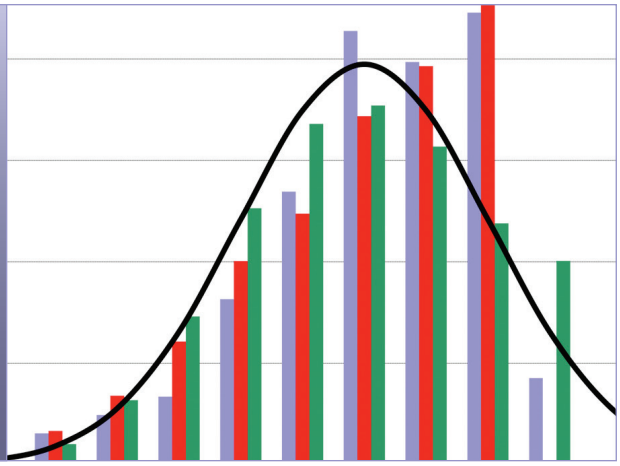
ACT Research
Report Series
2012 (10)



Statistical Considerations in Choosing a Test Reliability Coefficient

David Woodruff
Yi-Fang Wu

December 2012

ACT®

For additional copies, write:
ACT Research Report Series
P.O. Box 168
Iowa City, IA 52243-0168

Statistical Considerations in Choosing a Test Reliability Coefficient

David Woodruff
Yi-Fang Wu

Abstract

The purpose of this paper is to illustrate alpha's robustness and usefulness, using actual and simulated educational test data. The sampling properties of alpha are compared with the sampling properties of several other reliability coefficients: Guttman's λ_2, λ_4, and λ_6; test-retest reliability; as well as congeneric reliability. The comparisons are based on different sample sizes and test models comprising dichotomous item and polytomous item tests. It is concluded that alpha is indeed a lower bound to reliability except under the assumption of essential-tau-equivalence; however, it is robust to violations of this condition, and its values are competitive with other coefficients' values based on splitting a test into parallel halves or repeating a test to estimate test-retest reliability. Because it is not always possible to construct parallel half-tests or obtain test-retest data, it is very useful to have reliability coefficients, such as alpha, which are free of these constraints.

Statistical Considerations in Choosing a Test Reliability Coefficient

Introduction

Several articles have criticized the use of coefficient alpha to estimate test reliability or test internal consistency (Bentler, 2009; Green, Lissitz, & Mulaik, 1977; Green & Yang, 2009a; Green & Yang, 2009b; Hattie, 1985; Revelle & Zinbarg, 2009; Schmitt, 1996; Sijtsma, 2009a; Sijtsma 2009b). In general, these articles criticize alpha on two counts. The first is that alpha is a lower bound to reliability, and not a very good one unless stringent assumptions are satisfied. The second is that the actual test attribute measured by alpha, often labeled internal consistency, unidimensionality, or item homogeneity, is either poorly defined or poorly measured by alpha. For an alternative view, see Feldt and Qualls (1996). Although these articles demonstrate alpha's shortcomings, they ignore some of its advantages, such as conceptual simplicity, computational simplicity, and a known sampling distribution that is robust to violations of its assumption of a compound symmetric (CS) multivariate normal distribution (MVN) for the item scores. In addition, although alpha can be a poor estimate of test reliability for short tests with very heterogeneous item variances and inter-item covariances, it is fairly robust to violations of essential-tau-equivalence, and is a useful estimate of test reliability for many tests. This article uses actual and simulated educational test data to illustrate alpha's robustness and usefulness.

Although there are limitations to alpha as a measure of test reliability for some test data, it is argued here that test reliability is a well defined concept and, for a variety of test data, is well estimated by coefficient alpha. Simulations using dichotomous and polytomous educational data as well as covariance matrices derived from MVN distributed item scores are used to support these conclusions. Also, the sampling properties of alpha are compared with the sampling properties of several other reliability coefficients derived by Guttman (1945) and further

discussed by Revelle and Zinbarg (2009). These other coefficients are Guttman's λ_2, λ_4, and λ_6. Note that alpha is equal to Guttman's λ_3, and Guttman's λ_4 equals alpha for a test divided into odd and even split halves. In addition the correlation between two parallel tests, denoted ρ_{12} is investigated. Also investigated is reliability based on a one-factor factor analysis model, also called congeneric test score model (Jöreskog, 1971; Raykov, 1997), with its coefficient denoted ρ_c. All of the Guttman coefficients can be computed by SPSS (2006) and alpha can be computed by SAS (2008). Reliability coefficients that explicitly depend on determining the multifactor structure of a test are not considered. Determining the factor structure of a test is not always unequivocal especially for tests composed of dichotomous items or polytomous items with few score points, although specialized software such as TESTFACT (du Toit, 2003) does exist for tests composed entirely of dichotomous items.

Test reliability is commonly defined in two different ways. For the usual test score model $X = T + E$, the first definition which is used by Gulliksen(1950), Guttman (1945), and Sijtsma (2009a) is the test-retest model. In this model, reliability is defined as the correlation between X and a parallel measure of X, denoted $X' = T + E'$. This definition has the advantage of being defined in terms of observable quantities, namely X and X'. However, part of the definition of parallel measures is that both measures have the same reliability which leads to a circular definition. An alternate way to define reliability that is used by Lord and Novick (1968) is the squared correlation between *X* and *T*. This eliminates the circularity of the previous definition, and it depends on only a single test, but it involves the unobservable variable *T*. The latter definition is preferred in this paper as it offers a stronger rationale for methods of determining reliability based on a single test administration. However, coefficients derived form both definitions will be considered. Both definitions utilize the concept of a linear correlation in terms

of how well a test can predict itself, either its true self or a parallel version of itself, and so may be thought of as coefficients of internal linear predictiveness.

Data Source and Test Construction

A random sample of 100,000 examinees who took the same form of the ACT® test battery (ACT, 2007) was selected and treated as the population from which random samples of n = 50, 100, 200, 400, 1000, and 2000 were drawn 1500 times. The random sampling was done with replacement. The ACT test battery is composed of five subject tests: English (75 dichotomous items), Mathematics (60 dichotomous items), Reading (40 dichotomous items), Science (40 dichotomous items), and an optional Writing Essay polytomously scored on a scale of 2 to 12. Several dichotomous item content heterogeneous tests were constructed by selecting irregularly spaced samples of items from the first four subject tests, and combining them into a single test. Two content parallel 40-item tests each composed of different dichotomously scored items were constructed by choosing in the irregularly spaced item fashion 14 English items, 10 mathematics items, eight reading items, and eight science items per test. A single 22-item subtest composed of dichotomously scored items was constructed from one of the 40-item tests by choosing in the irregularly spaced item fashion eight English items, six mathematics items, four reading items, and four science items. The tests were constructed such that they could be divided into approximately parallel odd-even split halves. Finally, a single eight dichotomous item test was likewise created except that it was too short to be divided into parallel halves. The irregularly spaced sample of items was selected so that the sampled items represented not only a wide range of content but also a wide range of item difficulties. The 40 dichotomous item difficulties for the first 40 item test have a mean of .73, a standard deviation of .10, and they range from .42 to .89. Heterogeneity of item difficulties results in heterogeneous item variances

and heterogeneous inter-item covariances. Therefore the test items are neither tau-equivalent nor essentially-tau-equivalent.

Next, two content parallel 8-item tests each composed of polytomous items were constructed from the two 40-item tests by summing sets of dichotomous items. The 14 English dichotomous items were divided into two groups of seven items and then summed to create two polytomous English items each taking values from 0 to 7. Likewise, two polytomous mathematics items taking values from 0 to 5, two polytomous reading items taking values from 0 to 4, and two polytomous science items taking values of 0 to 4 were constructed by summing five dichotomous mathematics items, four dichotomous reading items, and four dichotomous science items, respectively. Differences in the number of dichotomous items making up the polytomous items in each content category resulted in the polytomous items having moderately heterogeneous item variances and inter-item covariances. So again the tests were neither tau-equivalent nor essentially-tau-equivalent.

Finally, the covariance matrix for one of the 8-item polytomous tests was used to generate the covariance matrices for two 8-item tests, where one formed a test with MVN distributed item scores and the other formed a test with CSMVN distributed item scores using the method of Odell and Feiveson (1966) as presented by Browne (1968). The compound symmetric covariance matrix was created by taking the average variance and average covariance from the original covariance matrix for the eight polytomous items. Therefore this test satisfied all of the assumptions necessary for alpha to equal reliability and for its sample estimate to have the *F* distribution discussed in the next section (van Zyl, Nuedecker, & Nel, 2000).

Distributions, Transformations, and Models

Let $r_\alpha = \hat{\lambda}_3$ denote the sample alpha coefficient for a test of *m* items administered to *n* examinees and $\rho_\alpha = \lambda_3$ denote the parameter value. Kristof (1963) and Feldt (1965) showed that the ratio $(1-r_\alpha)/(1-\rho_\alpha)$ has a *F* ((*n* − 1)(*m* −1), *n* − 1) distribution when the data meet the assumptions of a random effects (Type II) two-way items by examinees ANOVA. Van Zyl et al. (2000) extended the distribution theory for coefficient alpha as did Kistner and Muller (2004). Note that this *F* distribution also holds for Guttman's $\hat{\lambda}_4$ with *m* = 2. Sedere and Feldt (1977) found that under certain conditions the above *F* distribution held for an analogous ratio based on Guttman's $\hat{\lambda}_2$. For investigative purposes this *F* distribution also is hypothesized for Guttman's $\hat{\lambda}_6$ and $\hat{\rho}_c$. Although the *F* distribution should be more accurate than its normalizing transformation especially for smaller sample sizes than those considered in this paper, the normal transformation is used here for ease of interpretation and comparison among the different models and coefficients.

