# Chapter 10
# Examination of Test Characteristics' Effect on Coefficient $\alpha$ and Coefficient $\omega$


**Terry Ackerman, Ye Ma, and Richard Luecht**

**Abstract** In this study, five factors were simulated to determine their effect on three measures of reliability: coefficient $\alpha$, coefficient $\omega$, and the true scale reliability as defined in a classical test theory context as the ratio of true score variance over observed score variance. The factors were the number of items, the level of item discrimination, the number of dimensions, the correlations among dimensions, and the location of the items in relationship to the latent ability score distribution. In all higher-order dimensional conditions, simple structure was assumed. The data were generated using the multidimensional item response theory compensatory two-parameter logistic model. As expected, when the number of items, the magnitude of the item discriminations, and the correlations among the dimensions increased, the reliability correspondingly increased. Noticeable differences were observed across all higher dimensionality conditions with $\omega$ values being significantly lower than $\alpha$, a finding which could have been an artifact of the simulated simple structure.

## 10.1 Background

Reliability is one of the hallmark measures of an assessment's quality. It is a necessary condition for validity. Several authors have noted that a test's reliability is a function of the scores on a test, not the test itself or multiple forms of a test (Brennan, 2001; Thompson & Vacha-Haase, 2000). There are a host of measures

T. Ackerman (✉)
University of Iowa, Iowa City, IA, USA
e-mail: terry-ackerman@uiowa.edu

Y. Ma
Amazon Web Services (AWS), Chicago, Illinois, USA
e-mail: ymcheryl@amazon.com

R. Luecht
University of North Carolina at Greensboro, Greensboro, NC, USA
e-mail: rmluecht@uncg.edu

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023

L. A. van der Ark et al. (eds.), *Essays on Contemporary Psychometrics*,
Methodology of Educational Measurement and Assessment,
https://doi.org/10.1007/978-3-031-10370-4_10

that have been developed to estimate reliability (Feldt & Brennan, 1989; Kane, 1996). The basic definition of reliability is based on the classical test theory assumption that for individual $i$, a test score $X_i$ is the sum of two unobservable and uncorrelated components, $T_i$, a true score and measurement error, $E_i$:

$$X_i = T_i + E_i. \tag{10.1}$$

Reliability is then defined as the squared correlation between the observed test scores and the corresponding unobserved true scores which can be shown to be equal to the ratio of true score variance, $\sigma_T^2$, to total observed score variance, $\sigma_X^2$:

$$\rho_{TX}^2 = \frac{\sigma_T^2}{\sigma_X^2} \tag{10.2}$$

As noted by Sijtsma (2009a, b), over the years, the one standard reliability index that researchers and psychologists have adopted is coefficient alpha (Cronbach, 1951), further referred to as $\alpha$. Although Cronbach's name is tied to the statistic, this measure can be traced through the works of Kuder and Richardson (1937), who published a version of $\alpha$ for dichotomous items—the KR-20 coefficient. Hoyt (1941) proposed an equivalent statistic using of analysis of variance with dichotomous responses.

Finally, Guttman (1945) derived a series of reliability coefficients. One coefficient, denoted as $\lambda_3$, was equivalent to $\alpha$.

Assuming a test composed of $J$-items, where a random variable, $Y_j$, represents a score on item $j$, and the total score on the test for an examinee is defined as the sum,

$$X = \sum_{j=1}^{J} Y_j, \tag{10.3}$$

$\alpha$ for a group of examinees can be expressed as

$$\alpha = \frac{J}{(J-1)}\left[1 - \frac{\sum_{j=1}^{J}\sigma_{Y_j}^2}{\sigma_X^2}\right] \tag{10.4}$$

where $\sigma_{Y_j}^2$ represents the item variances and $\sigma_X^2$ is the variance of the total scores.

If the item scores are standardized, the formula for $\alpha$ can be expressed in terms of the mean of the inter-item correlations, $\overline{\rho}$; that is,

$$\alpha = \frac{J\overline{\rho}}{1 + (J-1)\overline{\rho}}, \tag{10.5}$$

or equivalently as the average of the inter-item covariances, $\overline{\sigma_{YY}}$,

$$\alpha = \frac{J\left(\frac{\overline{\sigma_{YY}}}{\overline{\sigma_X}}\right)}{1 + (J-1)\left(\frac{\overline{\sigma_{YY}}}{\overline{\sigma_X}}\right)}. \tag{10.6}$$

It should be noted that $\alpha$ also approximates the mean of all possible Spearman-Brown split-half coefficients (Spearman, 1910; Brown, 1910) where the split-half coefficients, $r_{12}$, are adjusted, pairwise Pearson product-moment correlations between the two half-test scores:

$$r_{\text{split-half(SB)}} = \frac{2r_{12}}{1 + r_{12}}. \tag{10.7}$$

Coefficient $\alpha$ equals the mean of the split-half coefficients when the standard deviations of all possible halves are equal and smaller when the standard deviations are heterogeneous (Cortina, 1993). Feldt and Brennan (1989) and Lord and Novick (1968) further noted that $\alpha$ will be equal to the mean of all split-half correlations when the split-half correlations are calculated by the Flanagan-Rulon formula:

$$r_{\text{split-half(FR)}} = \frac{4r_{12}s_1s_2}{s_T^2}, \tag{10.8}$$

where $s_1$ and $s_2$ are the standard deviations of each half and $s_T^2$ is the variance of the total test (Flanagan, 1937; Rulon, 1939).

