therefore, we should not simply use the same classical words when they mean something entirely different. For example, we

should immediately change theory or hypothesis → idea; test → examine; prove → demonstrate; data revealed → we noticed; significance → accuracy; predicted → connected; and controlled → adjusted. In the SFA context, we should substitute correlated errors → correlated specifics; confirmatory → overidentified, not rotatable; exploratory → exactly identified, rotatable; and a factor in FA is a thing → a factor in FA is evidence for the existence of a thing (Cattell, 1978). And if we do not know the basis of the probability statements we wish to make, we should drop them from our language and our analyses entirely.

4. *Primary analyses should use existing data*. There are very few barriers to the analysis of existing data, and this will allow most anyone to learn how to carry out analyses and demonstrate that we know how to analyze complex data problems. The analysis of existing data should be a formal requirement before anyone collects any new data on any individual. Of course, the APA publication system and the National Institutes of Health and National Science Foundation federal grant systems need to be willing to recognize this as valid research, too. One helpful hint: We can almost always think of the question in advance of the data selection—"*We cannot analyze a database, but we can analyze a question using a database!*"
5. *Any study should confirm, THEN explore*. In phase 1, confirm. Try to come into an analysis with a plan about the set of ideas you are going to examine and the data you are going to use to do so. This will permit a full and appropriate use of statistical probability tests and other indices of fit. Remember that we do not want the "best" model; we want the "set" of models that fit well separated from those that are "average" and "poor." In a subsequent phase 2, explore. Whether or not your favorite model fits the data on hand, try to improve the fit using any aspect of the data on hand. Do this completely so you can find something better than you had in phase 1. Who knows, maybe new results will then be able to be replicated by others.

In the merger of CFA and EFA into SFA, we are in the awkward position of trying to compromise two different statistical traditions: one old and one new. As long as nothing is lost, the newer techniques (CFA) offer improvements and should be favored over the older techniques (EFA). A key point here is that a lot can be lost in the blind application of CFA in situations where EFA might tell us a lot more, or at least the same thing (the HRS example here). In the classical and rigid approach to confirmation via hypothesis testing, we are taught to disdain the use of separate *t* tests in favor of the more rigorous one-way ANOVA (Scheffe, 1959). In an

exploratory mode, we are asked to wonder whether we missed anything important in the data we have collected (Tukey, 1962, 1977). Obviously, these are all valid points in an extended conversation about data analysis. However, we can all agree that it is wise to know exactly what we are doing ourselves, what boundary guidelines we need to follow, and to make sure we follow these rules ourselves.

Ethical guidelines in the area of FA can be as clear as in any other area of science. The main requirement is to report the sequence of analyses carried out so the reader can repeat or improve on these steps. Odd behaviors can emerge when the scientists forget to report a crucial step in the procedure, but this becomes an ethical problem when we do so on purpose or when we use a statistical test with known violations. This is as much an ethical violation as omitting a relevant reference because we do not like the author (i.e., we do not want to add to his or her h-index!). Unfortunately, there is little way to know when this is going on in these cases, so we must rely on the ethical behavior of the individual scientist. Of course, anyone who observes these behaviors—our students, our colleagues, our children—know what we are doing, and this alone may provide some needed ethical corrections.

In the SFA approach advocated here, we start with a strict CFA and move toward a more relaxed EFA—this is exactly what we typically need to do, and there is nothing unethical about it! This approach turns unethical when the sequence of procedures we use is not reported, perhaps in the hope that we can retain the illusion of the precision and power of the newest CFA-based statistical tests. As I have tried to show here, pretending to use CFA when we are really doing a form of EFA is fool-hardy at best—and devious at worst.[1]

References

Abelson, R. (1995). *Statistics as principled argument*. Mahwah, NJ: Erlbaum.

Albert M., Blacker D., Moss M. B., Tanzi R., & McArdle J. J. (2007). Longitudinal change in cognitive performance among individuals with mild cognitive impairment. *Neuropsychology, 21,* 158–169.

American Psychological Association. (2002). *Ethical principles of psychologists and code of conduct* (5th ed.). Washington, DC: APA Press.