Bonett (2002) employed an approximate normalizing transformation to develop formulas for the sample size needed to obtain a certain level of power when making statistical inferences about coefficient alpha. See also Romano, Kromrey, and Hibbard (2010) for a comparison of different methods for computing confidence intervals for alpha. Two normalizing transformations are used in this paper. The first, proposed by Fisher (Johnson, Kotz, & Balakrishnan, 1995), is for the *F* distribution. For a random variable, *X*, with a $F\ (\nu_1, \nu_2)$ distribution the transformation $(\ln X)/2$ is approximately normally distributed with mean $=\left(\nu_2^{-1} - \nu_1^{-1}\right)/2$ and variance $=\left(\nu_2^{-1} + \nu_1^{-1}\right)/2$. This transformation is applied to the above *F*

distributed random variable $(1-r_\alpha)/(1-\rho_\alpha)$, as well as analogous ratios for the other variables hypothesized to have *F* distributions. After the transformation is applied the resulting variables are standardized so that they should have *N* (0, 1) distributions given that the original variables have the hypothesized *F* distributions.

To exemplify the use of the first normalizing transformation for statistical inference, the lower and upper bounds for a confidence interval for coefficient alpha are given. Let $\zeta_{(1-\pi/2)}$ denote the $100(1-\pi/2)$ percentile of the standardized normal distribution with a mean of zero and a standard deviation of one. Then the lower and upper bounds of a $100(1-\pi)$ percent confidence interval for ρ_α are given, respectively, by

$$Lower(\rho_\alpha) = 1-(1-r_\alpha)\exp\left[\frac{m-2}{(n-1)(m-1)} - \xi_{(1-\pi/2)}\left(\frac{2m}{(n-1)(m-1)}\right)^{\frac{1}{2}}\right]^{-1} \qquad (1)$$

and

$$Upper(\rho_\alpha) = 1-(1-r_\alpha)\exp\left[\frac{m-2}{(n-1)(m-1)} + \xi_{(1-\pi/2)}\left(\frac{2m}{(n-1)(m-1)}\right)^{\frac{1}{2}}\right]^{-1}. \qquad (2)$$

The second normalizing transformation is Fisher's inverse hyperbolic tangent transformation for the correlation coefficient (Johnson et al., 1995) that here is applied to the sample test-retest correlation coefficient and is given by $\tanh^{-1}(r_{12}) = \ln\left[(1+r_{12})/(1-r_{12})\right]/2$. This transformation has approximate mean equal to $\ln\left[(1+\rho_{12})/(1-\rho_{12})\right]/2$ and approximate variance equal to $1/(n-3)$ and hence is variance stabilizing. Again, the resulting transformed variable is standardized to have an *N* (0, 1) distribution. This transformation is applied to the test-

retest reliability coefficient and it assumes that the two sets of test scores are normally distributed.

A formula for a $100(1-\pi)$ percent confidence interval for ρ_{12} in terms of the hyperbolic tangent and inverse hyperbolic tangent is

$$\tanh\left[\tanh^{-1}(r_{12})-\frac{\xi_{(1-\pi/2)}}{(n-3)^{1/2}}\right]\le\rho_{12}\le\tanh\left[\tanh^{-1}(r_{12})+\frac{\xi_{(1-\pi/2)}}{(n-3)^{1/2}}\right]. \quad (3)$$

Simplified lower and upper bounds for a $100(1-\pi)$ percent confidence interval for ρ_{12} derived from the above are, respectively,

$$Lower(\rho_{12})=\left[\frac{1+r_{12}}{1-r_{12}}-\exp\left(\frac{2\xi_{(1-\pi/2)}}{(n-3)^{1/2}}\right)\right]\left[\frac{1+r_{12}}{1-r_{12}}+\exp\left(\frac{2\xi_{(1-\pi/2)}}{(n-3)^{1/2}}\right)\right]^{-1} \quad (4)$$

and

$$Upper(\rho_{12})=\left[\exp\left(\frac{2\xi_{(1-\pi/2)}}{(n-3)^{1/2}}\right)-\frac{1-r_{12}}{1+r_{12}}\right]\left[\exp\left(\frac{2\xi_{(1-\pi/2)}}{(n-3)^{1/2}}\right)+\frac{1-r_{12}}{1+r_{12}}\right]^{-1}. \quad (5)$$

Six test models are investigated and their abbreviations along with brief descriptions are as follows:

(a) CSMVN8 denotes the eight item test consisting of eight MVN distributed items with compound symmetric covariance matrix.

(b) MVN8 denotes the eight item tests with eight MVN distributed items with heterogeneous covariance matrix.

(c) POLY8 denotes the eight item tests with polytomous items constructed from summing varying numbers of dichotomous items.

(d) DICH8 denotes the eight dichotomous items test.

(e) DICH22 denotes the 22 dichotomous items test.

(f) DICH40 denotes the 40 dichotomous items test.

Except for the first model, which serves as a baseline for comparison to the other models, all the models display some violation of alpha's distributional assumptions along with the assumption of essential-tau-equivalence. As mentioned previously, the dichotomous items used in models (d), (e), and (f) have varying difficulty values and hence varying variances and covariances because of the way in which the spaced sample of items was selected. The population covariance matrix for model (c) that is also used for generating the data for Models (a) and (b) is:

$$\Sigma = \begin{bmatrix} 2.18 & 1.28 & 0.67 & 0.93 & 0.55 & 0.52 & 0.59 & 0.61 \\ 1.28 & 2.33 & 0.69 & 0.98 & 0.63 & 0.58 & 0.65 & 0.65 \\ 0.67 & 0.69 & 1.35 & 0.84 & 0.36 & 0.33 & 0.47 & 0.49 \\ 0.93 & 0.98 & 0.84 & 1.90 & 0.52 & 0.48 & 0.68 & 0.70 \\ 0.55 & 0.63 & 0.36 & 0.52 & 1.10 & 0.46 & 0.40 & 0.41 \\ 0.52 & 0.58 & 0.33 & 0.48 & 0.46 & 1.05 & 0.39 & 0.38 \\ 0.59 & 0.65 & 0.47 & 0.68 & 0.40 & 0.39 & 1.11 & 0.53 \\ 0.61 & 0.65 & 0.49 & 0.70 & 0.41 & 0.38 & 0.53 & 1.23 \end{bmatrix}.$$

As can be seen the variances and covariances are moderately heterogeneous with a largest to smallest variance ratio of about two and a similar covariance ratio of almost four. Clearly the items are not essentially tau equivalent. Of course, to compute the test-retest reliability for the various test models, covariance matrices with dimension twice as large as the number of items is needed, but they are not shown for space considerations.

Results

In the text, tables (Appendix A), and figures (Appendix B) that follow Guttmans' lambda coefficients will be denoted L2, L3, L4, and L6. Their standardized normally transformed counterparts sometimes will be denoted Z2, Z3, Z4, and Z6. Test-retest reliability sometimes will

be denoted R12 and its standardized normal transform Z12. Congeneric reliability and its standardized normal transform may be denoted R Congeneric and Z Congeneric, respectively.

Table A1 displays population values of the reliability coefficients along with means computed over the 1500 simulations for the various sample sizes and models. Results are reported to only two decimals for ease of comparison. Finer assessments are made in later tables and figures. Not all reliability coefficients are computed for all models for reasons of appropriateness or redundancy. The population values of all the coefficients for all the models are relatively homogeneous although alpha and L6 tend to be a bit smaller than L4 and R12 which generally have the largest population values for all the models though not by much.

Bias for the sample means is large for L6 in the test models with dichotomous items. In those models it is clearly positively biased for the smaller sample sizes. It also is slightly positively biased for the other models at the smallest sample size. L2 also shows a small amount of bias for the two shorter dichotomous item test models. All the other coefficients appear relatively unbiased for all models and all sample sizes even though their sampling distributions are, in general, negatively skewed.

Table A2 presents the sampling standard errors for the six test models and six sample sizes. L4 tends to have the largest standard error at the smaller sample sizes which is not surprising because its distribution has the smallest degrees of freedom. Otherwise, the different coefficients tend to have similar standard errors that, as expected, decrease as sample size increases and the number of items increases.

Tables A3 through A8 present statistics that give a concise summary of how well the normalized reliability statistics fit an $N(0, 1)$ distribution for the various test models and sample sizes. The tables contain the means with 95% confidence interval bounds, the standard deviations

with 95% confidence interval bounds, and the Shapiro-Wilk (1965) statistics (W) for testing normality along with its p-value. Values of W range from 0 to 1 with smaller values indicating that normality should be rejected. The statistics in these six tables are based on simulated samples of 1500 observations and therefore are quite sensitive to small differences. A statistical significance level of .05 is loosely used for statistical significance and numbers in bold denote rejection of the null hypothesis at this level, but statistical significance does not necessarily imply practical significance. QQ plots for a subset of conditions are presented in Figures B1 through B6. Each plot contains the theoretical normal quantile line along with the observed quantile points for quantiles between −4.0 and 4.0. The plots can be compared to their corresponding statistics to give an alternative, more practical assessment of the normalized sampling distributions fit to an N (0, 1) distribution. The sample sizes chosen for the figures are the ones where a good fit to an N (0, 1) distribution begins to occur for at least some of the models. The fit generally improves for larger sample sizes as can be seen from the corresponding tables.