Many researchers have criticized the pervasive use of $\alpha$ (Green, et al., 1977; Green and Yang, 2009; Rodriguez & Maeda, 2006; Sijtsma, 2009a, b) or even wrote about the shortcomings of the statistic and its interpretations (Cronbach & Shavelson, 2004; Ten Berge & Socan, 2004). One drawback is the ubiquitous interpretation of $\alpha$ as a measure of internal consistency. Internal consistency is a characteristic of the test items, not the test, and does not reflect the length of the test (i.e., the pattern of inter-item covariances). Another caveat is that calculations of $\alpha$ can yield values that are outside the range of possible values of the score reliability that should be derivable from a single test administration (Cho & Kim, 2015; Sijtsma, 2009a).

It is often thought that $\alpha$ requires the test to be unidimensional and that it can be used as a measure signifying the degree of multidimensionality. Cronbach (1951) did address the test dimensionality issue when he wrote that for a test:

> to be interpretable, . . . it is not essential that all the items be factorially similar. What is required is that a large proportion of the test variance be attributable to the first principal factor running through the test.

Several authors have noted that multidimensional tests can exhibit high values of $\alpha$ (Davenport, et al., 2015; Davison & Davenport, 2015). When a test has been empirically demonstrated to be multidimensional, it is important the test developer

be able to articulate the meaning of the composite scale which $\alpha$ is characterizing (e.g., that the total test score is a weighted linear composite of two or more subscores by design). In any case, it has been well documented that a multidimensional test does not necessarily have a lower $\alpha$ than a unidimensional test.

Friedman and Weisberg (1981) demonstrated that if all the inter-item correlations are positive, the first principal component eigenvalue is approximately proportional to the average correlation of the $J$ items

$$\lambda_1 \approx 1 + (J-1)\,\overline{r}. \tag{10.9}$$

Using this relationship, $\alpha$ can be approximated as

$$\alpha \approx \frac{J\overline{r}}{\lambda_1}. \tag{10.10}$$

Another approach that tries to capture the underlying possibly multidimensional nature is to assess reliability using a factor-analytic approach such as coefficient $\omega_h$ (McDonald, 1985, 1999; Zinbarg et al., 2005), further referred to as $\omega_h$. The subscript $h$ denotes that this measure of reliability is derived from the hierarchical factor analytic model. That is, it is assumed that all items measure a common factor that accounts for a major proportion of variance in the scaled scores. In addition, it is assumed that each item measures a unique skill uncorrelated with the common scale. For the purposes of this study, we used a bifactor model in which all items load on a general factor and on a unique factor. All unique factors are uncorrelated. The $\omega_h$ statistic used is calculated as

$$\omega_h = \frac{\left(\sum_{j=1}^{J} \lambda_{gj}\right)^2}{\sigma_X^2}, \tag{10.11}$$

where $\lambda_{gj}$ are the factor loadings on the general factor.

The goal of this research is to examine and compare the performance of $\alpha$ and $\omega_h$ under several different test conditions including the correlations between dimensions, number of items, discrimination power of the items, and whether the difficulty of the items is optimal given the ability distribution of the examinees.

The response data were generated using the compensatory multidimensional two-parameter IRT model (M2PL) (Reckase, 2009). The M2PL can be expressed as

$$\boldsymbol{p}_j(\boldsymbol{\theta}) = \boldsymbol{P}\left(\boldsymbol{u}_j = \mathbf{1}|\boldsymbol{\theta}\right) = \frac{\mathbf{1}}{\mathbf{1} + \mathbf{e}^{-\left(\sum_{k=1}^{m} \boldsymbol{a}_{jk}\boldsymbol{\theta}_{ik} + \mathbf{d}_j\right)}}, \tag{10.12}$$

where $\theta = (\theta_1, \theta_2, \cdots \theta_k, \cdots \theta_m)$ is a $m$-length vector of the latent scores with elements indexed as $\theta_{ik}$(the score of person $i$ on dimension $k$), $a_{jk}$ is a discrimination for item $j$ on dimension $k$, respectively, and $d_j$ is an intercept term denoting the composite difficulty of each item. The MDISC index is the multidimensional analog

to unidimensional discrimination parameter, *a*. It is a composite discrimination index for each that can be expressed as

$$\text{MDISC}_i = \sqrt{\sum_{k=1}^{m} a_{jk}^2} \tag{10.13}$$

where $a_{jk}^2$ is defined above.

