Demonstration of How Score Reliability is Integrated into SEM with Multiple Exogenous Latent Variables

Lee M. Wolfle
Virginia Polytechnic Institute and State University

Yetkiner and Thompson (2010) demonstrated, within a Structural Equation Model (SEM) framework, that Spearman's (1910) corrected correlation coefficient generalized to the case of multiple manifest indicators. In the current study, the argument was taken a step further by demonstrating the effect of unreliable measurement on structural parameters in a three-variable SEM model. In general, the results from this study are in accord with previous findings, but there are considerable variations depending on the mix of high and low reliability specifications.

A century ago, Spearman (1910) "called attention to the fact that the apparent degree of correspondence between any two series of measurements is largely affected by the size of the 'accidental' errors in the process of measurement" (p. 271). He proposed that one could obtain the "desired correct correlation" with:

$$r_{x'y'} = r_{yx}/\sqrt{r_{yy}}$$

(simplified for one group) where $r_{x'y'}$ is the estimated correlation corrected for randomly generated imperfect score reliability, r_{xy} is the uncorrected correlation coefficient, and r_{xx} and r_{yy} are the reliability coefficients for the *X* and *Y* scores respectively. At an extreme, for example, if one obtained a measured correlation of 0.25 with two measures whose reliabilities were just 0.50 each, the desired corrected correlation would be estimated to be near unity.

Combining these so-called corrected measures of association into multiple regression analyses was seen infrequently since their inception. For example, Kerlinger and Pedazur (1973) did not even address the issue in their popular text. Although Wright (1925) was able to incorporate unmeasured, or corrected, variables in his extraordinary analysis of corn and hog correlations, it was not until the early 1970s that a practical way was found to incorporate corrected measures of association with regression and path analysis; generally called structural equation models (SEM). The work of Jöreskog (1973), Jöreskog and van Thillo (1972), Keesling (1972), and Wiley (1973) provided breakthroughs and there are now any number of computer software applications available with names such as LISREL (Jöreskog & Sörbom, 1999), EQS (Bentler, 1995), AMOS (Arbuckle, 2007), or Mplus (Muthén & Muthén, 2010).

To illustrate how measurement error can impact estimates of association in the framework of structural equation models and the general linear model, Yetkiner and Thompson (2010) have provided an exemplary analysis of real data. Using the data collected by Holzinger and Swineford (1939) in the late 1930's from two schools in the Chicago area, Yetkiner and Thompson extended that analysis by considering multivariate reliability estimates and their effect on the underlying corrected correlation between latent, or corrected, variables. In their analysis, they chose three measures of spatial ability and three measures of verbal ability (these data will be described in further detail below). Their model thus consisted of two latent variables, each with three indicators. They examined the influence of various manipulations of reliability estimates on the estimated correlation between the two latent factors: spatial and verbal ability. The zero-order correlations among the three measures of spatial ability and the three measures of verbal ability (nine such correlations) ranged from 0.14 to 0.37, with a simple average around 0.24. Yetkiner and Thompson considered four different scenarios for reliability estimates. When the reliabilities were set equal to near zero, the estimated correlation between spatial and verbal abilities was estimated to be .80. When the reliabilities were set to be nearly perfect, the estimated correlation was .38. They then considered two more practical and realistic scenarios for estimating the reliabilities. In one case, they estimated the reliabilities using previously published (Harman, 1976) estimates and found the corrected correlation between spatial and verbal ability to be 0.46. Finally, they estimated the reliabilities from the data in the model and found the corrected correlation to be 0.52. All of the corrected estimates were larger than the raw-score correlations and as one would expect, the scenarios that corrected for very low reliabilities inflated the estimated correlation between the latent factors the most.

Yetkiner and Thompson's (2010) analysis elucidated the bivariate case, but did not address the more complicated questions about the influence of various reliability estimates in the multivariate case. For example, what might happen when an imperfectly measured dependent variable is regressed on two imperfectly measured independent variables? The outcome is not at all immediately evident. As Wolfle (1979), among others, has shown, corrected structural coefficients (analogous to regression coefficients) are not simply inflated by correcting for random measurement error. Indeed, such estimates may be inflated or deflated depending on the mixture of reliability and structural components of the model.

