TWO PATH ANALYTIC MODELS OF TEST-RETEST RELIABILITY:
Examples from Social Psychology

Jaime B. Valera
Department of Agricultural Education
University of the Philippines
at Los Baños

Two models of estimating test reliability are illustrated using the path analytic methods of Heise (1969) and Wiley & Wiley (1970). The method is briefly described and empirical data from two studies are used to demonstrate how test-retest reliability may be separated from respondents' stability.

The first case is from an evaluation of effects of training in a family planning method involving 74 adult subjects. The measurement was on a knowledge test administered before and twice more after a training. The same general design (one indicator three-wave model) was applied to an attitude test towards the environment on 30 high school subjects in the second case. The original purpose of this second study was to validate the effects of an attitude change slide show to increase pro-environment attitudes.

This report shows that the first model, Heise's, gives a similar reliability coefficient as the second model of the Wileys'. The second model offers three separate estimates of reliability, once for every testing. The major difference lies in the assumptions in the two models which can be tested in the second model. The estimated reliabilities from the two models were larger than simple re-test correlations. In the second case, an illustration of low reliabilities and stability coefficients is given. Some likely explanations for this are discussed.

The test-retest method of determining reliability of tests is not frequently used because administering the same test to the the same set of subjects twice is problematic (Nunnaly, 1970; Downie and Heath, 1974; Kidder, 1981) although intuitively, it is "the simplest approach to determining the reliability coefficient '... (Nunnaly, 1970, p. 22)." There are two major disadvantages: (1) the retest coefficient, usually the Pearson coefficient of correlation, does not reflect error due to sampling and content and, (2) the retest method does not take into account memory and other similar factors within the individual. The problem of unreliability due to memory and other time dependent sources of error is a consequence of either immediate or delayed retesting. When two test administrations are close to each other, say a day or a few days, the retest estimates may be spuriously high. When the two testing sessions are separated by a time interval of several months, the opposite effect of an underestimation of true reliability may occur. For these reasons and the obvious cost in re-administering tests, the test-retest method is relegated to an academic principle, a text-book issue. It is not normally proposed nor used in routine psychological and other behavioral studies.

Purpose

The following exposition outlines two path analytic models as techniques of estimating true reliability separate from test-retest stability. The technique presented here is not new as it has been pioneered in the sociological literature by Heise (1969). This report presents occasions where judgment of a test's reliability may be enhanced through the use of path analysis and to offer some contemporary examples in the Philippine setting. To our knowledge, this technique is not well-known among psychological researchers and testers.

Briefly, Heise's and later, Wiley and Wiley's (1970) methods are techniques that address the problem of simple correlations as estimates of time-bound sources of error. If there are under- or over-estimates of reliability, then the researcher's responsibility is to determine the extent of such errors. The second purpose is to illustrate the two models using empirical examples having three administrations of parallel tests: a pre-test, a first post-test and a second post-test. The two examples also illustrate how training may be evaluated for relatively long-term and short-term effects.

The first model due to Heise (1969) was explained and illustrated in an earlier paper (Valera, 1985). The same model will be repeated here for clarification and substantive purposes since Wiley and Wiley's 1970 model will be difficult to understand without the historical and substantive background which is in the work of Heise.

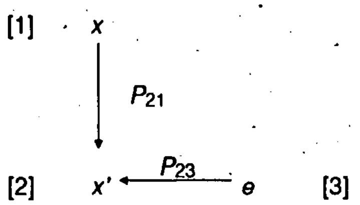
Path Analysis and Reliability

The intractability of the test-retest situation may be resolved by the use of path analytic methods which is the main contribution of Heise (1969). This section outlines the application of path analysis although a detailed presentation will not be done here. (Those interested may look into the work done by Land, 1969; Duncan, 1975; Asher, 1976 and Kenny, 1979).