## 10.2 Research Design

This is a simulation study. The response data were generated under prescribed testing conditions with multiple replications. Three coefficients were computed for each data set and then the comparative results aggregated across replications: (i) $\rho_{TX}^2$, the true scale reliability when the true score and error variances are known (through simulation), (ii) $\alpha$ (Eq. 10.4), and (iii) $\omega_h$ (Eq. 10.11). This design demonstrates how logically influential test design considerations such as test length, item discrimination, and the homogeneity of items relative to the population mean(s) impact those three reliability coefficients. The study included five completely crossed design factors:

- Number of items ($J = 24$, $J = 48$)
- Levels of MDISC (low MDISC, 0.4–0.8; moderate MDISC, 0.8–1.2; high MDISC, 1.2–1.6)
- Number of dimensions ($m = 1, 2, 3, 4$)
- Location of mean item difficulty ($d = 0, 1$) given the examinee distribution will always be centered at the origin
- Correlation of abilities ($\rho = .0, .5$)

The sample size for each simulation was fixed at 1000 randomly generated examinees sampled from a standard normal univariate or multivariate normal distribution centered at the origin for each simulation. Each condition was further replicated 100 times to provide empirical sampling distributions of each reliability coefficient for comparative purposes.

## 10.3 Reliability Estimation and Evaluation

Three reliability coefficients were calculated for each of the simulated data sets: the true scale reliability, $\rho_{TX}^2$, coefficient $\alpha$, and $\omega_h$ (based on a fitted bi-factor model). As noted earlier, the true scale reliability was calculated using Eq. 10.1 where the

true score variance is the variance of the expected scores of the N-examinees over J-items:

$$\sigma_T^2 = \sigma^2 \left( \sum_{i=1}^{N} \sum_{j=1}^{n} P\left(u_i = 1 | \theta_{i1}, \theta_{i2}\, a_{j1}, a_{j2}, d_j\right) \right) \tag{10.14}$$

using the generated $N \times m$ matrix, $\boldsymbol{\theta}$, and the $J \times (m + 1)$ matrix of generated item parameters. The raw score variance is calculated using the total score for each person and including all the items in the test. The $\alpha$ and $\omega_h$ were calculated using the corresponding functions in the R package **psych** (Revelle, 2021). That package calculates the three reliabilities given in Eqs. 10.4 and 10.11. In aggregate, there were 96 conditions ($2 \times 3 \times 4 \times 2 \times 2$), and each condition was replicated 100 times to provide empirical sampling distributions of the three coefficients. In particular, the means and standard deviations of those sampling distributions were computed across the 100 replications per condition, and graphical visualizations were created using the R package **ggplot2** (Wickham, 2016). All the simulations, data management, and analytical aspects of this study were carried out using R (R Core Team, 2021).

## 10.4 Results

The 5 design factors produced 96 simulation test design conditions. These factors were expected to have direct or indirect impact on the three reliability indices, $\rho_{TX}^2$, $\alpha$, and $\omega_h$. The impact of the number of items (test length) on reliability is well-known given the extensive body of research on the Spearman-Brown formula (e.g., Angoff, 1953; Traub, 1997),

$$\rho_{XX'}^{*} = q\rho_{XX'} / \left[1 + (q - 1)\, \rho_{XX'}\right] \tag{10.15}$$

where $\rho_{XX'}$ is the original reliability index and $q$ is the ratio of new to original (old) test lengths. In contrast, the average MDISC (composite item discrimination) and item location were generated to either *offset* or *match* to the population centroids' impact the contribution of each item to the score variance (e.g., Gulliksen, 1950). These two factors also directly and indirectly reflect on item quality—especially the item discrimination parameters and MDISC, which act as weights for the latent scores. Finally, the number of underlying dimensions and the correlation between those dimensions represent the dispersion of the measurement *signal* across the apparent latent structures representing the item covariances. Including these latter two conditions in the simulation directly speaks to the motivation for $\omega_h$, that is, to have a reliability index that responds to untended or idiosyncratic dimensionality, or to a test that includes multiple dimensions by design and perhaps reports the total score as weighted linear composite of subscores. Increasing the dimensionality