The current article thus extends Yetkiner and Thompson's (2010) analysis by considering a new model in which a latent variable, grades, is regressed on the latent variables of spatial and verbal ability, and examines differences in the structural estimates resulting from different scenarios concerning the reliabilities with which the manifest indicators were obtained. As with Yetkiner and Thompson, the purpose is to illustrate exactly how changes in score reliability can impact the structural coefficient estimates with a structural equation model.

The Model

The model to be estimated in these illustrations is shown in Figure 1. There are two exogenous latent variables, spatial and verbal ability, each measured with three manifest indicators. There is one endogenous latent variable, grades, also with three manifest indicators. Spatial and verbal ability are seen to be correlated for reasons unspecified in this model. The latent grade variable is shown to be caused by spatial and verbal ability, plus an unanalyzed residual disturbance term. The latent variables are depicted within ellipses. The manifest variables are shown without being enclosed. All of the manifest variables have errors of measurement, e_i, associated with them. The curved, double-headed arrow indicates a correlation. The straight arrow points toward a dependent variable from its independent cause.

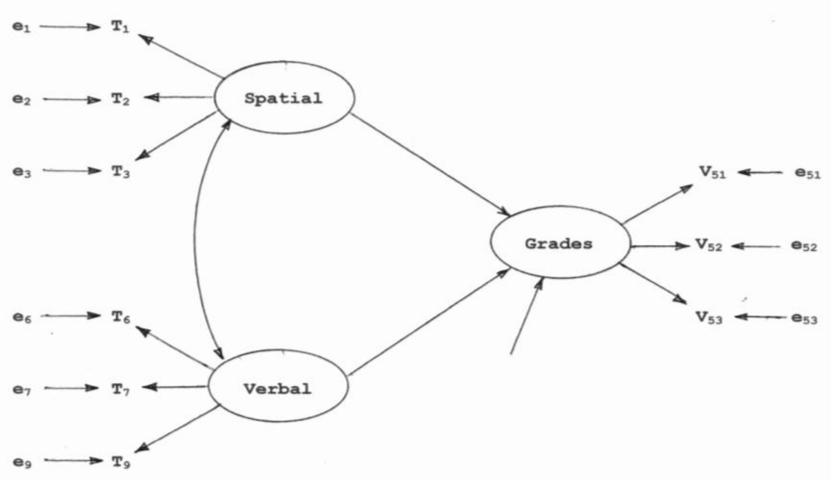

Figure 1. Model of spatial and verbal ability with grades and varying assumptions about error variance.

Data

In their exercise, Yetkiner and Thompson (2010) used real data from the widely-used dataset collected by Holzinger and Swineford (1939). They used test scores of 301 children collected from two schools in the Chicago area. With some modifications, the same data were used in the current study.

Scores on six tests from Holzinger and Swineford (1939) were used. The first three variables were used as indicators of spatial ability. The last three variables were used as indicators of verbal ability.

In addition to the test scores collected by Holzinger and Swineford (1939), school marks were obtained for students in the Grant-White school in Forest Park, Illinois in nine subjects: citizenship, literature, reading, language, history, elementary science, arithmetic, drawing, and music. Three of these school marks were used in the current analysis as indicators of grades.

Holzinger and Swineford (1939) Tests		School Marks	
Label	Description	Label	Description
T_1	Visual Perception Test	V_{51}	Literature
T_2	Cubes	V_{52}	Reading
T_3	Paper Form Board	V_{53}	Language
T_6	Paragraph Comprehension		
T_7	Sentence Completion		
T_9	Word Meaning		

Because data on grades were only collected at the Grant-White school, the analysis reported here is based on 144 students and not the 301 used by Yetkiner and Thompson (2010).

Correlations for the six test scores were obtained from Holzinger and Swineford's Table 9 (1939, p. 30). Correlations among the three school marks and their association with test scores were obtained from Holziner and Swineford's Table 23 (p. 53). Means and standard deviations were calculated from raw data (Holzinger & Swineford, pp. 52, 86-90). Those correlations, means, and standard deviations are reproduced in Table 1.