The fundamental equation for reliability is:

$$x' = x + e \qquad [1]$$

Figure 1. Path model of the fundamental measurement equation x' = x + e.



x = true score
x' = observed score
e = error component
Path model equation: $x' = P_{21}x + P_{23}\,e$

From the domain sampling model (Nunnaly, 1967) of parallel tests, we can express this equation in a path model as it is shown in Figure 1. The observed score (x') is presumed to be generally caused by two events: (1) the true score (x) and (2) errors represented by the random component, e. The arrows in Figure 1 are the **paths** labelled $P21$ and $P23$. These represent effect parameters of a linear relationship from the presumed true score (x) and its error component (e). These coefficients or paths are determinable by using linear regression. In this illustration, $P21$ is simply the standardized regression coefficient when we predict x' from x, the true score. Theoretically, the path from the error is estimable given that $P21$ is r^2. The path $P23$ is the contribution to x' not attributable to x, that is $1 - r^2$: The practical problem, however, is that we never know the true scores (x), and if we did, we do not bother with the fallible observed score. The major contribution of path analysis in reliability estimation is to make available a method of estimation of the contributions (paths) from the true score if there were multiple observations of the trait or characteristic being measured.

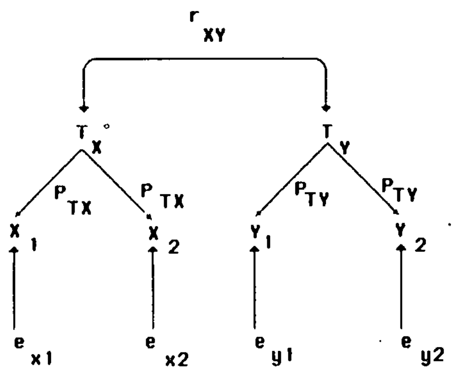
A correction for attenuation model can be illustrated using path analysis. This was shown by Heise (1969) and repeated by Valera (1985). It will be noted that the familiar correction for attenuation formula for the correlation can be mathematically derived from a model of parallel tests illustrated in Figure 2 which shows two measurements ($x1$ and $x2$, $y1$ and $y2$) for two traits or variables (x and y). The correlation between x and y can be shown to be equal to the correlation of the first measures of x and y divided by the product of the roots of the cross-correlations of the two parallel measures of the respective traits:

$$r_{xy} = \frac{r_{x1y1}}{\sqrt{r_{x1x2}}\,\sqrt{r_{y1y2}}} \qquad [2]$$

The main assumption is that the two measures for each of the traits are identical or parallel.

Furthermore, even if the contributions of the true scores are not equal, the paths are still determinable by path analytic estimation, i.e. using ordinary least squares regression.

Figure 2. Path model of a parallel test, a multiple indicator system.



Test-retest Models

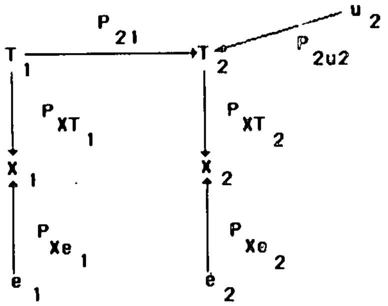
The test-retest situation is illustrated in Figure 3. Here we have two fallible measures over time, $x1$ and $x2$. The corresponding true scores are $t1$ and $t2$. The errors for the two measurements are also indicated. The path $P21$ is the coefficient which indicates the stability of the variable over time. If there are other "unknown" factors that may affect the stability of the true scores, these are explicitly "captured" in the error term, $u2$. This error, just as in any other path model, is assumed to be random. Its effect on the true score is the path coefficient $P2u2$.

An integral assumption for the test-retest model is that the contribution of the true scores on the observed scores remains essentially the same. Thus, the path is equal to P_{xt} in either time periods. Another simplifying assumption for this and similar path models is that $e1$ and $e2$ are not correlated with the true scores. Hence, there is no connection (arrows) between errors and true scores just as the error term, u, is assumed to be uncorrelated with $x1$ and $x2$ ($r_{x1u} = 0$). Given these assumptions, it can be shown that the correlation between the first and the second measurements ($r_{x1x2} = P_{xt}P21P_{xt}$) is the product of the paths connecting or joining $x1$ and $P2$ through the two true scores. In short, we have the correlation:

$$r_{x1x2} = P^2_{xt}P21 \qquad [3]$$

If the variable remains stable over time, which is a strong assumption, then $r_{x1x2} = P^2 21$. This strong assumption means that $P21$ is 1.00. The characteristic or, more appropriately, the persons did not change from the first to the second administration.

