Center for Advanced Studies in
Measurement and Assessment

CASMA Technical Note

Number 2

Some Perspectives on KR–21

Robert L. Brennan
Won-Chan Lee†

March 2006

†Robert L. Brennan is E. F. Lindquist Chair in Measurement and Testing and Director, Center for Advanced Studies in Measurement and Assessment (CASMA), 210D Lindquist Center, College of Education, University of Iowa, Iowa City, IA 52242 (email: robert-brennan@uiowa.edu). Won-Chan Lee is a research scientist in CASMA, 210E Lindquist Center, University of Iowa, Iowa City, IA 52242 (email: won-chan-lee@uiowa.edu).

Center for Advanced Studies in
Measurement and Assessment (CASMA)
College of Education
University of Iowa
Iowa City, IA 52242
Tel: 319-335-5439
Web: www.education.uiowa.edu/casma

Kuder and Richardson (1937) was one of the first papers to introduce internal consistency reliability coefficients. Based on the equation numbers in their paper, the two coefficients introduced by Kuder and Richardson are universally referred to as KR–20 and KR–21, both of which assume that items are dichotomously scored, with KR–21 being less than or equal to KR–20. These coefficients were developed long before most of classical test theory was formalized according to today's conventions (see, for example, Feldt & Brennan, 1989). Consequently, relating these coefficients to current ones has usually tended to be more of an algebraic exercise than a conceptual one. For example, it is well known that algebraically KR–20 is a special case of Cronbach's (1951) coefficient alpha for the case of dichotomously-scored items, although the assumptions used by Kuder and Richardson to develop KR–20 are substantially different from the assumptions used by Cronbach.

For the most part, the status of KR–21 in the measurement literature has been marginal and/or confusing. For example, in the original Kuder and Richardson development, KR–21 is simply KR–20 when items are equally difficult. By contrast, in the beta-binomial model, which can be derived without assuming equally difficult items, KR–21 is typically taken as the "appropriate" estimate of reliability (see, for example, Huynh, 1976).

A relationship between KR–20 and KR–21 was exploited by Keats (1957) in order to obtain a correction factor for Lord's (1955, 1957) conditional error variance (CEV) for an individual. Keats considers Lord's CEV to be biased (too big), but multiplying it by the ratio of (1 - KR–20) to (1 - KR–21) removes the bias in the sense that the average of the squared values of the "corrected" CEVs is approximately the error variance in KR–20. Two assumptions are evident in this logic: Lord's CEV is biased, and the correct error variance is that incorporated in KR–20. From the perspective of classical theory, these assumptions are supportable; from the perspective of generalizability theory (Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Brennan, 2001) they are not. For example, as discussed later, Lord's CEV is identical to the estimated conditional absolute error variance (for a single-facet design) in generalizability theory, and KR–21 is identical to a reliability-like coefficient that involves absolute error variance.

In other words, from the perspective of generalizability theory, KR–20 and KR–21 are reliability-like coefficients that incorporate different error variances.[1] There is more to the story than just this fact, but that is the most important part. For example, there is a need to distinguish carefully between biased and unbiased estimates of particular variances in order to establish certain equivalences. This technical note delves into these and related matters in some detail.

It is assumed here that readers are familiar with the conventions and basic results of generalizability theory, particularly as discussed by Brennan (2001). In accordance with these convention, the mean-score metric is used here rather than the total-score metric. That is, we deal with person mean scores over items as opposed to person total scores over items.

[1]This perspective on these two coefficients was partially foreshadowed in an unpublished paper by Buros (1963) that lays out a framework for viewing reliability coefficients as different types of intraclass correlation coefficients.

