Faking of Non-Cognitive Measures:

Factor Invariance Using Multiple Groups LISREL

Richard L. Frei, Richard L. Griffith,

Andrea F. Snell, Michael A. McDaniel, and Elizabeth F. Douglas

Department of Psychology

The University of Akron

Paper presented in G. Alliger (Chair) Faking Matters, symposium conducted at the 12th Annual Meeting of the Society of Industrial and Organizational Psychology, St. Louis, April, 1997.

Abstract

The authors tested for structural invariance between fakers and nonfakers on a biodata inventory using a four factor model (Dependable, Hard Working, Pleasant, and Tolerant) with LISREL 8. An overall chi-square test was calculated as a measure of fit of the model for both groups independently. The GFI for the honest group was .86, whereas the GFI was .69 for the faking group. In order to test for specific differences between the two groups, a nested model approach was used and a series of increasingly stringent hypotheses was tested (Alwin & Jackson, 1981). The authors found that: 1) the groups had different relationships among the observed variables, 2) the groups had a different number of latent factors, 3) the error variances were not the same between the faking and honest conditions, and 4) the relationships among the latent variables were different between the faking and honest conditions.

Faking of Non-Cognitive Measures:
Factor Invariance Using Multiple Groups LISREL

Response distortion is one of the most frequently cited criticisms of personality testing for personnel selection. Response distortion has been described in the literature with terms such as social desirable responding, claiming unlikely virtues, impression management, faking, intentional distortion, and self-enhancement (Hough, Eaton, Dunnette, Kamp, & McCloy, 1990; Hough & Paullin, 1994; Lautenschlager, 1994; Ones, Viswesvaran, & Korbin, 1995). Research on faking has found that subjects can, in fact, distort responses on non-cognitive measures when asked to do so (Hough et al., 1990; Lautenschlager, 1994). Ones et al. (1995) found that, when instructed to fake good, respondents were able to change their responses by almost half a standard deviation for the Big Five factors.

Although it is generally accepted that applicants have the ability to distort their self-descriptions, there is no consensus as to whether response distortion affects the reliability, predictive validity, and construct validity of non-cognitive selection tests. Hogan, Hogan, and Roberts (1996) argued that the ability to enhance scores on a personality inventory is, itself, a personality variable, one associated with good adjustment. Ones, Viswesvaran, and Reiss (1996) found that social desirability was related to stable personality constructs, specifically emotional stability and conscientiousness. They also found that correcting for faking on personality measures removed relevant variance from the measure. Therefore, Ones et al. (1996) concluded that socially desirable responding did not pose a problem in personnel screening. However, a

recent study by Douglas, McDaniel, and Snell (1996) revealed that faking can have substantial consequences with respect to reliability, construct validity, and criterion-related validity. Biodata and personality scales that measured four constructs (Dependable, Hard Working, Pleasant, and Tolerant) were administered to students under two conditions. In the first condition, subjects were instructed to answer the questionnaires as honestly as possible. In the second condition, subjects were instructed to answer the questionnaires in a way that would make a good impression. The authors also collected job performance data from students' employers. Using a between-subjects design, they found that, although faking increased the coefficient alpha of the scale, it also resulted in lower construct and criterion-related validity.

Why might making result in lower criterion-related validity? Douglas et al. (1996) offered two explanations. First, faking results in changes in subjects' rank-ordering, indicating that faking does not raise or lower everyone's scores in a consistent manner. Zickar, Rosse, and Levin (1996) used item response theory to demonstrate the effects on rank-ordering of applicants. Using a Monte Carlo design, they found that relatively few fakers are needed for the top end of the distribution to contain a high percentage of fakers. This is problematic because most employers use a top-down selection approach. Replicating the methodology of Zickar et al. (1996) with their own data, Douglas et al. (1996) found that, as the percentage of applicants who faked increased from 0-25%, 1) the mean validity of the measures dropped from .34 to .19 and 2) the number of fakers ranked in the top ten applicants rose from 0 to 9.

Second, recent research indicates that faking alters the underlying factor structure of the scale which, in turn, harms construct validity. Using an exploratory factor analytic framework,

Douglas et al. (1996) found that both the factor structure and the multitrait-multimethod (MTMM) matrix in the honest condition differed from that in the good impression condition. The current analysis improves upon the Douglas et al. (1996) study by examining the psychometric equivalence of the scales between groups using LISREL 8 (Joreskog & Sorbom, 1993).