Figure 3. Path model of a simple test-retest measurement.



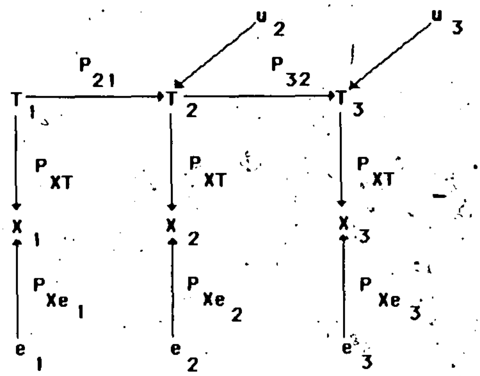
x = observed score
T = true score
e = error

If we allow that $P21$ to be fallible, or less than 1.00, then we would need to estimate P_{xt}. It is clear by now that the stability coefficient is not the same as the path P_{xt}. Given these, it is expected that the correlation between xs may be estimated if we know the two paths. As it is, there is only one correlation (one equation) and there are two unknowns. Thus, a simple test-retest model will not be sufficient. We need additional equations (observations).

Heise's Model:

Figure 4 is Heise's model which solves this inadequacy. There is an additional observation, $x3$ which implies an additional true score component, $t3$ and its error term, $u3$

Figure 4. Path model of two retests measurement (Adopted from Heise, 1969)



This extension allows us to generate three path model equations for a "just identified" system of equations:

$$r_{x1x2} = P_{xt}P_{21}P_{xt} = P^2_{xt}P_{21} \qquad [4]$$

and, similarly:

$$r_{x2x3} = P^2_{xt}\,P_{32} \qquad [5]$$

$$r_{x1x3} = P^2_{xt}P_{21}P_{32} \qquad [6]$$

From the above we can obtain,

$$P_{21} = r_{x1x2}/P^2_{xt} \qquad [7]$$

and,

$$P_{32} = r_{x2x3}/P^2_{xt} \qquad [8]$$

Substituting equations [7] and [8] into equation [6] we have:

$$r_{x1x3} = \frac{r_{x1x2}\,(r_{x2x3})}{P^2_{xt}}$$

so that we can estimate P_{xt} from the square root of:

$$P^2_{xt} = \frac{r_{x1x2}\,(r_{x2x3})}{r_{x1x3}} \qquad [9].$$

The reliability coefficient, then, is the correlation of the first and second measures times the correlation of the second and the third divided by the correlation of the first and the third observations or measures. This estimate is separate and different from the changes in the true scores which are presumably due to person characteristics (like memory and other factors) and not due to the test itself.

Using this path model, Heise showed that we have:

(1) stability coefficients (P_{21} and P_{32}) and, a separate

(2) reliability estimate (P_{xt}).

There is a third stability coefficient that can be estimated, that is, the stability from the first to the third measurement. By appropriate algebraic manipulation (after determining the values for P_{21} and P_{bxt} from equation [9]), this coefficient is:

$$P_{31} = P_{21}P_{32} \qquad [10]$$

This can also be computed directly from the correlations by:

$$\frac{(r_{x1x3})^2}{r_{x1x2}\,(r_{x2x3})} \qquad [10.1]$$

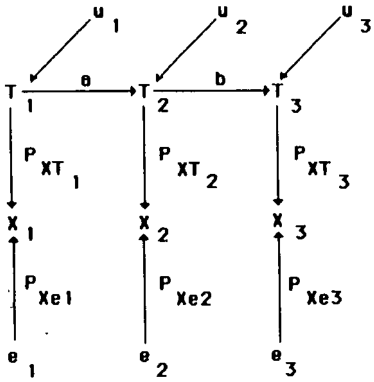
The Model of Wiley and Wiley

Wiley and Wiley improved on the model of Heise to what is now generically called a "three-wave, one-indicator model."