Proof that $\widehat{\Phi}(\lambda = \overline{X}) = $ KR–21

For domain-referenced contexts involving a cutting score λ, Brennan (2001, p. 48) states that, "(w)hen $\lambda = \overline{X}$, $\widehat{\Phi}(\lambda) =$ KR–21, which is indicative of the fact that KR–21 involves absolute error variance, not the relative error variance of classical theory." Letting $\hat{\rho}_{21}$ denote the KR–21 reliability coefficient,

$$\begin{aligned} \hat{\rho}_{21} &= \widehat{\Phi}(\lambda = \overline{X}) \qquad (1) \\ &= \frac{\hat{\sigma}^2(p) - \hat{\sigma}^2(\overline{X})}{\hat{\sigma}^2(p) - \hat{\sigma}^2(\overline{X}) + \hat{\sigma}^2(\Delta)}, \qquad (2) \end{aligned}$$

where

$$\hat{\sigma}^2(\overline{X}) = \frac{\hat{\sigma}^2(p)}{n_p} + \frac{\hat{\sigma}^2(i)}{n_i} + \frac{\hat{\sigma}^2(pi)}{n_p\, n_i}. \qquad (3)$$

Equation 2 has the usual form of an estimated reliability-like coefficient, namely, estimated universe score variance divided by itself plus estimated error variance, which is absolute error variance in this case. The appendix in Brennan and Kane (1977a) essentially proves Equation 2. Provided next is a somewhat more detailed proof.

A typical expression for $\hat{\rho}_{21}$ in terms of person mean scores (over items) is

$$\hat{\rho}_{21} = \frac{n_i}{n_i - 1}\left[1 - \frac{\overline{X}(1-\overline{X})}{n_i\, S^2(\overline{X}_p)}\right], \qquad (4)$$

where $\overline{X}$ is the mean (over persons) of the person mean scores, and $S^2(\overline{X}_p)$ is observed score variance in the sense of a *biased* estimate[2]; i.e.,

$$S^2(\overline{X}_p) = \frac{\sum_p (\overline{X}_p - \overline{X})^2}{n_p} = \frac{SS(p)}{n_i\, n_p} = \left[\frac{n_p - 1}{n_p\, n_i}\right] MS(p). \qquad (5)$$

For our purposes here, it is convenient to express Equation 4 as

$$\hat{\rho}_{21} = \frac{\dfrac{n_i}{n_i - 1}\left[S^2(\overline{X}_p) - \dfrac{\overline{X}(1-\overline{X})}{n_i}\right]}{S^2(\overline{X}_p)}, \qquad (6)$$

where the numerator is an estimate of universe score variance, and the denominator is obviously observed score variance. Note, as well, that observed score variance minus the estimated universe score variance in the numerator is

$$S^2(\overline{X}_p) - \frac{n_i}{n_i - 1}\left[S^2(\overline{X}_p) - \frac{\overline{X}(1-\overline{X})}{n_i}\right] = \frac{\overline{X}(1-\overline{X}) - S^2(\overline{X}_p)}{n_i - 1} = \hat{\sigma}^2(\Delta), \qquad (7)$$

which is a result originally provided by Brennan and Kane (1977b). The next paragraph proves that the denominators of Equations 2 and 6 are the same; the subsequent paragraph proves that the numerators are the same.

[2]Note that in Brennan (2001) $S^2(\overline{X}_p)$ or $S^2(p)$ is observed score variance in the sense of an *unbiased* estimate.

Since $\hat{\sigma}^2(pi) = MS(pi)$, $\hat{\sigma}^2(i) = [MS(i) - MS(pi)]/n_p$, and

$$\hat{\sigma}^2(p) = \frac{MS(p) - MS(pi)}{n_i}, \tag{8}$$

it is easy to show that

$$\hat{\sigma}^2(\overline{X}) = \frac{MS(p) + MS(i) - MS(pi)}{n_p\, n_i}, \tag{9}$$

and

$$\hat{\sigma}^2(\Delta) = \frac{MS(i) - MS(pi)}{n_p\, n_i} + \frac{MS(pi)}{n_i}. \tag{10}$$

It follows from Equations 8–10 that the denominator of Equation 2 is

$$\hat{\sigma}^2(p) - \hat{\sigma}^2(\overline{X}) + \hat{\sigma}^2(\Delta) = \frac{MS(p)}{n_i} - \frac{MS(p)}{n_p\, n_i} = \frac{1}{n_i}\left(\frac{n_p - 1}{n_p}\right) MS(p) = S^2(\overline{X}_p),$$

which is the denominator in Equation 6.