LISREL8 is a confirmatory method which can be used to test whether the underlying factor structures between two groups have the same number of latent variables with the same indicators and the same specification of fixed and free parameters. Unlike exploratory factor analysis, the researcher specifies the expected factor model and tests to see how well each data set fits the model. LISREL allows you to not only specify the number of factors in the model, but also the factor loadings, factor intercorrelations, and error variances. For example, Schmit and Ryan (1993) found that the five factor model often used to describe personality fit the data provided by student respondents better than the data provided by job applicant respondents. Upon further analysis, they identified a sixth factor (the "good employee" factor) that was present in the applicant sample but not in the student sample.

In another study that used multiple groups LISREL analysis, Schmit, Ryan, Stierwalt, and Powell (1995) attempted to improve the predictive and face validity of personality-based selection tests used to select customer service personnel. They did this by altering both the test instructions and test items of the NEO-PI to make them work-specific. They hypothesized that by making the inventory more work-related, they would provide the applicants with a common frame of reference from which to base their answers and, therefore, increase criterion-related validity. The authors used LISREL 8 to compare the factor structures of the original vs. modified inventory.

They found that the modified work-specific instructions/items resulted in higher predictive validity with no change in factor structure.

These results suggest that the structure of a non-cognitive instrument changes under different instructional sets. The following analyses were conducted to empirically test the measurement differences found by Douglas et al. (1996). The authors used LISREL 8 and a confirmatory framework to examine differences in the overall goodness of fit (GIF) and the factor structure between the honest and the good impression conditions. If the preliminary analyses of Douglas et al. (1996) are correct, the four factor model should better fit the data collected under the honest instructional set than the data collected under the good impression instructional set. In addition, there should be structural differences between the honest and good impression conditions.

Hypothesis I: As indicated by the Goodness of Fit (GFI), the four factor model will fit the data in the honest condition better than the data in the good impression condition.

Hypothesis II: The structure of the biodata scale in the honest condition will significantly differ from the structure of the scales in the good impression condition.

IIa: The covariance matrices of the honest and good impression conditions will not be equal.

IIb: The factor pattern (number of factors) will differ between the honest and good impression conditions.

IIc: Factor loadings will differ between the honest and good impression conditions.

IId: Error variances will differ between the honest and good impression conditions.

IIe: Correlations between the latent variables will differ between the honest and good impression conditions.

Method

Data for this analysis was collected by Douglas, McDaniel, and Snell (1996). 600 subjects had been randomly assigned to one of two conditions. In the honest condition, subjects were asked to complete the biodata questionnaire as honestly as possible. In the faking condition, subjects were told to answer the questions in such a way as to make as favorable an impression as possible. We tested for differences between the honest and good impression groups using a four factor model (Dependable, Hard Working, Pleasant, and Tolerant). Sixteen observed variables (i.e., item composites) were linked to the four corresponding factors. Each observed variable was allowed to load on only one of the four factors.

Analyses

An overall chi-square was calculated as a measure of fit of the model for both groups independently. The test for measurement invariance was conducted to determine whether the model for the two samples has the same number of latent variables with the same loadings of indicators, the same measurement errors, and the same factor intercorrelations. In order to test for differences between the two groups, a nested model approach was used and a series of increasingly stringent hypotheses was tested (Alwin & Jackson, 1981). For each hypothesis, chi-

square values from two models were compared; one in which parameters were freely estimated separately for the two groups and one in which the parameters were estimated with the constraint that they be the same for both groups. If the constrained model fits the data significantly worse than the unconstrained model, the parameters estimated for the two groups are judged to be significantly different.

Results

The GFI for the honest group was .86, whereas the GFI was .69 for the good impression group. These findings suggest that the four factor model fits the honest group relatively well but does not fit the good impression group very well. This also indicates that there are possible differences in structure between the honest and good impression conditions. In order to test this hypotheses, we conducted a test of measurement invariance using multiple groups LISREL analysis (Joreskog, 1971).

Hypothesis IIa examined whether the covariance matrices of the two groups were equal. To test this hypothesis, an overall chi-square for the two groups was calculated. If the overall chi-square value is significant, there is variance between the two groups. The overall chi-square for the honest/good impression groups was significant χ^2=(1113.45, 136df), indicating that the groups have a different relationship between the observed variables. In order to investigate where the differences in the measurement model lie, four additional hypotheses were tested. For each of these hypotheses, two nested models were compared, one which constrained certain parameters in the model to be equal across the groups and one that did not. Hypothesis IIb

stated that the number of factors is the same for both groups. A significant chi-square difference $_{s}^{2}$=(510.46, 48df) between the constrained and unconstrained models indicated that the groups had a different number of latent factors. Hypothesis IIc stated that the factor loadings were the same across the groups. A significant chi-square difference $_{s}^{2}$=(61.15, 16df) indicated that the groups had differential factor loadings. Hypothesis IId stated that the error variances were the same across the groups. A significant chi-square difference $_{s}^{2}$=(98.64, 16df) indicated that the error variances were not the same across the honest and good impression conditions. Hypothesis Ie stated that the correlations between the latent variables were the same across groups. Again, a significant chi-square difference $_{s}^{2}$=(60.45, 16df) between the constrained and unconstrained models indicated that the relationships among the latent variables are different across the honest and good impression conditions.