Table 1. Correlations, Means, and Standard Deviations for Variables in Model of Spatial and Verbal Ability on Grades (N = 144)

	T_1	T_2	T_3	T_6	T_7	T_9	V_{51}	V_{52}	V_{53}
T_1	1.000								
T_2	0.318	1.000							
T_3	0.379	0.191	1.000						
T_6	0.335	0.234	0.260	1.000					
T_7	0.304	0.157	0.269	0.722	1.000				
T_9	0.326	0.195	0.261	0.714	0.685	1.000			
V_{51}	0.260	0.228	0.292*	0.576	0.594	0.597	1.000		
V_{52}	0.291	0.221	0.327*	0.665	0.613	0.652	0.697	1.000	
V_{53}	0.161	0.079	0.174*	0.551	0.496	0.454	0.554	0.675	1.000
Mean	29.517	24.800	14.303	9.952	18.848	17.283	2.140	2.217	2.387
S.D.	6.959	4.445	2.823	3.375	4.649	7.947	1.202	0.913	0.988

Source: Holzinger and Swineford (1939): Tables 9, 22, 23 (correlations indicated by * are for T_{25} in Table 23), and Appendix II.

In their analyses, Yetkiner and Thompson (2010) manipulated the reliabilities of the measured variables in such fashion to show the effects of high and low reliability. In addition, they used as input previously-reported reliability estimates for the six tests scores taken from Harman (1976, p. 123). The Cronbach's alpha coefficients for the six test scores were: 0.756, 0.568, 0.544, 0.651, 0.754, and 0.870, respectively.

Previously-published reliability estimates for the three school marks used as indicators of grades were less easily obtained. They were apparently not reported in Holzinger and Swineford (1939), Swineford and Holzinger (1942), Holzinger and Harman (1941), or Swineford (1947). Nor were the raw data contained in Holzinger and Swineford's appendices. Instead, I conducted a simple confirmatory factor analysis of the nine school marks reported by Holzinger and Swineford (p. 53) and used the resulting reliability estimates from that analysis. For the three grade reports used in the following analyses; therefore, as previously obtained reliability estimates, the values of 0.488, 0.730, and 0.682 were used respectively for marks in literature, reading, and language.

Results

In order to examine the influence of reliability estimates upon the structural parameters shown in Figure 1, several difference scenarios were considered. As with Yetkiner and Thompson (2010), high and low reliability specifications were used with previously-reported reliability estimates. With three reliability scenarios specified for each of the three latent variables, there were 27 different combinations to consider.

Let us first consider what we might expect to find. Holzinger and Swineford (1939, pp. 55-56) found that the most important factor contributing to school marks was a halo factor. Of their four group factors, only the verbal factor had any statistically significant effect. The effect from their spatial factor was mostly small and negative. For the data used here, a confirmatory analysis was performed using LISREL 8.8 with no external constraints applied to the manifest variables; the reliabilities were thus estimated within the analysis. For the structural portion of the model, this resulted in the following:

$$G' = 0.051\ S + 0.824\ V$$

where the estimated parameters are reported in standardized form, with an estimated R^2 of .73 and a respectable measure of fit of $\chi^2 = 19.60$ with 24 degrees of freedom.

The overwhelming influence on grades (as indexed by literature, reading, and language) is from verbal ability with a small (and statistically insignificant) positive effect from spatial ability.

To estimate the influence of manipulating the reliability estimates upon these parameters, three scenarios were considered. High reliabilities were set equal to 0.90 and low reliabilities were set equal to 0.10. Previously reported reliabilities were all of intermediate range as reported above. Error variances were specified in the analysis as (1 – r_{xx}), in standardized form, since it was the standardized results that were of interest. For example, to specify the previously estimated error variance for literature, the value (1 - 0.488) = 0.512 was used. The interested reader can replicate the analysis using previously reported reliability estimates using the data and Lisrel syntax reported in the Appendix. Substitutions of error variances of .10 for high reliability and 0.90 for low reliability would allow the replication of all of the results reported here. Of course, with the data provided, the interested reader could supply any combination of error variance estimates.

The results of the 27 different scenarios are reported in Table 2. In Table 2, the designation of H indicates high reliability (0.90), L indicates low reliability (0.10), and P indicates previously-reported reliability estimates. The estimated correlation, r_{sv}, between the spatial and verbal ability factors is reported in Table 2 along with the standardized estimates of the effects of spatial and verbal ability on grades. Also shown are the R^2 for the structural equations and the χ^2 goodness-of-fit, L^2, for the model. Since the structural portion of the model is fully specified, lack of fit can be assumed to rest in the measurement portion of the model.