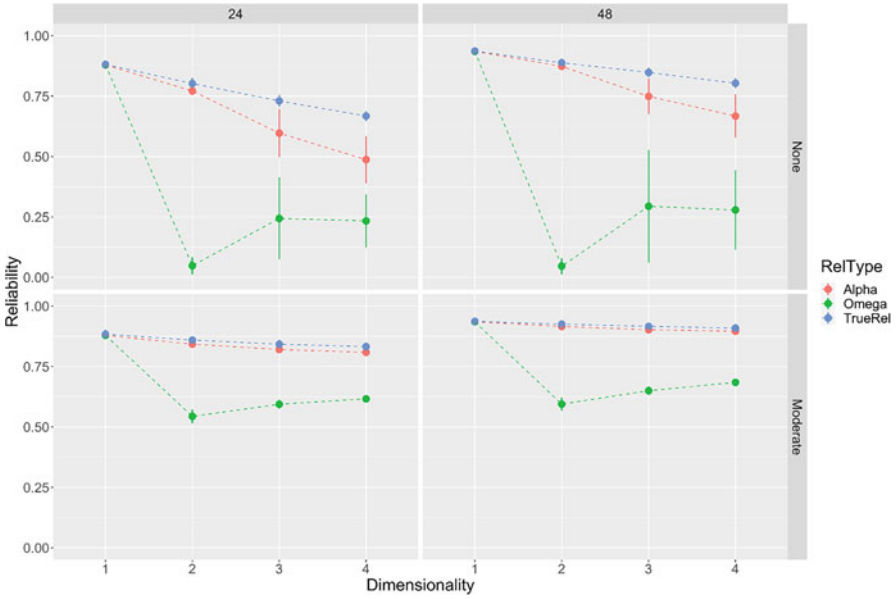


**Fig. 10.1** Summary of reliability coefficients for high MDISC and item difficulty matched to the population proficiency score centroids: $\mu(d) - \mu(\theta_k) = 0$ (100 replications per condition)

and covariance(s) among the underlying factors should disperse the "measurement signal" relative to a reported total score.

For the most part, these factors produced results that met expectations. Figures 10.1, 10.2, 10.3, 10.4, 10.5 and 10.6 include "trellis" or facetted multi-plots that embed a bivariate plot conditioned on the number of items (columns) and the magnitude of correlation between the underlying dimensions or factors (*none* implies a zero correlation between the factors; *moderate* implies a correlation of .5 between all factors). The number of dimensions is shown along the horizontal axis for each plot, and the vertical axis represents the magnitude of the correlation. The three plotted outcomes in each cell of the multi-plot denote the three reliability indices: $\rho^2_{TX}$, $\alpha$, and $\omega_h$. These results are summarized as the mean and standard error of the reliability coefficients across 100 replications per combination of simulation conditions.

As Fig. 10.1 shows (high MDISC, with the mean item difficulty matched to the population centroids, $\mu(d) - \mu(\theta_k) = 0$ for all $k$), there is a noticeable increase in the $\rho^2_{TX}$ and $\alpha$ coefficients as the test length increased from 24 to 48 items, and a decrease in the coefficients as the number of dimensions increased from 1 up to 4 due to the amount of total test score signal dispersion among the dimensions. The three coefficients are all highly similar in the unidimensional case ($m$=1) with $\alpha$ and $\omega_h$ essentially being identical. The coefficients only start to decline as the total score signal is dispersed across two or more underlying factors. Note that the

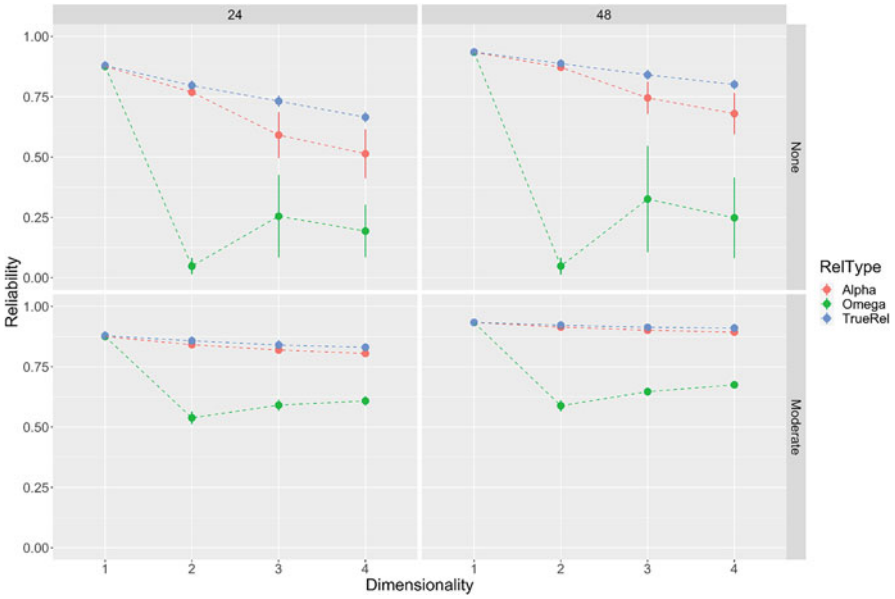


**Fig. 10.2** Summary of reliability coefficients for high MDISC with item difficulty offset from the population proficiency score centroids: $\mu(\theta_k) - \mu(d) = 1$ (100 replications per condition)