Figure B1 presents sample size 100 results for Z2 under all six test models. The plots for the three non-dichotomous item test models show good fits to N (0, 1) distributions though Z2 is slightly under-estimated due to L2 being slightly over-estimated. For the three test models with dichotomous items, the fit to an N (0, 1) distribution is not as good with greater under-estimation of Z2 due to the larger positive bias of L2, and the standard deviation of Z2 is over-estimated for the DICH8 model as can been seen in Table A3. The results in Table A3 for larger sample sizes show that the fit improves, but there remains some under-estimation especially for the dichotomous item models.

Figure B2 shows the Z3 results for sample size 50. Except for some discrepancies in the tails the fit is generally good except for the DICH8 model where the lack of parallelism between the empirical and theoretical quantile lines indicates the standard deviation of Z3 is too large. Table A4 shows that the fit generally improves as the sample size increases. The Z4 plots for a sample size of 50 are presented in Figure B3 where the empirical quantiles show very close agreement with the N (0, 1) quantiles except for a small amount of disparity in the tails. The fit remains excellent for larger sample sizes as shown in Table A5.

The fit for Z6 is shown in Figure B4 for a sample size of 2000. From Table A1 and Table A6, it is obvious that $\hat{\lambda}_6$ is over estimating λ_6 even at this large sample size. However, the transformed and standardized distribution of $\hat{\lambda}_6$ does appear to be normally distributed for sample size 2000 and smaller sample sizes as can be seen from Table A6. Results for the test-retest coefficient are given in Table A7 and for sample size 50 in Figure B5. The plots for models MVN8 and DICH40 indicate good fit except at the lower tails, but the fit for POLY8 is not as good. The fit improves as the sample size increases as can be seen in Table A7. The transformed congeneric reliability coefficient does not quite fit the normal distribution for sample size 100 as shown in Figure B6, but the fit improves as sample size increases as shown in Table A8.

Discussion and Conclusions

The results indicate that although alpha is a lower bound to reliability except under essential-tau-equivalence, it is robust to violations of this condition, and its values are competitive with other coefficients' values based on splitting a test into parallel halves or repeating a test to estimate test-retest reliability. It is not always possible to split tests into parallel-half tests, and obtaining test-retest data is often difficult in practice; therefore, having

reliability coefficients that are free of these constraints is useful. The distribution of alpha also is robust to violations of its assumptions at least for the sample sizes and test models considered in this paper, so confidence intervals based on the normal distribution theory can be computed. For small sample sizes or tests with few items, especially dichotomous items, statistical inference can be based on the *F* distribution theory that has been extensively developed by Feldt (Feldt, Woodruff, & Salih, 1987) for just such situations. It should be more accurate than the normalizing transformation used here. However, results, not reported here, obtained with a test constructed from one polytomous item with many categories and many dichotomous items, show that if item variances and inter-item covariances differ dramatically in scale with largest to smallest ratios greater than 10, then alpha can perform poorly. In this situation where many items with small variances and covariances are combined with one item with very large variance, alpha does underestimate reliability. An alternative in this situation may be an estimate of reliability based on a congeneric model even though larger sample sizes are needed for normal theory based inference.

This paper used factor analysis to compute congeneric reliability for its polytomous item data models, and factor analysis of dichotomous data can be problematic in some situations. However, Gilmer and Feldt (1983) and Feldt (2002) present an alternative, computationally simpler method for computing congeneric reliability that avoids having to do factor analysis. In addition, Feldt (2002) presents an interesting formula comparing coefficient alpha to congeneric reliability. This can be expressed by the inequality

$$\rho_c = \frac{m}{m-1-m^2\sigma_\gamma^2}\left(1-\frac{\sum_{i=1}^{m}\sigma_{x_i}^2}{\sigma_X^2}\right) > \frac{m}{m-1}\left(1-\frac{\sum_{i=1}^{m}\sigma_{x_i}^2}{\sigma_X^2}\right) = \rho_\alpha$$

that holds whenever σ_γ^2, the variance of the factor loadings, is greater than zero. This inequality is not based on the usual factor analysis constraint that the true score variance is one, but rather on the constraint that the factor loadings sum to one which is the case in Feldt's method for computing congeneric reliability. From this the ratio of alpha to congeneric reliability is approximately equal to $1 - m\sigma_\gamma^2$ when *m* is large. However, when *m* is large and the γ's are all positive σ_γ^2 will tend to be small because the γ's sum to one, and hence the ratio will approach unity.

Another alternative to alpha is L2, which is especially suited to situations where there are negative inter-item covariances. For larger sample sizes ($n > 100$) and polytomous items, L2 is approximately normally distributed, but for tests composed of many dichotomous items, it is less accurate until the sample size becomes very large ($n > 2000$). If confidence intervals are desired when items are dichotomous, then Feldt's *F* distribution theory may be more accurate.

If a test can be divided into parallel-half tests then L4 is highly recommended. Its transformed sampling distribution is very robust to violations of its assumptions. If it is possible to administer two parallel tests to the same sample of examinees, preferably in a counter-balanced design, then test-retest reliability can be assessed. The transformed test-retest reliability coefficient also has a sampling distribution robust to violations of its assumptions, but its values were not appreciably larger than the values of the other coefficients included in this study. Therefore, the added expense of developing an additional form of the test and the additional examinee testing time required may not be worthwhile.

Finally, L6 is not generally recommended unless the sample size is very large and the items are not dichotomous. Although its transformed value does approach normality, its sample estimate tends to be positively biased, especially for tests composed of many dichotomous items.

References

ACT (2007). *The ACT technical manual*. Iowa City, Iowa: Author.

Bentler, P. M. (2009). Alpha, dimension-free, and model-based internal consistency reliability. *Psychometrika*, *74*, 137-143.

Bonett, D. G. (2002). Sample size requirements for testing and estimating coefficient alpha. *Journal of Educational and Behavioral Statistics*, *27*, 353-340.

Browne, M. W. (1968). A comparison of factor analytic techniques. *Psychometrika, 33*, 267-334.

du Toit, M. (2003). *IRT from SSI: BILOG-MG, MULTILOG, PARSCALE, TESTFACT*. Lincolnwood, IL: Scientific Software International.

Feldt, L. S. (1965). The approximate sampling distribution of Kuder-Richardson reliability coefficient twenty. *Psychometrika*, *30*, 357-370.

Feldt, L. S. (2002). Estimating the internal consistency reliability of tests composed of testlets varying in length. *Applied Measurement in Education*, *15*, 33-48.

Feldt, L. S., & Qualls, A. L. (1996). Bias in coefficient alpha arising from heterogeneity of test content. *Applied Measurement in Education*, *9*, 277-286.

Feldt, L. S., Woodruff, D. J., & Salih, F. A. (1987). Statistical inference for coefficient alpha. *Applied Psychological Measurement*, *11*, 93-103.

Gilmer, J. S., & Feldt, L. S. (1983). Reliability estimation for a test with parts of unknown lengths. *Psychometrika*, *48*, 99-111.

Green, S. B., & Yang, Y. (2009a). Commentary on coefficient alpha: A cautionary tale. *Psychometrika*, *74*, 121-135.

Green, S. B., & Yang, Y. (2009b). Reliability of summed item scores using structural equation modeling: An alternative to coefficient alpha. *Psychometrika*, *74*, 155-167.

Green, S. B., Lissitz, R. W., & Mulaik, S. A. (1977). Limitations of coefficient alpha as an index of test unidimensionality. *Educational and Psychological Measurement*, *37*, 827-838.

Gulliksen, H. (1950). *Theory of mental tests*. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika*, *10*, 255-282.

Hattie, J. A. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement*, *9*, 139-164.

Johnson, N. L., Kotz, S., & Balakrishnan, N. (1995). *Continuous univariate distributions, volume 2* (2nd ed.). New York: John Wiley & Sons.

Jöreskog, K. G. (1971). Statistical analysis of congeneric sets of tests. *Psychometrika*, *36*, 109-133.

Kistner, E. O., & Muller, K. E. (2004). Exact distribution of interclass correlation and Cronbach's alpha with Gaussian data and general covariance. *Psychometrika*, *69*, 459-474.

Kristof, W. (1963). The statistical theory of stepped-up reliability coefficients when a test has been divided into several equivalent parts. *Psychometrika*, *28*, 221-238.

Lord, F. M., & Novick, M. R. (1968). *Statistical theory of mental tests*. Reading, MA: Addison-Wesley.

Odell, P. L., & Feiveson, A. H. (1966). A numerical procedure to generate a sample covariance matrix. *Journal of the American Statistical Association*, *61*, 199-203.

Raykov, T. (1997). Estimation of composite reliability for congeneric measures. *Applied Psychological Measurement*, *21*, 173-184.