Discussion

Douglas et al. (1996) found that faking increased the coefficient alpha of the biodata scale. These additional analyses suggest that the increase in reliability is an artifact of the common response distortion method factor and is not an indication that the constructs are more reliably measured. In addition, the previous evidence that faking harms criterion-related validity is supported by these findings. If everyone was faking at the same rate, there would be no difference in criterion-related validity between the honest and good impression groups. Faking would be comparable to adding a constant to everyone's score. However, this evidence suggests that criterion-related validity is harmed because everyone is not faking at the same rate. Some

possible influences on propensity to fake may be motivation to fake, opportunity to fake, and various characteristics of the test, such as warnings or verifiability of the items.

Previous research has suggested that situational effects may change the trait-factor structure of non-cognitive tests (e.g., Douglas et al., 1996; Stokes, Hogan, and Snell, 1993). This study improves on past research, which relied on exploratory factor analysis to pinpoint differences across groups. By utilizing a confirmatory approach, we provide conclusive evidence that faking does alter the factor structure (and, thus, the construct validity) of personality inventories. These findings imply that inventories validated using a student or incumbent sample may not be valid with an applicant sample, given differences in motivation to fake. Further research is needed to pinpoint how these structural differences affect predictive validity.

References

Alwin, D. F. & Jackson, D. F. (1981). Applications of simultaneous factor analysis to issues of factor invariance. In D. F. Jackson and E. F. Borgatta (Eds.) *Factor analysis and measurement in sociological research.* Beverly Hills: Sage.

Douglas, E. F., McDaniel, M. A., & Snell, A. F. (1996). The validity of non-cognitive measures decays when applicants fake. Paper presented at the 14th annual Academy of Management Conference.

Hogan, R., Hogan, J., & Roberts, B. W. (1996). Personality measurement and employment decisions: Questions and answers. *American Psychologist*, 51, 469-477.

Hough, L. M., Eaton, N. K., Dunnette, M. D., Kamp, J. D., & McCloy, R. A. (1990). Criterion-related validities of personality constructs and the effect of response distortion on those validities. *Journal of Applied Psychology*, 75, 581-595.

Hough, L. M. & Paullin, C. (1994). Construct-oriented scale construction: the rational approach. In G. S. Stokes, M. D. Mumford, & W. A. Owens (Eds.) *Biodata Handbook: Theory, Research, and Use of Biographical Information in Selection and Performance Prediction.* Palo Alto: Consulting Psychologist Press.

Joreskog, K. G. & Sorbom, D. (1993). *LISREL 8: Analysis of linear structural relationships by the method of maximum likelihood.* Chicago: International Education Services.

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

Lautenschlager, G. J. (1994). Accuracy and faking of background data. In G. S. Stokes,

M. D. Mumford, & W. A. Owens (Eds.) *Biodata Handbook: Theory, Research, and Use of Biographical Information in Selection and Performance Prediction.* Palo Alto: Consulting Psychologist Press.

Ones, D. S., Viswesvaran, C., & Korbin, W. P. (1995). Meta-analysis of fakability estimates: Between-subjects versus within-subjects designs. Paper presented in F. L. Schmidt (Chair) *Response Distortion and Social Desirability in Personality Testing and Personnel Selection.* Symposium conducted at the 10th annual meeting of the Society of Industrial and Organizational Psychology, Orlando, April.

Ones, D. S., Viswesvaran, C., & Reiss, A. D. (1996). The role of personality testing for personnel selection: The red herring. *Journal of Applied Psychology*, 84, 1021-1035..

Schmit, M. J. & Ryan, A. M.(1993). The big five in personnel selection: Factor structure in applicant and nonapplicant populations. *Journal of Applied Psychology*, 78, 966-974.

Schmit, M. J., Ryan, A. M., Stierwalt, S. L., Powell, A. B. (1995). Frame-of-reference effects on personality scale scores and criterion-related validity. *Journal of Applied Psychology*, 80, 607-620.

Stokes, G. S., Hogan, J. B., & Snell, A. F. (1993). Comparability of incumbent and applicant samples for the development of biodata keys: The influence of social desirability. *Personnel Psychology*, 46, 739-762.