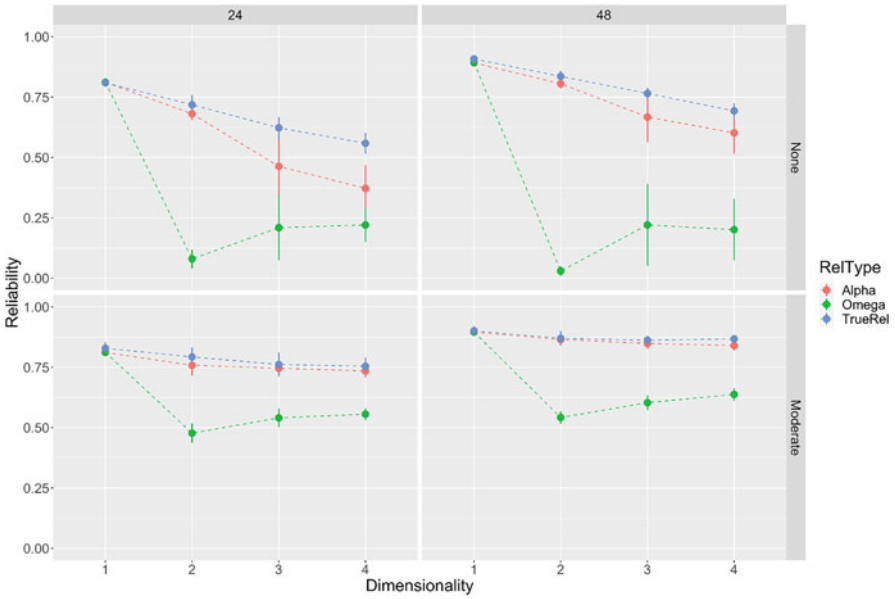


**Fig. 10.3** Summary of reliability coefficients for moderate MDISC and item difficulty matched to the population proficiency score centroids: $\mu(\theta_k) - \mu(d) = 0$ (100 replications per condition)

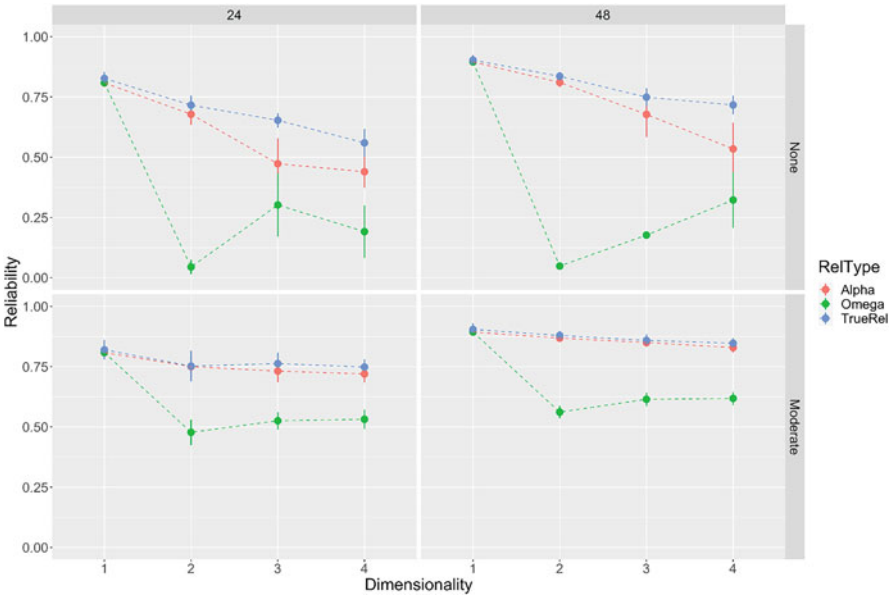


**Fig. 10.4** Summary of reliability coefficients for moderate MDISC with item difficulty offset from the population proficiency score centroids: $\mu(\theta_k) - \mu(d) = 1$ (100 replications per condition)