Revelle, W., & Zinbarg, R. E. (2009). Coefficient alpha, beta, omega, and the GLB: Comments on Sijtsma. *Psychometrika*, *74*, 145-154.

Romano, J. L., Kromrey, J. D., & Hibbard, S. T. (2010). A Monte Carlo study of eight confidence interval methods for coefficient alpha. *Educational and Psychological Measurement*, *70*, 376-393.

SAS Institute Inc. (2008). *SAS® 9.2 Software* [computer software]. Cary, NC: Author.

Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment*, *8*, 350-353.

Sedere, M. U., & Feldt, L. S. (1977). The sampling distribution of the Kristof reliability coefficient, the Feldt coefficient, and Guttman's lambda-2. *Journal of Educational Measurement*, *14*, 53-62.

Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, *52*, 591-611.

Sijtsma, K. (2009a). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika*, *74*, 107-120.

Sijtsma, K. (2009b). Reliability beyond theory and into practice. *Psychometrika*, *74*, 169-173.

SPSS Inc. (2006). *SPSS 14.0 for Windows* [computer software]. Chicago: Author.

van Zyl, J. M., Nuedecker, H., & Nel, D. G. (2000). On the distribution of the maximum likelihood estimator of Cronbach's alpha. *Psychometrika, 65*, 271-280.

Appendix A

Tables A1 – A8

Table A1.

Population Means and Sampling Means for the Six Test Models and Six Sample Sizes

		Population	Results from 1500 random samples					
Test Model	Coefficient	N = 100,000	n = 50	n = 100	n = 200	n = 400	n = 1000	n = 2000
CSMVN8	Lambda 2	.84	.83	.84	.84	.84	.84	.84
	Lambda 3	.84	.83	.83	.84	.84	.84	.84
	Lambda 6	.82	.84	.83	.82	.82	.82	.82
MVN8	Lambda 2	.84	.84	.84	.84	.84	.84	.84
	Lambda 3	.84	.83	.83	.84	.84	.84	.84
	Lambda 4	.87	.87	.87	.87	.87	.87	.87
	Lambda 6	.83	.85	.84	.84	.83	.83	.83
	Test-retest	.86	.86	.86	.86	.86	.86	.86
	Congeneric	.84	.84	.84	.84	.84	.84	.84
POLY8	Lambda 2	.84	.84	.84	.84	.84	.84	.84
	Lambda 3	.84	.83	.83	.84	.84	.84	.84
	Lambda 4	.87	.86	.87	.87	.87	.87	.87
	Lambda 6	.83	.85	.84	.84	.83	.83	.83
	Test-retest	.86	.86	.86	.86	.86	.86	.86
	Congeneric	.84	.84	.84	.84	.84	.84	.84
DICH8	Lambda 2	.54	.56	.55	.54	.54	.54	.54
	Lambda 3	.54	.51	.52	.53	.53	.53	.54
	Lambda 6	.51	.57	.53	.52	.51	.51	.51
DICH22	Lambda 2	.77	.78	.78	.77	.77	.77	.77
	Lambda 3	.77	.76	.76	.77	.77	.77	.77
	Lambda 4	.79	.78	.78	.79	.79	.79	.79
	Lambda 6	.77	.87	.82	.79	.78	.77	.77
DICH40	Lambda 2	.86	.86	.86	.86	.86	.86	.86
	Lambda 3	.86	.85	.85	.86	.86	.86	.86
	Lambda 4	.87	.86	.87	.87	.87	.87	.87
	Lambda 6	.86	.97	.92	.89	.88	.87	.86
	Test-retest	.86	.86	.86	.86	.86	.86	.86

Table A2.

Sampling Standard Errors for the Six Test Models and Six Sample Sizes

		Results from 1500 random samples					
Test Model	Coefficient	$n = 50$	$n = 100$	$n = 200$	$n = 400$	$n = 1000$	$n = 2000$
CSMVN8	Lambda 2	.036	.025	.017	.012	.008	.006
	Lambda 3	.038	.026	.017	.012	.008	.006
	Lambda 6	.036	.026	.018	.013	.008	.006
MVN8	Lambda 2	.036	.023	.016	.012	.007	.005
	Lambda 3	.038	.024	.017	.012	.007	.006
	Lambda 4	.039	.026	.019	.013	.008	.006
	Lambda 6	.035	.024	.017	.012	.008	.006
	Test-retest	.037	.026	.018	.013	.008	.006
	Congeneric	.039	.026	.017	.012	.008	.005
POLY8	Lambda 2	.037	.024	.018	.012	.008	.006
	Lambda 3	.039	.025	.018	.012	.008	.006
	Lambda 4	.042	.027	.019	.013	.009	.006
	Lambda 6	.035	.025	.018	.013	.008	.006
	Test-retest	.036	.027	.018	.013	.008	.005
	Congeneric	.039	.027	.018	.013	.008	.006
DICH8	Lambda 2	.111	.079	.056	.042	.026	.018
	Lambda 3	.132	.087	.059	.043	.026	.018
	Lambda 6	.113	.081	.057	.043	.026	.018
DICH22	Lambda 2	.047	.033	.024	.016	.011	.007
	Lambda 3	.054	.036	.025	.017	.011	.008
	Lambda 4	.067	.045	.032	.022	.013	.010
	Lambda 6	.031	.028	.022	.016	.011	.007
DICH40	Lambda 2	.028	.020	.015	.011	.006	.005
	Lambda 3	.032	.022	.016	.011	.006	.005
	Lambda 4	.041	.028	.019	.013	.009	.006
	Lambda 6	.007	.013	.012	.010	.006	.005
	Test-retest	.038	.026	.018	.013	.008	.006

Table A3.

Fit Statistics for Z2: Means, SD's, CI's and Normality Test Statistics

Sample Size	Model	*Mean*[a]	95 % *CI*		*SD*[b]	95 % *CI*		*W*[c]	*p−value*
n = 50	CSMVN8	**−.112**	−.162 ,	−.061	.996	.962 ,	1.033	.9986	.2479
	MVN8	**−.134**	−.185 ,	−.083	1.002	.967 ,	1.039	**.9968**	.0038
	POLY8	**−.124**	−.176 ,	−.072	1.033	.998 ,	1.072	**.9978**	.0400
	DICH8	**−.384**	−.440 ,	−.328	**1.102**	1.064 ,	1.143	**.9959**	.0005
	DICH22	**−.308**	−.358 ,	−.258	.984	.950 ,	1.021	.9981	.0872
	DICH40	**−.377**	−.427 ,	−.327	.989	.955 ,	1.026	.9993	.8868
n = 100	CSMVN8	**−.115**	−.166 ,	−.064	1.001	.966 ,	1.038	**.9977**	.0318
	MVN8	**−.064**	−.113 ,	−.015	.968	.935 ,	1.004	.9991	.7348
	POLY8	**−.120**	−.171 ,	−.069	1.003	.968 ,	1.040	.9986	.2497
	DICH8	**−.250**	−.306 ,	−.193	**1.116**	1.077 ,	1.157	**.9979**	.0463
	DICH22	**−.255**	−.305 ,	−.205	.990	.956 ,	1.027	**.9971**	.0077
	DICH40	**−.277**	−.327 ,	−.226	.995	.961 ,	1.032	**.9940**	.0000
n = 200	CSMVN8	**−.079**	−.128 ,	−.031	**.959**	.926 ,	.995	.9983	.1285
	MVN8	**−.071**	−.120 ,	−.023	.966	.933 ,	1.002	.9984	.1594
	POLY8	**−.099**	−.152 ,	−.047	**1.038**	1.002 ,	1.077	.9981	.0788
	DICH8	**−.205**	−.263 ,	−.148	**1.143**	1.103 ,	1.185	.9985	.2121
	DICH22	**−.166**	−.217 ,	−.114	1.014	.979 ,	1.052	.9988	.4175
	DICH40	**−.177**	−.229 ,	−.125	1.032	.996 ,	1.070	.9988	.3819
n = 400	CSMVN8	**−.107**	−.157 ,	−.057	.990	.955 ,	1.026	.9984	.1775
	MVN8	**−.066**	−.115 ,	−.016	.982	.948 ,	1.018	.9985	.1960
	POLY8	**−.086**	−.138 ,	−.034	1.024	.989 ,	1.063	**.9977**	.0278
	DICH8	**−.133**	−.193 ,	−.073	**1.193**	1.151 ,	1.237	**.9974**	.0154
	DICH22	**−.184**	−.235 ,	−.134	.991	.957 ,	1.028	.9984	.1678
	DICH40	**−.170**	−.224 ,	−.117	**1.049**	1.013 ,	1.088	.9979	.0523
n = 1000	CSMVN8	**−.065**	−.116 ,	−.015	.998	.963 ,	1.035	.9991	.6524
	MVN8	**−.066**	−.116 ,	−.017	.974	.940 ,	1.010	.9991	.6609
	POLY8	**−.080**	−.132 ,	−.027	1.034	.998 ,	1.072	.9993	.8938
	DICH8	−.037	−.096 ,	.021	**1.151**	1.111 ,	1.194	.9990	.6034
	DICH22	**−.124**	−.177 ,	−.072	1.035	.999 ,	1.073	.9991	.6909
	DICH40	**−.077**	−.127 ,	−.027	.991	.957 ,	1.028	.9987	.3220
n = 2000	CSMVN8	−.042	−.094 ,	.009	1.013	.978 ,	1.051	.9991	.7324
	MVN8	−.008	−.059 ,	.044	1.024	.988 ,	1.062	**.9971**	.0074
	POLY8	−.043	−.096 ,	.010	**1.047**	1.011 ,	1.086	.9986	.2862
	DICH8	−.054	−.111 ,	.003	**1.126**	1.087 ,	1.168	.9982	.0946
	DICH22	**−.085**	−.136 ,	−.035	1.004	.969 ,	1.041	.9989	.4870
	DICH40	**−.060**	−.111 ,	−.008	1.014	.979 ,	1.052	.9990	.5962

[a] Values in bold are significantly different from 0 at the .05 level.
[b] Values in bold are significantly different from 1 at the .05 level.
[c] The Shapiro-Wilk statistic as computed by SAS for sample size ≤ 2000

Table A4.