[1] Author note: Thanks to Drs. A. T. Panter and Sonya K. Sterba for creating this opportunity, to Drs. Daniel and Lynda King for their insightful comments, and to Dr. John L. Horn for his classic advice in dealing with these complex technical and ethical issues: "People often underestimate the dangers of overplanning" (08/24/1990). The work reported here was initially presented at the APA symposium of the same title, Boston, August 2008. This work has been supported by National Institutes of Health Grant AG-007137.

Bowles, R. P., Grimm, K. J., & McArdle, J. J. (2005). A structural factor analysis of vocabulary knowledge and relations to age. *Gerontology: Psychological Sciences, 60B*, 234–241.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York: Guilford.

Browne, M., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. Bollen & S. Long (Eds.), *Testing structural equation models* (pp. 136–162). Beverly Hills, CA: Sage.

Browne, M. W. (2001). An overview of analytic rotation in exploratory factor analysis. *Multivariate Behavioral Research, 36,* 111–150.

Canter, M. B., Bennett, B. E., Jones, S. E., & Nagy, T. F. (1994). *Ethics for psychologists: A commentary on the APA ethics code.* Washington, DC: American Psychological Association.

Cattell, R. B. (1966). Psychological theory and scientific method. In R. B. Cattell (Ed.), *Handbook of multivariate experimental psychology* (pp. 1–18). Chicago: Rand McNally & Co.

Cattell, R. B. (1978). *The scientific use of factor analysis in behavioral and life sciences*. New York: Plenum.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists.* Mahwah, NJ: Erlbaum.

Fisher, R. A. (1925). *Statistical methods for research workers*. (14th ed., 1973). New York: Hafner.

Harlow, L. L., Mulaik, S. A., & Steiger, J. H. (1997). *What if there were no significance tests?* Hillsdale, NJ: Erlbaum.

Horn, J. L. (1972). State, trait, and change dimensions of intelligence. *The British Journal of Mathematical and Statistical Psychology, 42,* 159–185.

Horn, J. L., & McArdle, J. J. (1980). Perspectives on mathematical and statistical model building (MASMOB) in research on aging. In L. Poon (Ed.), *Aging in the 1980s: Psychological issues* (pp. 503–541). Washington, DC: American Psychological Association.

Horn, J. L., & McArdle, J. J. (1992). A practical guide to measurement invariance in aging research. *Experimental Aging Research, 18,* 117–144.

Horn, J. L., & McArdle, J. J. (2007). Understanding human intelligence since Spearman. In R. Cudeck & R. MacCallum (Eds.), *Factor analysis at 100 years* (pp. 205–247). Mahwah, NJ: Erlbaum.

Jennrich, R. I. (2007). Rotation methods, algorithms, and standard errors. In R. C. MacCallum & R. Cudeck (Eds.), *Factor analysis at 100: Historical developments and future directions*. Mahwah, NJ: Erlbaum.

Jöreskog, K. G. (1966). Testing a simple structure hypothesis in factor analysis. *Psychometrika, 31,* 165.

Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika, 34,* 183.

Jöreskog, K. G. (1977). Factor analysis by least-squares and maximum-likelihood methods. In K. Enslein, A. Ralston, & H. S. Wilf (Eds.), *Statistical methods for digital computers* (pp. 125–153). New York: Wiley.

Juster, F. T., & Suzman, R. (1995). *The Health and Retirement Study: An overview.* HRS Working Papers Series 94-1001. *Journal of Human Resources*, *30*, S7–S56.

Kaiser, H. (1976). [Review of the book *Factor analysis as a statistical method*]. *Educational and Psychological Measurement, 36,* 586–589.

Lawley, D. N., & Maxwell, A.E. (1971). *Factor analysis as a statistical method.* New York: Macmillan.

Lecoutre, B., Lecoutre, M.-P., & Poitevineau, J. (2001). Uses, abuses and misuses of significance tests in the scientific community: Won't the Bayesian choice be unavoidable? *International Statistical, 69,* 399–417.

Lykken, D. T. (1968). Statistical significance in psychological research. *Psychological Bulletin, 70,* 151–159.