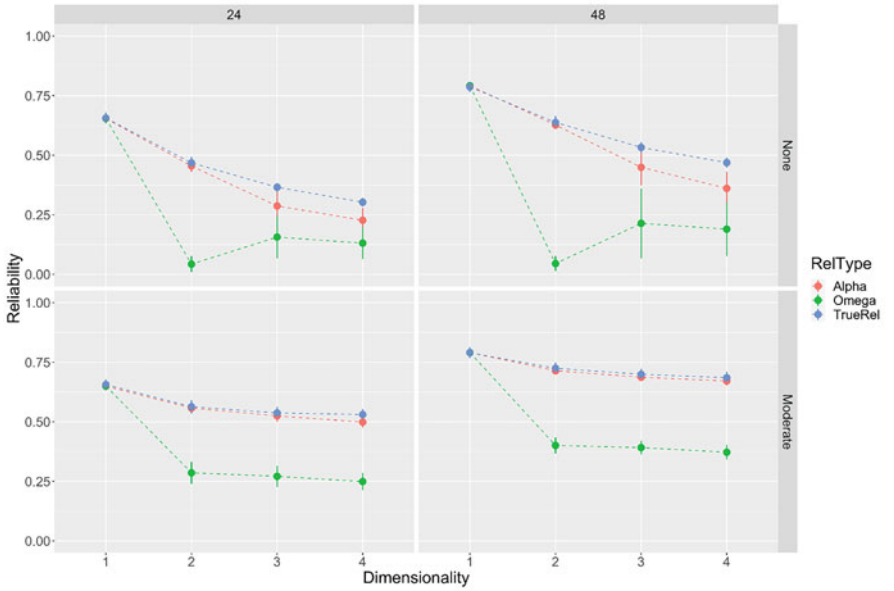


**Fig. 10.5** Summary of reliability coefficients for low MDISC and item difficulty matched to the population proficiency score centroids: $\mu(\theta_k) - \mu(d) = 0$ (100 replications per condition)

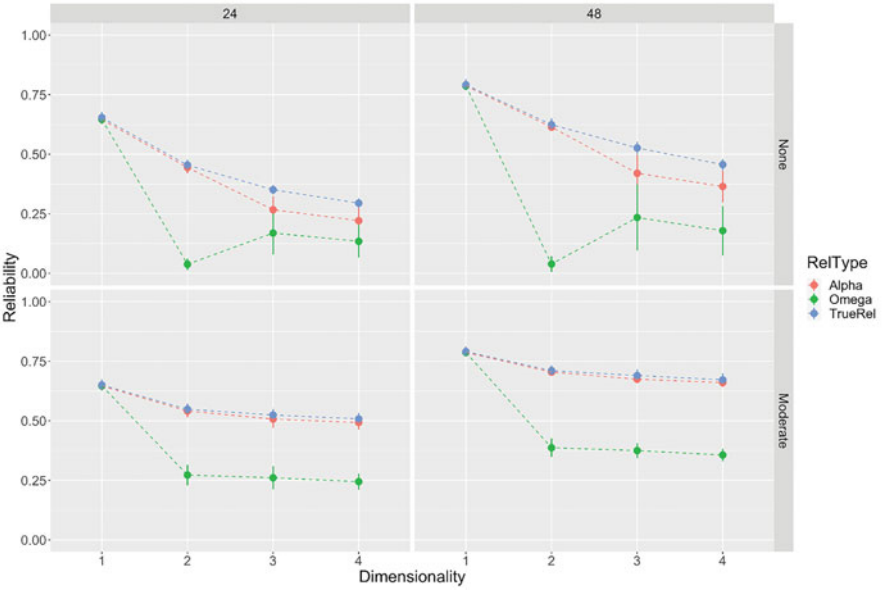


**Fig. 10.6** Summary of reliability coefficients for low MDISC with item difficulty offset from the population proficiency score centroids: $\mu(\theta_k) - \mu(d) = 1$ (100 replications per condition)

zero-correlation condition is rather unrealistic in a practical sense[1], but provides a reasonable baseline under "maximum dispersion" conditions. Interestingly, the mean values of $\omega_h$ tend to somewhat track with the inter-factor correlations (.0 = *none* or .5 = *moderate*).

Figure 10.2 (high MDISC, with the mean item difficulty offset from the population centroids, $\mu(\theta_k) - \mu(d) = 1$ for all dimensions) shows a pattern that is very consistent with Fig. 10.1. Cronbach's $\alpha$ values tend to be smaller than the "true reliabilities" with known true scores, $\rho^2_{TX}$. This likely reflects some sampling error when estimating the item error variances (see Eq. 10.3). The $\omega_h$ coefficients, again, somewhat track with the magnitude of the inter-factor correlations, although the mean values are also confounded by the high MDISC present in the items.

Figure 10.3 (moderate average MDISC, with the mean item difficulty matched to the population centroids, $\mu(\theta_k) - \mu(d) = 0$ for all dimensions) begins to show an interesting pattern where the mean $\alpha$ and $\rho^2_{TX}$ values respond to the reduced composited item discrimination, but the $\omega_h$ coefficients do not.

Figure 10.4 (moderate average MDISC, with the mean item difficulty offset from population centroids, $\mu(\theta_k) - \mu(d) = 1$ for all dimensions) confirms the coefficient

[1] In practice, it would be very rare to encounter a test designed to measure two or more underlying traits with NO covariance between the traits. Even tests measuring distinctly different traits like mathematics and English language arts tend to positively correlate in the moderate range.

patterns of Fig. 10.3; that is, the $\omega_h$ coefficients respond more to the amount of total score signal dispersion than to the reduced composite item discrimination. The mean $\alpha$ and $\rho^2_{TX}$ values respond to the reduced composited item discrimination and, to a lesser degree, to the signal dispersion across dimensions.