Table 2. Structural Equation Estimates for Grades with Varying Specifications for Reliabilities

r_{xx}		Standardized Coefficients			
S V G	r_{SV}	Spatial	Verbal	R^2	L^2
H H H	0.425	0.062	0.746	0.600	1815.5
H H P	0.425	0.062	0.788	0.666	1445.1
H H L	0.425	0.085	0.940	0.959	1535.3
H P H	0.428	0.058	0.751	0.604	1603.9
H P P	0.428	0.058	0.792	0.670	1233.8
H P L	0.428	0.080	0.948	0.970	1323.5
H L H	0.525	-0.170	1.045	0.934	1699.1
H L P	0.520	-0.162	1.075	1.000	1328.9
H L L	0.501	-0.057	1.027	1.000	1429.1
P H H	0.460	0.039	0.755	0.598	709.9
P H P	0.460	0.038	0.796	0.644	339.5
P H L	0.460	0.055	0.950	0.954	429.9
P P H	0.464	0.033	0.760	0.603	498.4
P P P	0.464	0.033	0.801	0.668	128.2
P P L	0.464	0.049	0.959	0.965	218.1
P L H	0.569	-0.232	1.088	0.949	593.6
P L P	0.558	-0.211	1.102	1.000	223.4
P L L	0.538	-0.096	1.048	1.000	323.9
L H H	0.625	0.122	0.696	0.606	662.2
L H P	0.625	0.122	0.738	0.672	291.9
L H L	0.625	0.167	0.871	0.970	382.1
L P H	0.629	0.115	0.704	0.610	450.7
L P P	0.630	0.115	0.744	0.675	80.6
L P L	0.629	0.159	0.882	0.980	170.3
L L H	0.773	-0.450	1.303	0.994	545.9
L L P	0.736	-0.298	1.198	1.000	176.0
L L L	0.721	-0.071	1.050	1.000	276.0

Table 2 is arranged to show nine configurations where the reliabilities for the three spatial variables are set to be high, with nine embedded configurations for the verbal and grade factors. Then, reliability estimates for the spatial factor are specified as previously-reported with nine scenarios for verbal and grades; followed by a specification of low reliability for the spatial factor. The results of the 27 scenarios are shown.

From Table 2, several patterns are evident. For the correlation estimates of spatial and verbal ability, in general, the lower the reliability estimates were specified for these two factors, the higher was the estimated correlation between the factors. For example, when reliability coefficients were set to be high for the manifest indicators of all three latent factors (HHH), the estimated correlation between spatial and verbal ability was seen to be 0.425, the estimated correlation increased to 0.464 when previously reported reliabilities were specified (PPP), and when the reliabilities were all set to be low (LLL), the estimated correlation was 0.721.

For the effects of spatial and verbal ability on grades, the effect of the verbal factor was always greater than the effect of the spatial factor. However, there were considerable variations depending on the mix of reliability specifications. Probably the most profound effect can be seen when the reliability estimates for the verbal factor were set to be low. On those occasions, the effect of verbal ability on grades was inflated to an extent that may even be unreasonable. The coefficients for verbal were greater than 1.0, the corresponding estimated effect of spatial ability was negative, and the R^2's for the equation were near unity.

That being said, there were some uniform changes observed in the resulting standardized coefficients primarily as a result of changes in the specification of the reliabilities of the dependent variable. Examining the estimated coefficients for spatial ability, when the reliabilities for spatial were high and verbal were low, and when the reliabilities for grades were set to be high and as previously reported (HLH and HLP), respectively, we see the results were -0.170 and -0.162. But when the reliabilities for grades were set to be low (HLL), the estimated coefficient for spatial ability was -.057, which was a considerable change. Similar changes in the estimated spatial coefficient can be seen when spatial reliabilities were set as previously-reported and verbal reliabilities were low (PLH and PLP); the coefficients were -0.232 and -0.211, but -0.096 when reliabilities for grades were low (PLL). The pattern was repeated when both spatial and verbal reliabilities were set to be low (LLH and LLP); the estimated spatial coefficients were -0.450 and -0.298, but -0.071 when reliabilities for grades were low (LLL).