Fit statistics for Z3: Means, SD's, CI's and normality test statistics

Sample Size	Model	*Mean*[a]	95 % *CI*		*SD*[b]	95 % *CI*		*W*[c]	*p−value*
n = 50	CSMVN8	.033	−.018,	.085	1.014	.979 ,	1.052	.9985	.2004
	MVN8	.003	−.047,	.053	.995	.961 ,	1.032	**.9954**	.0002
	POLY8	.029	−.023,	.081	1.025	.990 ,	1.063	**.9972**	.0091
	DICH8	−.018	−.078,	.042	**1.181**	1.140 ,	1.224	**.9944**	.0000
	DICH22	**.063**	.011,	.116	1.035	.999 ,	1.074	**.9978**	.0419
	DICH40	.007	−.046,	.059	1.033	.998 ,	1.072	.9993	.8826
n = 100	CSMVN8	−.015	−.066,	.036	1.010	.975 ,	1.048	**.9975**	.0207
	MVN8	.028	−.020,	.076	**.948**	.916 ,	.983	.9992	.8058
	POLY8	−.016	−.066,	.034	.979	.946 ,	1.016	.9984	.1523
	DICH8	.027	−.032,	.087	**1.174**	1.133 ,	1.217	**.9972**	.0087
	DICH22	.023	−.028,	.075	1.025	.990 ,	1.064	**.9969**	.0042
	DICH40	.013	−.039,	.064	1.025	.990 ,	1.063	**.9938**	.0000
n = 200	CSMVN8	−.009	−.058,	.039	**.963**	.930 ,	.999	.9983	.1224
	MVN8	−.012	−.060,	.036	**.949**	.917 ,	.985	.9984	.1812
	POLY8	−.026	−.078,	.025	1.015	.980 ,	1.053	.9981	.0927
	DICH8	−.002	−.062,	.057	**1.178**	1.137 ,	1.222	.9987	.3407
	DICH22	.037	−.016,	.089	**1.037**	1.002 ,	1.076	.9988	.3846
	DICH40	.036	−.017,	.089	**1.051**	1.015 ,	1.090	.9987	.3635
n = 400	CSMVN8	**−.058**	−.109,	−.008	.992	.958 ,	1.029	.9984	.1911
	MVN8	−.016	−.065,	.033	.968	.934 ,	1.004	.9982	.1011
	POLY8	−.031	−.082,	.020	1.002	.967 ,	1.039	**.9976**	.0232
	DICH8	.016	−.045,	.078	**1.215**	1.173 ,	1.260	**.9972**	.0090
	DICH22	−.039	−.090,	.012	1.005	.971 ,	1.043	.9983	.1488
	DICH40	−.017	−.071,	.037	**1.062**	1.025 ,	1.101	.9980	.0587
n = 1000	CSMVN8	−.034	−.084,	.017	.998	.964 ,	1.035	.9990	.6284
	MVN8	−.038	−.086,	.010	**.953**	.920 ,	.988	.9992	.7509
	POLY8	−.046	−.097,	.005	1.007	.973 ,	1.045	.9993	.8795
	DICH8	.056	−.003,	.115	**1.165**	1.124 ,	1.208	.9989	.5011
	DICH22	−.030	−.083,	.023	**1.044**	1.008 ,	1.083	.9990	.6123
	DICH40	.022	−.029,	.072	.998	.963 ,	1.035	.9987	.3145
n = 2000	CSMVN8	−.020	−.072,	.031	1.013	.978 ,	1.051	.9991	.7272
	MVN8	.008	−.042,	.059	1.002	.968 ,	1.040	**.9969**	.0041
	POLY8	−.023	−.074,	.028	1.014	.979 ,	1.052	.9985	.1959
	DICH8	.013	−.044,	.071	**1.135**	1.096 ,	1.177	.9982	.1056
	DICH22	−.020	−.071,	.032	1.012	.977 ,	1.049	.9989	.4901
	DICH40	.011	−.040,	.063	1.018	.983 ,	1.056	.9990	.5798

[a] Values in bold are significantly different from 0 at the .05 level.
[b] Values in bold are significantly different from 1 at the .05 level.
[c] The Shapiro-Wilk statistic as computed by SAS for sample size ≤ 2000

Table A5.

Fit Statistics for Z4: Means, SD's, CI's and Normality Test Statistics

Sample Size	Model	*Mean*[a]	95 % *CI*		*SD*[b]	95 % *CI*		*W*[c]	*p−value*
n = 50	MVN8	−.021	−.072,	.030	1.002	.967 ,	1.039	.9989	.5167
	POLY8	−.011	−.064,	.041	**1.039**	1.003 ,	1.077	.9983	.1209
	DICH22	.049	−.002,	.101	1.019	.984 ,	1.057	.9991	.6437
	DICH40	−.011	−.063,	.041	1.032	.996 ,	1.070	.9980	.0691
n = 100	MVN8	.015	−.034,	.065	.974	.941 ,	1.011	.9990	.5705
	POLY8	−.003	−.054,	.048	1.010	.975 ,	1.048	.9987	.3651
	DICH22	.005	−.046,	.057	1.016	.980 ,	1.053	.9986	.2877
	DICH40	.047	−.005,	.098	1.014	.979 ,	1.052	.9983	.1359
n = 200	MVN8	.002	−.049,	.052	.996	.962 ,	1.033	.9989	.4881
	POLY8	−.023	−.076,	.029	1.033	.997 ,	1.071	.9989	.4789
	DICH22	.012	−.041,	.066	**1.052**	1.015 ,	1.091	**.9974**	.0163
	DICH40	.014	−.038,	.066	1.031	.996 ,	1.070	.9985	.1972
n = 400	MVN8	.012	−.038,	.062	.989	.955 ,	1.026	.9993	.8723
	POLY8	−.034	−.085,	.017	1.007	.972 ,	1.044	.9989	.4829
	DICH22	−.040	−.091,	.011	1.015	.980 ,	1.053	.9987	.3325
	DICH40	−.013	−.065,	.039	1.024	.988 ,	1.062	.9987	.3416
n = 1000	MVN8	−.031	−.081,	.018	.977	.944 ,	1.014	.9983	.1436
	POLY8	.008	−.044,	.061	1.036	1.000 ,	1.074	.9983	.1352
	DICH22	**−.057**	−.108,	−.006	1.009	.974 ,	1.046	.9991	.6788
	DICH40	.023	−.029,	.075	1.024	.988 ,	1.062	.9987	.3373
n = 2000	MVN8	.034	−.017,	.085	1.009	.974 ,	1.046	.9989	.5205
	POLY8	−.029	−.082,	.024	**1.041**	1.005 ,	1.080	.9991	.7329
	DICH22	−.016	−.070,	.037	**1.055**	1.019 ,	1.094	.9991	.6445
	DICH40	**.056**	.005,	.107	1.007	.972 ,	1.044	.9996	.9890

[a] Values in bold are significantly different from 0 at the .05 level.
[b] Values in bold are significantly different from 1 at the .05 level.
[c] The Shapiro-Wilk statistic as computed by SAS for sample size ≤ 2000

Table A6.