Figures 10.5 and 10.6 show an overall decline in mean $\alpha$ and $\rho^2_{TX}$ values proportional to both the low average MDISC values and the dimensional dispersion of the total score signal. Interesting, and similar to Figs. 10.3 and 10.4, the latter dispersion has less impact across the increasing number of dimensions than under the high discrimination condition. Increasing the test length helps to somewhat offset the decline in the reliability coefficients, but the recommendation to write high-quality items and monitor that the level of composite item discrimination remains as high as possible seems to be good advice.

## 10.5 Conclusion

In this study, we varied testing conditions that we felt would influence the performance of the three reliability coefficients: (1) true reliability, (2) Cronbach's $\alpha$, and (3) $\omega_h$. As the number of items was doubled from 24 to 48, there was the expected proportional increase in reliability. Likewise, as the discrimination of the items, MDISC, increased, the magnitude of the reliability coefficients also unilaterally increased. The simulation response data were generated relative to an underlying multidimensional simple structure for three of the four simulation conditions. As the correlations between the multidimensional latent abilities increased from 0 to .5, thus "collapsing" the latent space—the reliability coefficients also proportionally increased. The effect of increasing the average difficulty of the items, that is, increasing the amount of offset between the location of maximum measurement information relative to the centroid of the examinee ability, joint latent distributions did not induce any prominent change in reliability.

The simulation condition that appeared to demonstrate the greatest impact on the reliability coefficients was multidimensionality. As the number of dimensions increased, coefficient $\omega$ dropped considerably in comparison to the true scale reliability and coefficient $\alpha$. This was anticipated because $\omega_{\mathrm{h}}$ was computed using the sum of the loadings on the general factor in the hierarchical, orthogonal bifactor model, where all factors are uncorrelated. Because the data were generated using simple structure, the loadings on the unique factors were higher than the loadings on the general factor, creating significant dispersion in the measurement "signal"—specifically, inducing "noise" relative to the general factor. That is, the R-packages that were used estimated $\omega_h$ using the bi-factor model versus a common factor or component model.

In the unidimensional case, $\alpha$ and $\omega$ were always equal. In some cases, these coefficients exceeded the true scale reliability. As dimensionality increased, $\alpha$ like the $\rho^2_{TX}$ decreased though not nearly as much $\omega$. It appeared that $\alpha$ was not affected as much as $\omega_h$ by the increase in dimensionality. There was one notable

inconsistency. In the two-dimensional case, $\omega$ was consistently lower than in the three- and four-dimensional cases across all conditions. This may have been a function of the sampled item discrimination parameters.

It seems clear that testing practitioners must be advised always to conduct a thorough dimensionality analysis of their test results relative to the intended, reported score scale(s) and further evaluate the dimensionality analysis outcomes in terms of the test specification so that they can articulate the meaning of the observed score scale. Only evaluating a reliability coefficients or standard errors of measurement is not sufficient.

Future research will extend the current research to incorporate factorially complex item structures where the multidimensionality may relate to nuisance dimensions of idiosyncratic characteristics of the items (i.e., items loadings on both intended and unintended factors underlying the data). We also plan to examine reliability from a multidimensional IRT perspective and relate more directly to the concept of a unidimensional composite of intended multidimensional traits (i.e., Wang's (1985) reference composite). Lastly, we plan to experiment with the formulation of $\omega_h$ and determine if additional information about dimensionality and its effect on reliability can be delineated for testing practitioners.

## References

Angoff, W. H. (1953). Test reliability and effective test length. *Psychometrika, 18*(1), 1–14. https://doi.org/10.1007/BF02289023

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology, 1904–1920, 3*(3), 296–322. https://doi.org/10.1111/j.2044-8295.1910.tb00207.x

Brennan, R. L. (2001). An essay on the history and future of reliability from the perspective of replications. *Journal of Educational Measurement, 38*, 295–317. https://doi.org/10.1111/j.1745-3984.2001.tb01129.x

Cho, E., & Kim, S. (2015). Cronbach's coefficient alpha: Well known but poorly understood. *Organizational Research Methods, 18*(2), 207–230. https://doi.org/10.1177/1094428114555994

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology, 78*(1), 98–104. https://doi.org/10.1037/0021-9010.78.1.98

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297–334. https://doi.org/10.1007/BF02310555

Cronbach, L. J., & Shavelson, R. J. (2004). My current thoughts on coefficient alpha and successor procedures. *Educational and Psychological Measurement, 64*(3), 391–418. https://doi.org/10.1177/0013164404266386

Davenport, E. C., Davison, M. L., Liou, P.-Y., & Love, Q. U. (2015). Reliability, dimensionality, and internal consistency as defined by Cronbach: Distinct albeit related concepts. *Educational Measurement: Issues and Practice, 34*(4), 4–9. https://doi.org/10.1111/emip.12095

Davison, M. L., & Davenport, E. C. (2015, April 15–19). *Coefficient α and dimensionality* [Paper presentation]. Annual meeting of the National Council on Measurement in Education, Chicago, IL, United States.