As with the estimated effect of spatial ability, similar changes were observed for the estimates of verbal ability. With reliability estimates for verbal ability set to be low, along with low reliability estimates for grades, the estimated standardized coefficients for verbal were 1.027, 1.048, and 1.050 for conditions where the reliabilities for the spatial factor were high (HLL), previously-reported (PLL), and low (LLL), respectively. All three of these estimated coefficients were lower in absolute value than when the reliabilities for grades were set to be higher. When the estimated coefficients for spatial ability were examined for conditions other than when verbal reliabilities were low, a consistent pattern was revealed. We observed that there was a slight decline in value from when reliabilities for the spatial variables were changed from high to those previously reported, but increased as the imposed reliabilities for spatial were set to be low. For example, the estimated effect for high reliabilities for spatial, coupled with previously-reported reliabilities for verbal and grades (HPP), was 0.058. It declined slightly to .033 for the imposition of previously-reported reliabilities for all three variables (PPP), but when spatial reliabilities were set to be low, the corresponding estimate (LPP) was 0.115.

There was a consistent effect on the estimated standardized coefficient of the spatial factor as a result of changing the imposed reliabilities of the dependent variable; grades, it always increased in value. For example, when reliabilities for spatial and verbal abilities were set to be high and grade reliabilities were either high or as previously-reported (HHH and HHP), respectively, the effect was 0.062. When grade reliabilities were set to be low (HHL), the estimated effect was .085. The pattern repeated itself when the reliabilities for spatial were set to those previously-reported (PHH and PHP); these were 0.039 and 0.038, respectively, but PHL was 0.055. The pattern repeats itself again when the reliabilities for spatial were set to be low, the spatial effects for LHH and LHP were 0.122. When the reliabilities for grades were set to be low (LHL), the estimated effect was 0.167.

There is a similar pattern evident for the estimated effect of verbal ability on grades. Without going into all of the detail, the numbers change, but the pattern is consistent. That is, when the reliabilities for grades were set to be low, the estimated effect of verbal ability on grades increased. For example, when the imposed reliabilities for the spatial and verbal factors were high, and the imposed reliabilities for grades were set to be high, or as previously-reported (HHH and HHP), respectively, the estimated effects of verbal ability were 0.746 and 0.788. When low reliabilities were imposed for grades (HHL), the effect was 0.940.

In general, we have seen that lower reliability specifications resulted in higher estimates of the structural parameters. But, when low reliability estimates were specified for the verbal factor, which had the highest degree of association with grades, the resulting effects overwhelmed the system; thus, causing the estimated structural estimate from verbal ability to grades to exceed 1.0 in value. That being said, we have also seen that low reliability estimates for the dependent variable, the grade factor, also had noticeable effects reverberating through the system.

Discussion

Yetkiner and Thompson (2010) demonstrated within a SEM framework that Spearman’s (1910) corrected correlation coefficient generalized to the case of multiple manifest indicators of two latent factors. In the current study, the argument was taken a step further by demonstrating the effect of unreliable measurement on structural parameters in a three-variable SEM model. In some respects, the results shown in this study generalize previous findings. The largest estimates of the correlation between verbal and spatial ability occurred when the reliabilities for these variables were low.

The estimated effects of the two independent factors on grades were more complicated. In general, when lower reliabilities were specified for the dependent variable, the estimated structural coefficients for the two independent variables increased. But, when low reliability estimates were imposed for one of the independent variables, verbal ability, which had the strongest association with grades, the estimated structural parameters were unreasonably large due to the overcompensation for the low reliability estimates for verbal ability.

In sum, as many have pointed out, it is crucially important to pay close attention to the reliability of measurement of variables in a SEM or any analytic framework. This can become even more important (e.g., Wolfle, 1979; Wolfle, 1985) when comparing structural coefficients across groups, such as whites and blacks, rural and urban, etc., because differences in reliability estimates across groups can have profound effects on the estimates of structural parameters within the groups.

References

Arbuckle, J. L. (2007). *AMOS 16.0 user’s guide.* Spring House, PA: Amos Development Corporation.

Bentler, P. M. (1995). *EQS structural equations program manual.* Encino, CA: Multivariate Software, Inc.

Harman, H. H. (1976). *Modern factor analysis* (3rd ed.). Chicago: The University of Chicago Press.

Holzinger, K. J., & Harman, H. H. (1941). *Factor analysis: A synthesis of factorial methods.* Chicago: The University of Chicago Press.

Holzinger, K. J., & Swineford, F. (1939). *A study in factor analysis: The stability of a bi-factor solution* (Supplementary Educational Monographs No. 48). Chicago: University of Chicago.