Fit statistics for Z6: Means, SD's, CI's and Normality Test Statistics

Sample Size	Model	*Mean*[a]	95 % *CI*	*SD*[b]	95 % *CI*	*W*[c]	*p−value*
n = 50	CSMVN8	**−.685**	−.737, −.634	1.009	.975 , 1.047	.9988	.3775
	MVN8	**−.715**	−.767, −.664	1.024	.989 , 1.062	**.9972**	.0085
	POLY8	**−.774**	−.827, −.720	**1.064**	1.027 , 1.103	.9980	.0605
	DICH8	**−.852**	−.911, −.793	**1.164**	1.124 , 1.207	**.9960**	.0006
	DICH22	**−2.939**	−2.995, −2.882	**1.116**	1.077 , 1.158	.9979	.0578
	DICH40	**−8.652**	−8.722, −8.582	**1.380**	1.332 , 1.432	**.9968**	.0037
n = 100	CSMVN8	**−.494**	−.544, −.443	.992	.958 , 1.029	.9983	.1320
	MVN8	**−.459**	−.509, −.409	.981	.947 , 1.018	.9991	.7032
	POLY8	**−.550**	−.602, −.499	1.012	.977 , 1.050	.9988	.4159
	DICH8	**−.526**	−.583, −.470	**1.120**	1.081 , 1.161	**.9976**	.0242
	DICH22	**−1.804**	−1.857, −1.751	**1.048**	1.012 , 1.087	**.9962**	.0009
	DICH40	**−3.907**	−3.962, −3.852	**1.086**	1.049 , 1.126	**.9944**	.0000
n = 200	CSMVN8	**−.341**	−.389, −.294	**.945**	.912 , .980	.9981	.0773
	MVN8	**−.335**	−.384, −.285	.976	.942 , 1.012	.9983	.1458
	POLY8	**−.402**	−.455, −.349	**1.045**	1.009 , 1.084	.9982	.0942
	DICH8	**−.389**	−.445, −.332	**1.118**	1.080 , 1.160	.9986	.2462
	DICH22	**−1.167**	−1.220, −1.114	**1.043**	1.007 , 1.082	.9987	.3227
	DICH40	**−2.393**	−2.447, −2.339	**1.072**	1.035 , 1.112	.9985	.2059
n = 400	CSMVN8	**−.288**	−.338, −.239	.973	.940 , 1.009	.9986	.2552
	MVN8	**−.255**	−.304, −.205	.975	.942 , 1.012	.9984	.1827
	POLY8	**−.298**	−.349, −.246	1.018	.983 , 1.056	**.9971**	.0067
	DICH8	**−.259**	−.317, −.201	**1.149**	1.109 , 1.191	**.9973**	.0127
	DICH22	**−.869**	−.920, −.818	1.001	.967 , 1.038	.9986	.2408
	DICH40	**−1.644**	−1.698, −1.590	**1.068**	1.032 , 1.108	**.9977**	.0287
n = 1000	CSMVN8	**−.179**	−.229, −.130	.977	.943 , 1.013	.9990	.5981
	MVN8	**−.187**	−.236, −.137	.972	.939 , 1.009	.9989	.5419
	POLY8	**−.207**	−.259, −.155	1.025	.989 , 1.063	.9993	.8943
	DICH8	**−.115**	−.171, −.060	**1.098**	1.060 , 1.139	.9991	.6585
	DICH22	**−.548**	−.601, −.495	**1.043**	1.007 , 1.081	.9992	.8000
	DICH40	**−.976**	−1.027, −.926	1.001	.967 , 1.039	.9988	.3723
n = 2000	CSMVN8	**−.122**	−.172, −.072	.991	.957 , 1.028	.9991	.7243
	MVN8	**−.087**	−.139, −.035	1.018	.983 , 1.056	**.9974**	.0167
	POLY8	**−.134**	−.187, −.081	**1.044**	1.008 , 1.083	.9991	.6464
	DICH8	**−.107**	−.161, −.053	**1.067**	1.030 , 1.106	.9980	.0722
	DICH22	**−.380**	−.431, −.329	1.003	.969 , 1.041	.9988	.3990
	DICH40	**−.686**	−.738, −.634	1.020	.984 , 1.058	.9990	.5921

[a] Values in bold are significantly different from 0 at the .05 level.
[b] Values in bold are significantly different from 1 at the .05 level.
[c] The Shapiro-Wilk statistic as computed by SAS for sample size ≤ 2000

Table A7.

Fit statistics for Test-Retest Reliability: Means, SD's, CI's and Normality Test Statistics

Sample Size	Model	*Mean*[a]	95 % *CI*		*SD*[b]	95 % *CI*		*W*[c]	*p−value*
	MVN8	**.071**	.021,	.121	.981	.947 ,	1.017	.9988	.4080
n = 50	POLY8	**.077**	.029,	.126	**.956**	.923 ,	.992	.9987	.3533
	DICH40	**.086**	.035,	.137	1.003	.968 ,	1.040	.9990	.5836
	MVN8	**.057**	.007,	.107	.990	.956 ,	1.027	.9988	.4327
n = 100	POLY8	.008	−.044,	.059	1.016	.981 ,	1.054	.9990	.5809
	DICH40	.040	−.010,	.091	.998	.963 ,	1.035	.9991	.7352
	MVN8	.010	−.040,	.061	.993	.959 ,	1.030	.9989	.4702
n = 200	POLY8	**.069**	.019,	.120	.991	.957 ,	1.028	.9988	.3810
	DICH40	.002	−.048,	.053	.995	.961 ,	1.032	.9988	.3796
	MVN8	**.060**	.009,	.110	.993	.959 ,	1.030	.9990	.5485
n = 400	POLY8	**.053**	.002,	.103	.990	.956 ,	1.027	.9991	.6566
	DICH40	.042	−.007,	.091	.971	.937 ,	1.007	.9983	.1355
	MVN8	−.014	−.066,	.037	1.014	.979 ,	1.052	.9985	.2091
n = 1000	POLY8	.017	−.034,	.069	1.021	.986 ,	1.059	.9987	.3257
	DICH40	−.029	−.079,	.021	.986	.952 ,	1.023	.9993	.8850
	MVN8	−.016	−.068,	.036	1.022	.987 ,	1.060	.9990	.5765
n = 2000	POLY8	.008	−.042,	.058	.995	.960 ,	1.032	.9986	.2865
	DICH40	.034	−.017,	.084	.994	.960 ,	1.031	.9987	.3399

[a] Values in bold are significantly different from 0 at the .05 level.
[b] Values in bold are significantly different from 1 at the .05 level.
[c] The Shapiro-Wilk statistic as computed by SAS for sample size ≤ 2000

Table A8.

Fit statistics for congeneric reliability: Means, SD's, CI's and normality test statistics

Sample Size	Model	*Mean*[a]	95 % *CI*		*SD*[b]	95 % *CI*		*W*[c]	*p−value*
$n = 50$	MVN8	**−.080**	−.134,	−.026	**1.065**	1.028 ,	1.104	**.9950**	.0001
	POLY8	−.019	−.073,	.035	**1.070**	1.033 ,	1.110	**.9965**	.0017
$n = 100$	MVN8	**−.074**	−.127,	−.021	**1.052**	1.015 ,	1.091	.9988	.4310
	POLY8	−.025	−.080,	.029	**1.078**	1.041 ,	1.118	**.9965**	.0016
$n = 200$	MVN8	−.025	−.076,	.027	1.011	.976 ,	1.048	.9985	.1993
	POLY8	−.042	−.097,	.012	**1.077**	1.040 ,	1.117	.9993	.8323
$n = 400$	MVN8	−.005	−.057,	.046	1.020	.985 ,	1.058	.9989	.5030
	POLY8	−.044	−.098,	.010	**1.075**	1.038 ,	1.115	**.9968**	.0032
$n = 1000$	MVN8	−.025	−.076,	.027	1.024	.988 ,	1.062	.9991	.7215
	POLY8	−.050	−.105,	.005	**1.088**	1.050 ,	1.128	.9989	.4653
$n = 2000$	MVN8	−.045	−.096,	.006	1.003	.968 ,	1.040	.9990	.5738
	POLY8	−.032	−.086,	.022	**1.068**	1.031 ,	1.107	.9984	.1794

[a] Values in bold are significantly different from 0 at the .05 level.
[b] Values in bold are significantly different from 1 at the .05 level.
[c] The Shapiro-Wilk statistic as computed by SAS for sample size ≤ 2000