MacCallum, R. C., Roznowski, M., & Necowitz, L. B. (1992). Model modifications in covariance structure analysis: The problem of capitalization on chance. *Psychological Bulletin, 111,* 490–504.

McArdle, J. J. (1991). Principles versus principals of structural factor analysis. *Multivariate Behavioral Research, 25,* 81–87.

McArdle, J. J. (1994a). Factor analysis. In R. J. Sternberg (Ed.), *The encyclopedia of intelligence* (pp. 422–430). New York: Macmillian.

McArdle, J. J. (1994b). Structural factor analysis experiments with incomplete data. *Multivariate Behavioral Research, 29,* 409–454.

McArdle, J. J. (1996). Current directions in structural factor analysis. *Current Directions in Psychological Science, 5,* 11–18.

McArdle, J. J. (2007). Five steps in the structural factor analysis of longitudinal data. In R. MacCallum & R. Cudeck (Eds.), *Factor analysis at 100 years* (pp. 99–130). Mahwah, NJ: Erlbaum.

McArdle, J. J. (2009). Latent variable modeling of longitudinal data. *Annual Review of Psychology, 60,* 577–605.

McArdle, J. J., & Cattell, R. B. (1994). Structural equation models of factorial invariance in parallel proportional profiles and oblique confactor problems. *Multivariate Behavioral Research*, *29*(1), 63–113.

McArdle, J. J., Fisher, G. G., & Kadlec, K. M. (2007). Latent variable analysis of age trends in tests of cognitive ability in the elderly U.S. population, 1993–2004. *Psychology and Aging, 22,* 525–545.

McArdle, J. J., & Lehman, R.S. (1992). A functionalist view of factor analysis. In D. F. Owens & M. Wagner (Eds.), *Progress in modern psychology: The contributions of functionalism to modern psychology* (pp. 167–187). Hillsdale, NJ: Erlbaum.

McArdle, J. J., & Nesselroade, J. R. (1994). Using multivariate data to structure developmental change. In S. H. Cohen & H. W. Reese (Eds.), *Life-span developmental psychology: methodological innovations* (pp. 223–267). Hillsdale, NJ: Erlbaum.

McArdle, J. J., & Prescott, C. A. (1992). Age-based construct validation using structural equation modeling. *Experimental Aging Research, 18,* 87–115.

McDonald, R. P. (1985). *Factor analysis and related methods.* Hillsdale, NJ: Erlbaum.

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Erlbaum.

Meredith, W., & Horn, J. L. (2001). The role of factorial invariance in measuring growth and change. In L. Collins & A. Sayer (Eds.), *New methods for the analysis of change* (pp. 201–240). Washington, DC: American Psychological Association.

Mulaik, S. A. (2009). *Foundations of factor analysis* (2nd ed.). New York: Chapman & Hall.

Scheffe, H. (1959). *The analysis of variance*. New York: Wiley.

Schwarz, N., Park, D., Knäuper, B., & Sudman, S. (Eds.). (1999). *Cognition, aging, and self- reports*. Philadelphia: Psychology Press.

Spearman, C. E. (1904). "General intelligence," objectively determined and measured. *American Journal of Psychology, 15,* 201–293.

Stapleton, L. M. (2002). The incorporation of sample weights into multilevel structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal, 9,* 475–502.

Tatsuoka, M. M., & Tiedeman, D. V. (1954). Discriminant analysis. *Review of Educational Research*, Washington, DC: AERA Press.

Tucker, L. R., & Lewis, C. (1973). The reliability coefficient for maximum likelihood factor analysis. *Psychometrika, 38,* 1–10.

Tukey, J. W. (1962) The future of data analysis. *Annals of Mathematical Statistics, 33,* 1–67.

Tukey, J. W. (1977). *Exploratory data analysis.* Reading, MA: Addison-Wesley.

Wilkinson, L., & the Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54,* 594–604.

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah, NJ: Erlbaum.

Yates, A. (1987). *Multivariate exploratory data analysis: A perspective on exploratory factor analysis.* Albany, NY: State University of New York Press.