Diagramatically, the test-retest model of the Wileys is the same as that of Heise (see Figure 5) with one important exception.

In Heise's model which is very similar to the one presented by the Wileys in Figure 5, the reliabilities across time, P_{xt}s are assumed to be the same so that the reliability of the test remains the same in all three testing occasions. Thus the reliability remains constant. Wiley and Wiley argued that this is not true in many instances. Briefly, the Wileys suggested that if the basic path model (Figure 1) is correct and if the reliability is interpreted as the proportion of the variance in the observed measurement (x) accounted for by the true component (t), or "the ratio of the true component variance to the ob-

Figure 5. Path model of a one indicator three-wave model after Wiley & Wiley (1970)



served variance" (Borhnstedt, 1983, p. 73), the total variance is the sum of the true score component and error variances as in the following formula for reliability:

$$r_{xx} = \frac{Var(t)}{Var(t) + Var(e)} \qquad [11]$$

It is clear in this equation [11], that as the error decreases, the reliability approaches 1.00. When the error is absolutely zero, we have perfect reliability. Equation [11] may be rewritten as the the ratio of the true score variance (*var* (*t*)) divided by the observed variance (*var* (*x*)) or, for computational purposes, a formula that does not involve the true scores, r_{xx} is:

$$1 - [\frac{Var(e)}{Var(x)}] \qquad [11.1]$$

With this clarification, Wiley and Wiley showed that it is possible for the reliability of an observed measurement to change without the error variance changing as it can be seen in equation 11. This situation may be obtained in cases where the tested population is not the same in the three occasions or when the variance changes due to social and psychological processes over time. In all these situations we cannot safely assume constant reliability. The resulting estimates from the model were shown to be biased by the Wileys.

Furthermore, Wiley and Wiley have shown that if the reliabilities are assumed to vary across time, the three-wave one-indicator model will be under-identified so that there will be more unknowns than available equations. However, they have shown that if variances and co-variances are used instead of correlations, it is possible to estimate the parameters of the model with the restriction of assuming only that the error variances are equal. The detailed proofs and discussion on this are in Wiley and Wiley (1970) and Werts et al. (1971).

The derivations of the computation formulas for the reliabilty and stability co-efficients follows:

Using Figure 5 and beginning with the normal predicting equations ([12], [13], [14]) for the true scores, as well as the equations ([15], [16], [17]) for the observed scores or measurements,

$$t_1 = u_1 \qquad [12]$$

$$t_2 = au_1 + u_2 \qquad [13]$$

$$t_3 = b(au_1 + u_2) + u_3 \qquad [14]$$

$$x_1 = t_1 + e_1 \qquad [15]$$

$$x_2 = t_2 + e_2 \qquad [16]$$

$$x_3 = t_3 + e_3 \qquad [17]$$

we can obtain the computational normal equations which do not need the true scores (by substituting the right hand side of equations [12], [13] and [14] into equations [15], [16] and [17]):

$$x_1 = u_1 + e_1$$

$$x_2 = au_1 + u_2 + e_1$$

$$x_3 = b(\ au_1 + u_2) + u_3 + e_3$$

Following the rules for path analysis (which are explained in detail by Duncan, 1975) we may break down these normal equations in terms of variances and co-variances with the necessary assumption that the variances of the errors are equal, i.e., $V(e1) = V(e2) = V(e3)$:

$$V(x1) = V(u1) + V(e)$$

$$V(x2) = a^2V(u1) + V(u2) + V(e)$$

$$V(x3) = b^2[a^2V(u1) + V(u2)] + V(u3) + V(e)$$

$$c(x_1x_2) = aV(u_1)$$

$$c(x_1x_3) = abV(u_1)$$

$$c(x_2X_3) = b[a^2(u_1) + V(u_2)]$$

From the above, we have six equations to estimate six unknowns: a, b, $V(e)$, $V(u1)$, $V(u2)$ and $V(u3)$:

$$a = [\frac{c(x_1x_2)}{V(u_1)}] \qquad [18]$$

$$b = [\frac{c\ (x_1x_3)}{c(x_1x_2)}] \qquad [19]$$

$$\frac{V(e) = V(x_2) - [c(x_2x_3)]}{b} \qquad [20]$$

$$V(u_1) = V(x_1) - V(e) \qquad [20.1]$$

$$V(u_2) = V(x_2) - [\ ac(x_1x_2) + v(e)] \qquad [20.2]$$

$$V(u_3) = V(x_3) - [\ bc(x_2x_3) + V(e)] \qquad [20.3]$$

V(x1), $V(x2)$ and $V(x3)$ are the variances of the measurements in the three testing occasions. The respective reliabilities of each testing may then be computed using variables defined above with the formula for r_{x1x1} equation [11]:

$$r_{x1x1} = \frac{V(u_1)}{V(u_1)+V(e)} \qquad [21]$$

$$r_{x2x2} = \frac{a^2V(u_1) + V(u_2)}{a^2V(u_1) + V(u_2) + V(e)} \qquad [22]$$

$$r_{x3x3} = \frac{b^2[a^2V(u_1) + V(u_2)] + V(u_3)}{b^2[a^2V(u_1) + V(u_2)] + V(u_3) + V(e)} \qquad [22.1]$$

To estimate the true stability coefficients a and b above are not sufficient since they are unstandardized regression slopes. We may estimate the stability coefficients by noting that the path coefficient or standardized path regression slope is equal to (in this example from variable 1 to 2):

$$P_{21} = b_{21}[\frac{\sqrt{V(x_1)}}{\sqrt{V(x_2)}}] \qquad [23]$$

Using a, b and ab for the respective unstandardized slopes, we may estimate the stability coefficients with the following computational equations:

$$S_{12} = a[\frac{\sqrt{V(u_1)}}{\sqrt{a^2V(u_1) + V(u_2)}}] \qquad [24]$$

$$S_{23} = b\ [\frac{\sqrt{a^2V(u_1) + V(u_2)}}{\sqrt{b^2(a^2V(u_1) + V(u_2) + V(u_3)}}] \qquad [25]$$

and, using the rules of path analysis, $S13$ is simply the product of stability from the first to the second testing and the stability of the second to the third testing:

$$S_{13} = S_{12}(S_{23}) \qquad [26]$$

This can serve as a check for the following derived equation for this stability coefficient which is :

$$S_{13} = ab\ [\frac{\sqrt{v(u_1)}}{\sqrt{b^2(a^2V(u_1) + V(u_2) + V(u_3)}}] \qquad [27]$$

From these formulas, the separate reliabilities as well as the stability coefficients could be determined.

Empirical Examples

Case 1: Evaluating Long Term Effects of Training

The data has been presented elsewhere (Valera, 1982; 1985) and was the result of an evaluation of a training for field workers of the Population Commission in 1980. The design is a pre-test, first post-test and a second post-test panel observation. The first post-test was administered immediately after a three-day training on the Natural Family Planning method, in particular the calendar method. The second post-test was given to the same subjects six months later to determine the relative long-term effects of training on the knowledge gained.

The tests were substantially the same multiple-choice items. Nineteen of the items were the same in the three tests administered. The pre-test and the first pre-test had 31 items wherein the 19 items were embedded. The second post-test had only 19 items in it. The data is part of larger analysis which involved a Solomon-four-group design. In this report, only those subjects who took the tests on all three occasions are relevant. The subjects were 74 adults of varying educational attainments. The tests were group administered.

Case 2:Validating an attitude scale

The second data set is from a recent unpublished study (Valera and Jerusalem, 1989) which attempts to develop environmental attitude scale in two Filipino languages (Pilipino and Cebuano). There were three administrations of an attitude test towards the environment. This involved splitting into two a 20-item (5 point Likert type scale) test previously selected through item analysis. A 10-item test was randomly administered for each of the testing occasions. The purpose of the test was to determine the validity of the attitude scale through an experimental exposure of subjects to a pro-environment slide tape show.