Jöreskog, K. G. (1973). A general method for estimating a linear structural equation system. In A. S. Goldberger & O. D. Duncan (Eds.), *Structural equation models in the social sciences* (pp. 85-112). New York: Seminar Press.

Jöreskog, K. G., & Sörbom, D. (1999). *Structural equation modeling with the SIMPLIS command language.* Lincolnwood, IL: Scientific Software International, Inc.

Jöreskog, K. G., & van Thillo, M. (1972). *LISREL: A general computer program for estimating a linear structural equation system involving multiple indicators of unmeasured variables* (Research Bulletin #RB-72-56). Princeton, NJ: Educational Testing Service.

Keesling, J. W. (1972). *Maximum likelihood approaches to causal analysis.* Ph.D. Dissertation. Department of Education: University of Chicago.

Kerlinger, F. N., & Pedhazur, E. J. (1973). *Multiple regression in behavioral research.* New York: Holt, Rinehart and Winston.

Muthén, L. K., & Muthén, B. O. (2010). *Mplus user’s guide* (6th ed.). Los Angeles: Author.

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology, 3,* 271-295.

Swineford, F. (1947). Examination of the purported unreliability of teachers’ marks. *The Elementary School Journal, 47*, 516-521.

Swineford, F., & Holzinger, K. J. (1942). *A study in factor analysis: The reliability of bi-factors and their relation to other measures* (Supplementary Educational Monographs No. 53). Chicago: University of Chicago.

Wiley, D. E. (1973). The identification problem for structural equation models with unmeasured variables. In A. S. Goldberger & O. D. Duncan (Eds.), *Structural equation models in the social sciences* (pp. 69-83). New York: Seminar Press.

Wolfle, L. M. (1979). Unmeasured variables in path analysis. *Multiple Linear Regression Viewpoints, 9,* 20-56.

Wolfle, L. M. (1985). Postsecondary educational attainment among whites and blacks. *American Educational Research Journal, 22*, 501-525.

Wright, S. (1925, January). *Corn and hog correlations.* Washington, DC: U.S. Department of Agriculture Bulletin 1300.

Yetkiner, Z. E., & Thompson, B. (2010). Demonstration of how score reliability is integrated into SEM and how reliability affects all statistical analyses. *Multiple Linear Regression Viewpoints*, *36*, 1-12.

Send correspondence to:
Lee M. Wolfle
Virginia Polytechnic Institute and State University
Email: lwolfle@vt.edu

APPENDIX

```
Lisrel Syntax for the Model Where Error Variances Were Set Equal to
(1 - Previous Reliability Estimates): Model P P P
Standardized Model with Grant-White Data
Observed Variables:
T1 T2 T3 T6 T7 T9 V51 V52 V53
Correlation Matrix
1.00
0.318  1.00
0.379  0.191  1.00
0.335  0.234  0.260  1.00
0.304  0.157  0.269  0.722  1.00
0.326  0.195  0.261  0.714  0.685  1.00
0.260  0.228  0.292  0.576  0.594  0.597  1.00
0.291  0.221  0.327  0.665  0.613  0.652  0.697  1.00
0.161  0.079  0.174  0.551  0.496  0.454  0.554  0.675  1.00
Standard Deviations
1  1  1  1  1  1  1  1  1
Sample Size = 144
Latent Variables: Spatial  Verbal  Grades
Relationships:
  T1 =     Spatial
  T2 =     Spatial
  T3 =     Spatial
  T6 =     Verbal
  T7 =     Verbal
  T9 =     Verbal
  V51 =    Grades
  V52 =    Grades
  V53 =    Grades
  Grades = Spatial Verbal
Set the Variance of Spatial to 1.0
Set the Variance of Verbal to 1.0
Set the Variance of Grades to 1.0
Set the Error Variance of T1 to 0.244
Set the Error Variance of T2 to 0.432
Set the Error Variance of T3 to 0.456
Set the Error Variance of T6 to 0.349
Set the Error Variance of T7 to 0.246
Set the Error Variance of T9 to 0.130
Set the Error Variance of V51 to 0.512
Set the Error Variance of V52 to 0.270
Set the Error Variance of V53 to 0.318
LISREL Output: sc nd=3 it=500
End of Problem
```