Appendix B

Figures B1 - B6

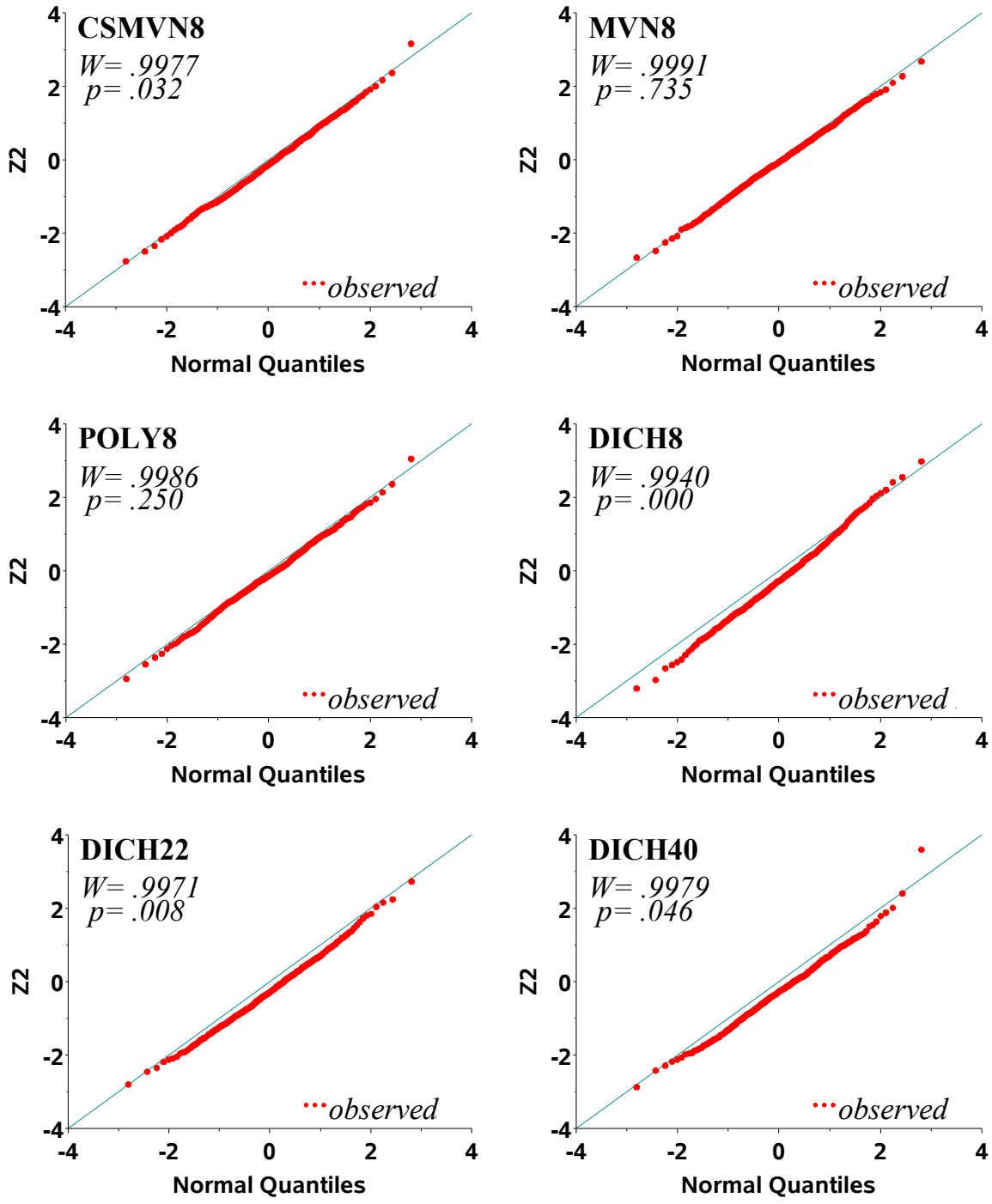


Figure B1. QQ plots of Z2 for sample size 100 under all six models.

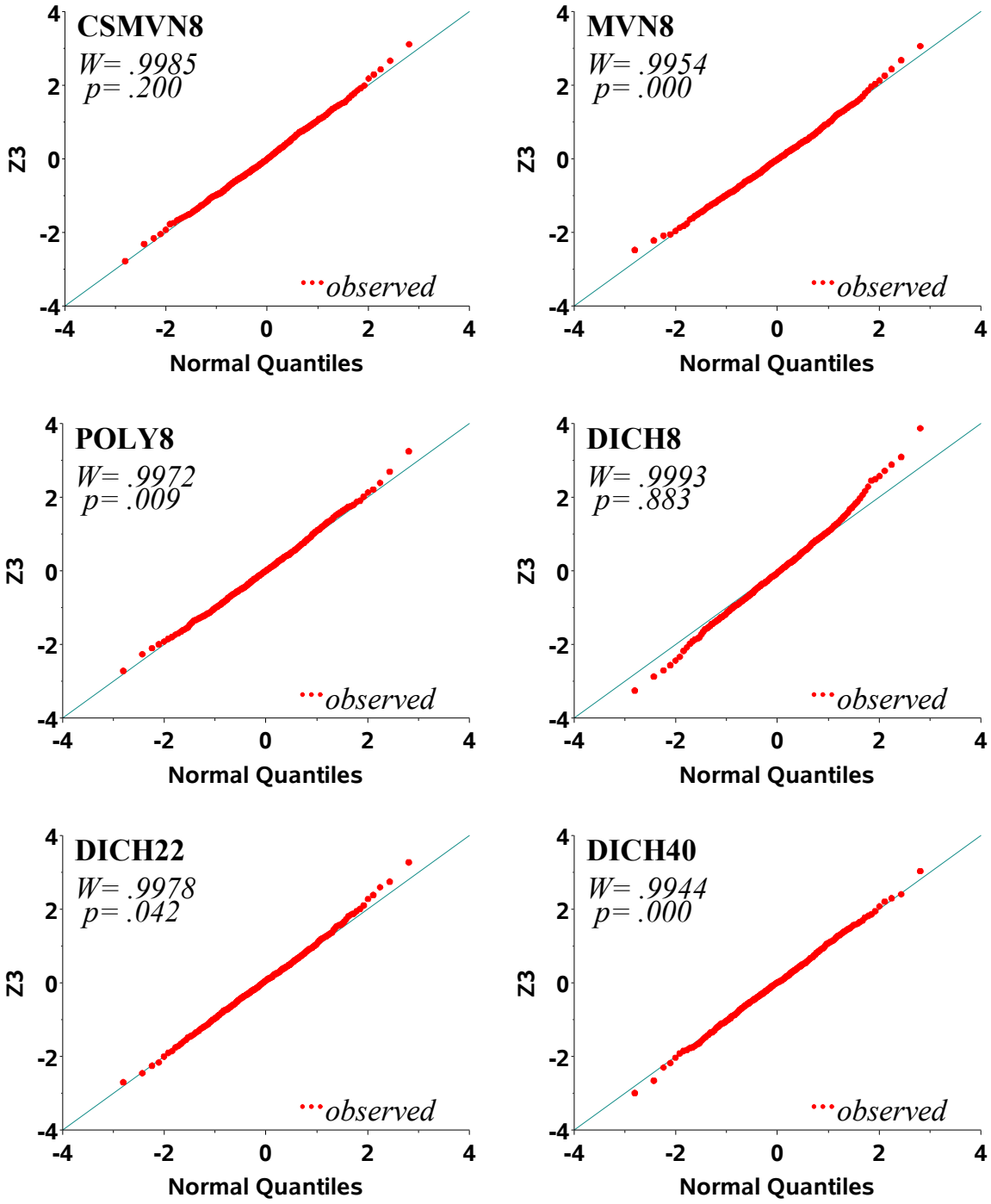


Figure B2. QQ plots of Z3 for sample size 50 under all six models.

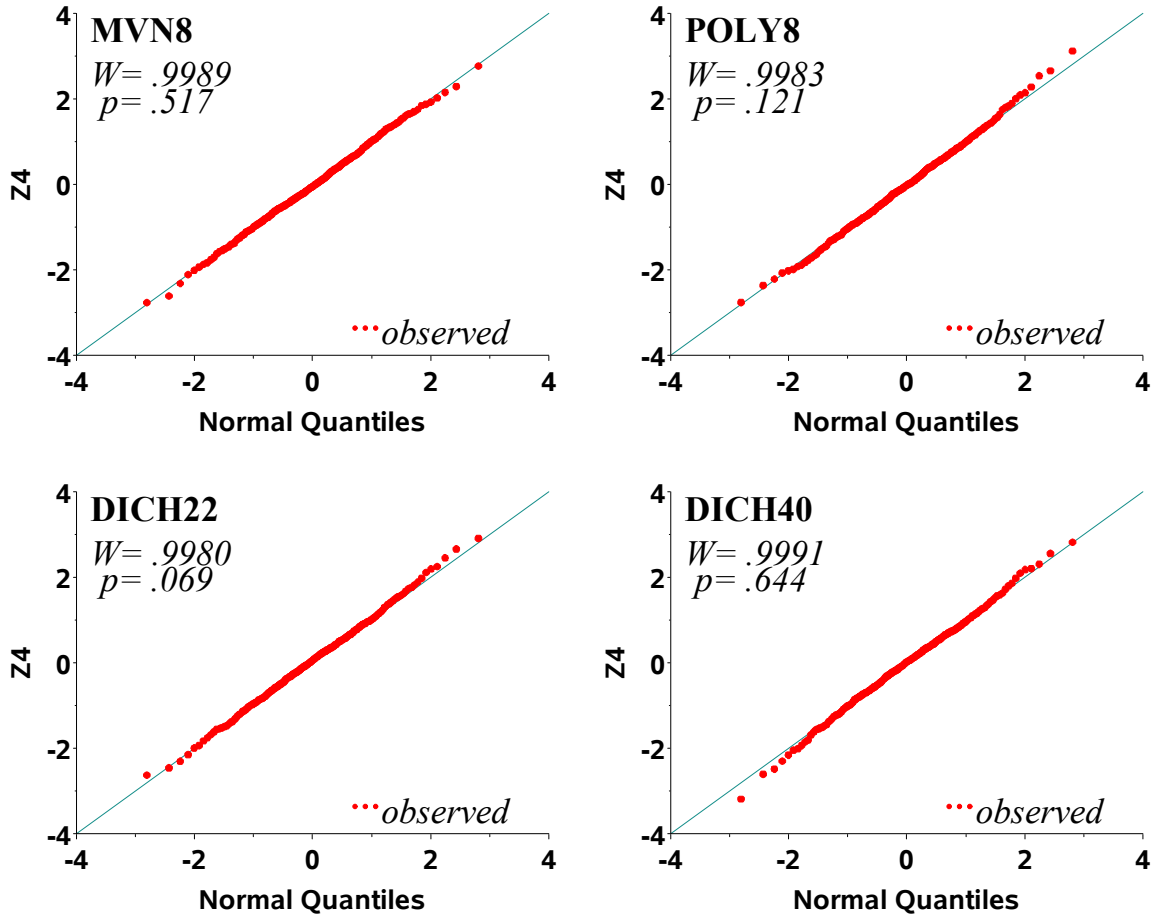


Figure B3. QQ plots of Z4 for sample size 50 under four models.