Briefly, if the scale was valid and if the slide tape show was successful in investing positive attitudes towards environmental conservation, then the post-tests should show a gain or an increased average (positive) attitude towards the environment. (The scores were such that the higher value indicated a positive attitude towards the environment). The first post-test was administered one week after the slide show. The second post-test was administered about three weeks after the first post-test to determine if the observed attitude change was also relatively long-term. The data used here is from 30 subjects using the Pilipino version of the scale.

The subjects were junior high school students of a vocational fishery school located on the western side of Laguna de Bay. The attitude scale was composed of two distinct parts, a general environmental attitude scale and a specific set of items for a coastal environment scale. Thus, there were two 10-item sets. The tests were given without dummy or other items besides the biographical information questionnaire. The items were not "hidden" unlike the knowledge item tests in Case 1 above. It should also be noted that since there were only a few items involved, the second post-test had the same items, albeit in a different order, as the pre-test. As in Case 1, the second post-test was not completely administered in a group. Some subjects had to be given the test individually having been absent during the group administration of the test. These were a minority however, and did not exceed five percent of the sample.

The Data

Table 1 summarizes the relevant data needed for the panels: (2A) a sub-scale concerning the coastal environment and (2B), another for the general attitudes towards the environment.

Case 1 has smaller absolute means since it is a correct or wrong type of test with a maximum of 17 points while case 2 scores are averages of 10 items with a maximum of 5 points per item.

Table 1. Variances, covariances, correlations and mean scores for the pre-test and post-test of the knowledge test and attitude to the environment scales.

	1. Pre-test	2. Post-test1	3. Post-test 2	Means
Case 1(n=74)				
1. Pre-test	V1=15.21	C12= 9.66	C13= 3.16	10.8
2. Post -test1	r12=0.59	V2= 17.64	C23= 5.67	12.3
3. Post-test 2	r13=0.30	r23= 0.50	V3= 7.29	11.1
Case 2A (n=30)				
1. Pre-test	V1= 7.07	C12= 1.83	C13= 1.46	33.4
2. Post-test 1	r12= 0.23	V2= 8.94	C23= 4.95	35.4
3. Post-test 2	r13= 0.18	r23= 0.55	V3= 9.06	35.2
Case 28				
1. Pre-test 1	V1=11.47	C12= 5.72	C3= 4.88	33.3
2. Post-test 1	r12= 0.54	V2= 9.87	C23= 3.56	35.3
3. Post-test 2	r13=0.43	r23= 0.34	V3= 11.15	34.3

* The diagonals are the variances(V's); the upper triangle contain the covariances(C's) while the lower triangle of the matrices for each case contains the bivariate correlations (r's).

Note: Case 2A is a summary of attitude scale scores specific to the coastal environment; Case 2B scores are for the general attitudes towards the environment.

Results and Discussion

Using the values in Table 1, Tables 2, 3 and 4, give the results for Heise's and the Wileys' models.

Table 2. The reliabilities and stability coefficients for Case 1 in the Heise and Wiley & Wiley models.

	Model 1 (Heise)		Model 2 (Wiley & Wiley)
Reliability*	(P_{xT}) =	0.983	rel 11 = 0.979
			rel 22 = 0.983
			rel 33 = 0.958
Stabilities**	S12	0.600	0.601
	S23	0.508	0.307
	S13	0.305	0.215
Path Coefficients***			V(e) = .307
			A =.648
			B =.327
			V(u1) = 14.90
			V(u2) = 11.07
			V(u3) = 5.13

* Using equations [9] for model 1; [21], [22], [23] for model 2.

**Using equations [7], [8] and [10] for model 1; [24], [25] and [27] for model 2.

*** Using equations [18], [19], [20], [20.1], [20.2] and, [20.3].

Note: To distinguish reliability from correlation, the following tables and the discussion, refer to reliability coefficients as "rel."