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement* (pp. 105–146). Macmillan.

Flanagan, J. C. (1937). A proposed procedure for increasing the efficiency of objective tests. *Journal of Educational Psychology, 28*(1), 17–21. https://doi.org/10.1037/h0057430

Friedman, S., & Weisberg, H. F. (1981). Interpreting the first eigenvalue of a correlation matrix. *Educational and Psychological Measurement, 41*(1), 11–21. https://doi.org/10.1177/001316448104100102

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*(4), 827–838. https://doi.org/10.1177/001316447703700403

Green, S. B., & Yang, Y. (2009). Commentary on coefficient alpha: A cautionary tale. *Psychometrika, 74*(1), 121–135. https://doi.org/10.1007/s11336-008-9098-4

Gulliksen, H. (1950). *Theory of mental tests*. Wiley. https://doi.org/10.1037/13240-000

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika, 10*(4), 255–282. https://doi.org/10.1007/BF02288892

Hoyt, C. (1941). Test reliability estimated by analysis of variance. *Psychometrika, 6*(3), 153–160. https://doi.org/10.1007/BF02289270

Kane, M. T. (1996). The precision of measurements. *Applied Measurement in Education, 9*(4), 355–379. https://doi.org/10.1207/s15324818ame0904_4

Kuder, G. F., & Richardson, M. W. (1937). The theory of estimation of test reliability. *Psychometrika, 2*(3), 151–160. https://doi.org/10.1007/BF02288391

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Addison-Wesley.

McDonald, R. P. (1985). *Factor analysis and related methods*. Psychology Press. https://doi.org/10.4324/9781315802510

McDonald, R. P. (1999). *Test theory: A unified treatment*. Psychology Press. https://doi.org/10.4324/9781410601087

R Core Team. (2021). *R: A language and environment for statistical computing*. [Computer Software]. R Foundation for Statistical Computing, http://www.R-project.org/

Reckase, M. D. (2009). *Multidimensional item response theory*. Springer. https://doi.org/10.1007/978-0-387-89976-3

Revelle, W. (2021). *Psych R package* (Version 2.1.3) [Computer Software]. https://personality-project.org/r/psych/

Rodriguez, M. C., & Maeda, Y. (2006). Meta-analysis of coefficient alpha. *Psychological Methods, 11*(3), 306–322. https://doi.org/10.1037/1082-989X.11.3.306

Rulon, P. (1939). A simplified procedure for determining the reliability of a test by split-halves. *Harvard Educational Review, 9*(1), 99–103.

Sijtsma, K. (2009a). On the use, misuse, and very limited usefulness of Cronbach's α. *Psychometrika, 74*(1), 107–120. https://doi.org/10.1007/s11336-008-9101-0

Sijtsma, K. (2009b). Reliability beyond theory and into practice. *Psychometrika, 74*(1), 169–173. https://doi.org/10.1007/s11336-008-9103-y

Spearman, C. (1910). Correlation calculated with faulty data. *British Journal of Psychology, 3*(3), 271–295. https://doi.org/10.1111/j.2044-8295.1910.tb00206.x

Ten Berge, J. M. F., & Sočan, G. (2004). The greatest lower bound to the reliability of a test and the hypothesis of unidimensionality. *Psychometrika, 69*(4), 613–625. https://doi.org/10.1007/BF02289858

Thompson, B., & Vacha-Haase. (2000). Psychometrics is datametrics: The test is not reliable. *Educational and Psychological Measurement, 60*(2), 174–195. https://doi.org/10.1177/0013164400602002

Traub, R. E. (1997). Classical test theory in historical perspective. *Educational Measurement: Issues and Practice, 16*(4), 8–14. https://doi.org/10.1111/j.1745-3992.1997.tb00603.x

Wang, M. (1985). *Fitting a unidimensional model to multidimensional item response data: The effects of latent space misspecification on the application of IRT.* Unpublished manuscript, University of Iowa.

Wickham, H. (2016). *Ggplot2: Elegant graphics for data analysis* (2nd ed.). Springer. https://doi.org/10.1007/978-3-319-24277-4

Zinbarg, R. E., Revelle, W., Yovel, I., & Li, W. (2005). Cronbach's α, Revelle's β, and Mcdonald's $\omega_H$: Their relations with each other and two alternative conceptualizations of reliability. *Psychometrika, 70*(1), 123–133. https://doi.org/10.1007/s11336-003-0974-7