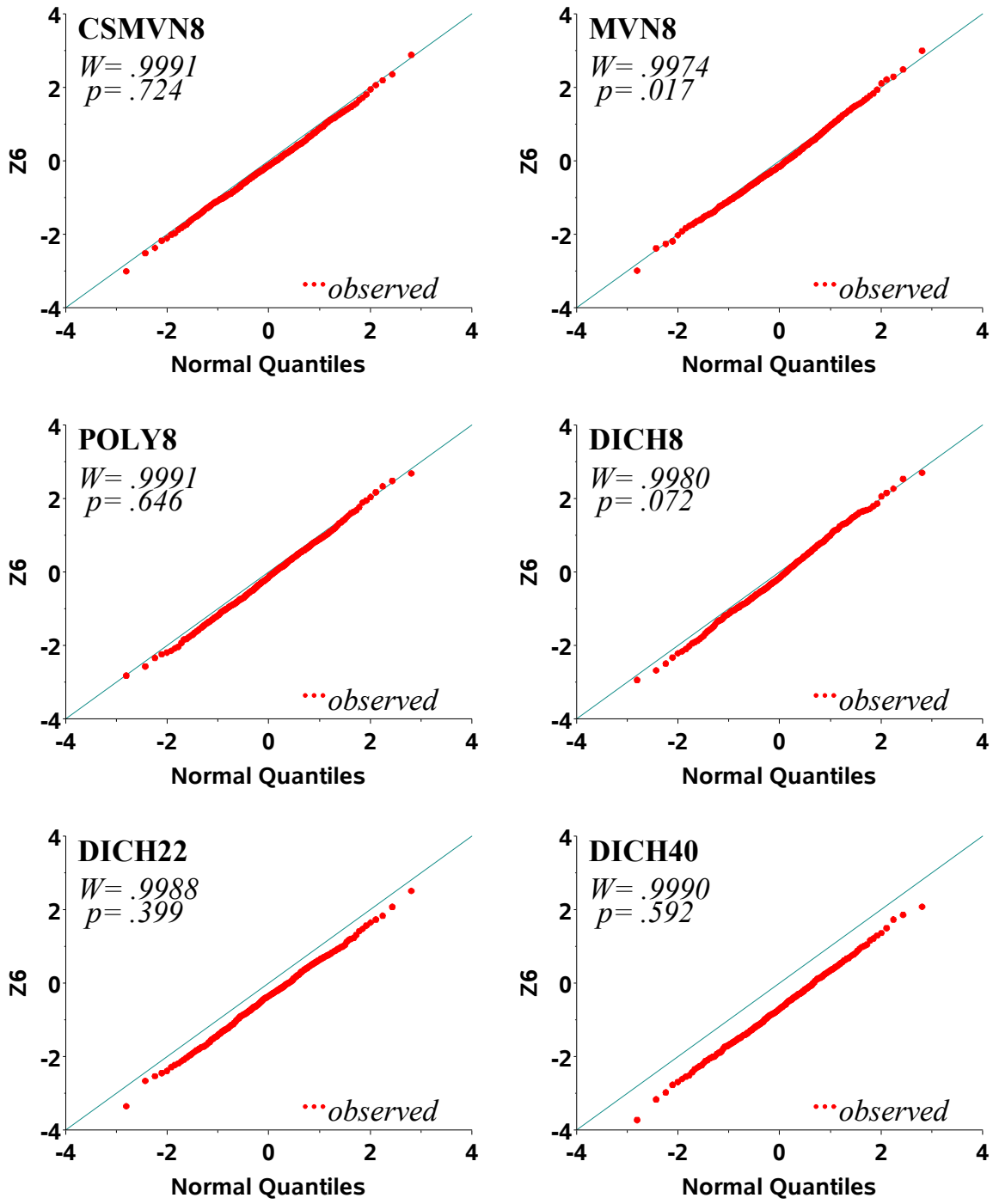


Figure B4. QQ plots of Z6 for sample size 2000 under all six models.

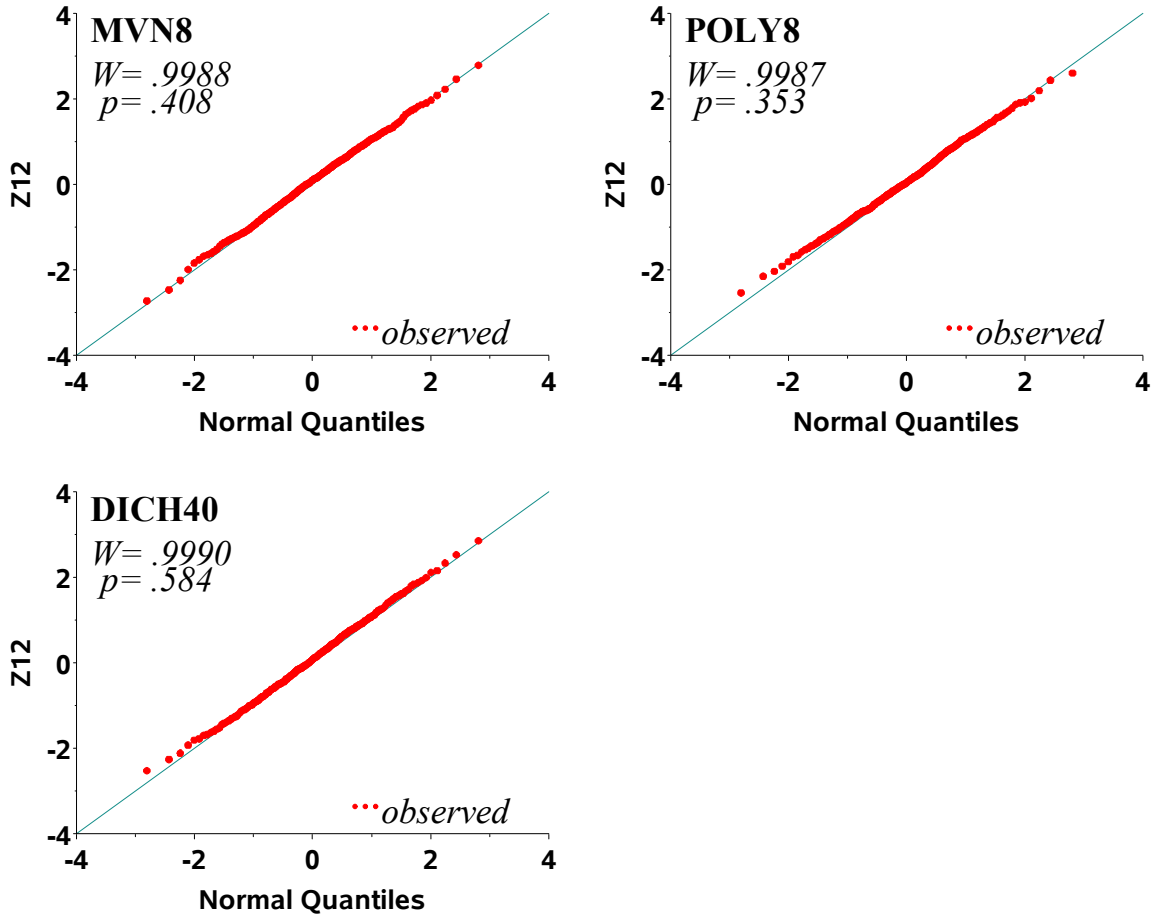


Figure B5. QQ plots of Z12 for sample size 50 under three models.

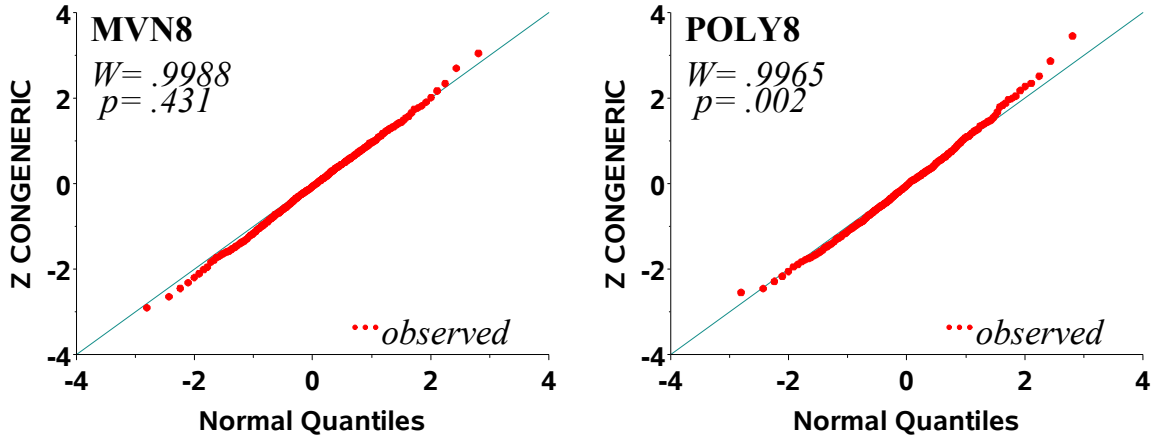


Figure B6. QQ plots of Z Congeneric for sample size 100 under two models.

0502101120 Rev 1