In Table 2, the coefficients from Model 1, Heise's model, are similar to those in Model 2. In particular, the reliability coefficient in Model 1 (Path P_{xt}) is exactly the same as the second reliability coefficient in Model 2 (rel 22), which is equal to 0.983. This is as it should be when the assumptions of Model 1 are correct. There are slight differences between the stability coefficients of the two models by the general trend and conclusions that could be made from these are substantively the same. This means that the largest stability, due to the short time interval, is between the first and second testing (pre-test and the first post-test). The least or smallest stability is between the first and the third testing (i.e., between the pre-test and the second post-test).

Table 3. Reliabilities and stability coefficients for Case 2 (Specific attitudes towards the coastal environment) forthe Heise and Wiley & Wiley models.

	Model 1 (Heise)		Model 2 (Wiley & Wiley)
Reliability	(P_{xT}) =	9.703	rel 11= 0.607
			rel 22 = 0.689
			rel 33 = 0.693
Stabilities	S12	0.327	0.356
	S23	0.783	0.581
	S13	0.256	0.215
Path Coefficients			V(e) = 2.777
			A= .426
			B= .803
			V(u1) = 4.29
			V(u2) = 5.38
			V (u3) = 2.31

Table 4. Reliabilities and stability coefficients for Case 2B (General attitudes towards the environment) scores for the Heise and Wiley & Wiley models.

	Model 1 (Heise)		Model 2 (Wiley & Wiley)
Reliability	(P_{xT}) =	0.427	rel 11 = 0.503
			rel 22 = 0.423
			rel 33 = 0.489
Stabilities	S12	1.007	1.165
	S23	0.796	0.912
	S13	1.264	0.768
Path Coefficients			V(e) = 5.687
			A = .991
			B = .853
			V(u1) = 5.77
			V(u2) = −1.49

The path coefficients found for Model 2 show the estimated values for the errors, $V(e)$ and the $V(u\text{'s})$. The coefficients A and B are the unstandardized path coefficients representing the stability from one period to another. The small error (0.307) is consistent with the high reliability coefficients in the three testing periods. This should be the case since the reliability is directly a function of the ratio of the error variance to the total variance (equation [11.1]). The errors of the true component from time 1 through time 3 are not unreasonable but their relative large values indicate why the stability coefficients are low.

Table 3 shows that the reliability coefficient in Model 1 is similar to the one obtained in Model 2 (rel 22). Model 2 shows the refinement that is introduced by the governing assumptions such that the reliability across the three administrations of the test cannot be said to be constant. The variance of the error is also relatively larger than that of Case 1. This error variance is up to 39 percent (2.77/7.07 x 100) of the lowest variance of the three tests. In Case 1 the relative size of the error variance is less than one half of one percent which explains the higher reliabilities in Case 1 when compared to that of Case 2A. The stability coefficients show that there was a declining stability among test-takers through time with high stability between the second to the third testings (S_{23}). In sum, the test for specific attitudes towards the coastal environment has been nominally if not moderately reliable.

Case 2B shows what happens when both the test reliabilities and the stability are low due to a relatively large error variance which was up to 56 percent of the observed test variances. It will be noted that one of the variances for the true component, $V(u_2)$, turned out to be negative—a theoretical impossibility. Variances are an average sum of squares which can never be negative. Thus, an assumption of the model has been violated. This may indicate that some interdependence could be operating so that the model as drawn in Figure 4 is not acceptable. A possible model is one that has some of the errors ($V(u$'s)) correlated with the true components or with other errors, or the measurement errors ($V(e)$) are not equal and independent of the x's, the observed variances. In any case, the coefficients show why the reliabilities and stabilities are not large.

It suggests that conclusions regarding the change in mean levels which imply the effect of the slide show on attitudes may not be supported since the measurement is unreliable. The fact that we have observed inconsistent error variances illustrates the importance of explicitly stating the assumptions regarding the measurements being performed. In Case 2B, we see the untenability of the conclusions about the changes between testing periods. The mean score differences are statistically significant, that is, the mean for the pre-test against the means of the post-tests are significant well beyond the usual level (p =.05). Thus, the validity of the change of the mean attitude levels is vitiated by poor reliability estimates.

One interesting result which seems to violate expectations is the fact that the last two reliabilities (rel 22 and rel 33) for Case 2A are larger than the reliability of the pre-test. A parallel situation is also observed for the stability coefficients. Stability S_{12} is lower than S_{23}. This appears unusual since the first two testing periods are closer to each other (one week) than the last two (three weeks). This is probably due to the fact that the second post-test and the pre-test had the same items. This explains why the third reliability is high and stability coefficient S_{23} is larger than S_{12}. It is possible that the three-week separation was not enough to erase recognition of the same items though the second test (first post-test) had a different set of items. It is apparent why the subjects appear to be more stable in second to the third test administration than between the first and the second administration of the test.

Furthermore, the 10-item attitude tests for Case 2 may have contributed to low reliabilities. In general, and other things being equal, tests with more items have higher reliabilities. However, Case 2A illustrates that so long as the assumptions are satisfied and the test-takers are not as unstable as they apparently were in their responses to the general environment items (Case 2B), this need not be the case..

In summary, Case 1 is an illustration of a study where the change in mean levels of response to a test is tenable as attested to by the reasonable coefficients and variances for the two models. Case 2A is also an illustration of a test with minimally acceptable reliability levels while Case 2B shows an instance where the coefficients are less than desirable. These may be due to the unreasonable levels of error variances suggesting that the measurement model may be something other than those postulated by the Heise or the Wiley and Wiley models. These cases also illustrate once more that simple test-retest correlations understate the actual reliabilities. Borhnstedt (1983) correctly emphasized that, "This difference undoubtedly results from the fact that *simple test-retest correlations confound change (stability) with unreliability* (p.83) [underscoring is Borhnstedt's]." In other words, simple correlation coefficients lead to the underestimation of the contribution of the hypothesized true score component on the empirical measurements.

References

Asher, H. 1976. *Causal Modeling*. Beverly Hills: Sage. Borhnstedt, G. W. 1983. Measurement. In P.H. Rossi, J.D. Wright & A.B. Anderson (eds.) *Handbook of Survey Research* (pp. 69–121). New York: Academic Press.

Downie, N. & Heath, R. (1974). *Basic Statistical Methods*. New York: Harper and Row.

Duncan, O.D. (1975). *Introduction to Structural Equation Models*. New York: Academic Press.

Heise, D.R. (1969). Separating reliability and stability in test retest correlation. *American Sociological Review*, 34: 93–101.

Kenny, D.A. (1979). *Correlation and Causality*. New York: John Wiley.

Kidder, L.H. (1981). *Research Methods in Social Relations*. Fourth Edition. New York: Holt Rinehart and Winston.

Land, K. (1969). Principles of path analysis. In E. Borgatta (ed.) *Sociological Methodology 1969* (pp. 3–37). San Francisco: Jossey-Bass.

Nunnaly, J. (1967). *Psychometric Theory*. New York: McGraw-Hill.

Nunnaly, J. (1970). *Introduction to Psychological Measurement*. New York: McGraw-Hill.

Valera, J.B. 1982. An evaluation of effects of strategies used to improve delivery of rhythm as a family planning method: Final Report. Population Center Foundation.

Valera, J.B. (1985). Test retest reliability: An empirical example using path analysis. Paper read at the 22nd Annual Convention of the Psychological Association of the Philippines. August 6–8. Manila.

Valera, J.B. & Jerusalem, F. (1989). A Filipino attitude towards the environment scale. Unpublished research. U.P. at Los Baños.

Wiley, D.E. & Wiley, J.A. (1970). The estimation of measurement error in panel data. *American Sociological Review*, 35: 112–117.

Werts, C.E., Joreskog, K.G. & Linn, R. (1971). Comment on 'The estimation of measurement error in panel data.' *American Sociological Review*, 36: 110–112.