Using Equations 8 and 9, the numerator of Equation 2 is

$$\begin{aligned}
\hat{\sigma}^2(p) - \hat{\sigma}^2(\overline{X}) &= \frac{n_p MS(p) - n_p MS(pi) - MS(p) - MS(i) + MS(pi)}{n_p\, n_i} \\
&= \frac{(n_p - 1)MS(p)}{n_p\, n_i} - \frac{(n_p - 1)MS(pi) + MS(i)}{n_p\, n_i} \\
&= \frac{SS(p)}{n_p\, n_i} - \frac{SS(pi) + SS(i)}{(n_i - 1)n_p\, n_i} \\
&= \frac{SS(p)}{n_p\, n_i} - \frac{1}{n_i - 1}\left[\frac{SS(\text{total}) - SS(p)}{n_p\, n_i}\right].
\end{aligned} \tag{11}$$

For dichotomously-scored items

$$\frac{SS(\text{total})}{n_p\, n_i} = \frac{\sum_p \sum_i (X_{pi} - \overline{X})^2}{n_p\, n_i} = \overline{X}(1 - \overline{X}).$$

Replacing this result, as well as Equation 5, in Equation 11 gives

$$\hat{\sigma}^2(p) - \hat{\sigma}^2(\overline{X}) = \frac{n_i}{n_i - 1}\left[S^2(\overline{X}_p) - \frac{\overline{X}(1 - \overline{X})}{n_i}\right], \tag{12}$$

which is the numerator in Equation 6.

The algebraic equalities reported in this section are a curious mixture of biased and unbiased estimates. In particular:

- The numerator of $\hat{\rho}_{21}$ in Equation 6 (i.e., Equation 12) is a biased estimate of universe score variance, $\sigma^2(p)$.
- The expression for overall absolute error variance in Equation 7 is an unbiased estimate with respect to item sampling, in the sense discussed in the next section.

Note, as well, that Equations 4 and 6 and other expressions for $\hat{\rho}_{21}$ typically use a biased estimate of observed score variance in the denominator, whereas in generalizability theory the convention is to use an unbiased estimate.

KR–21 and Lord's Binomial Error Variance

It is often stated that the error variance in $\hat{\rho}_{21}$ is equal to the average (over persons) of Lord's (1955, 1957) binomial error variances (see, for example, Brennan, 2001, p. 161). This fact is demonstrated next.

Lord's estimated conditional error variance for person p in terms of mean scores is

$$\hat{\sigma}^2(\Delta_p) = \frac{\overline{X}_p(1-\overline{X}_p)}{n_i - 1}, \tag{13}$$

and the average over persons is

$$\overline{\hat{\sigma}^2}(\Delta_p) = \frac{\sum_p \overline{X}_p(1-\overline{X}_p)}{n_p(n_i - 1)} = \frac{\sum_p \overline{X}_p - \sum_p \overline{X}_p^2}{n_p(n_i - 1)}. \tag{14}$$

According to Equation 7,

$$\begin{aligned}
\hat{\sigma}^2(\Delta) &= \frac{\overline{X}(1-\overline{X}) - S^2(\overline{X}_p)}{n_i - 1} \\
&= \frac{\overline{X}(1-\overline{X}) - \sum_p(\overline{X}_p - \overline{X})^2/n_p}{n_i - 1} \\
&= \frac{\overline{X}(1-\overline{X}) - (\sum_p \overline{X}_p^2/n_p) + \overline{X}^2}{n_i - 1} \\
&= \frac{\sum_p \overline{X}_p - \sum_p \overline{X}_p^2}{n_p(n_i - 1)},
\end{aligned} \tag{15}$$

which is identical to Equation 14.

The estimate of conditional error variance in Equation 13 is an *un*biased estimate of the absolute error variance for an examinee in the sense that $\hat{\sigma}^2(\Delta_p)$ is unbiased with respect to *item* sampling for the examinee. It follows that the estimated overall absolute error variance $\overline{\hat{\sigma}^2}(\Delta_p) = \hat{\sigma}^2(\Delta)$ is also unbiased with respect to item sampling. Since the parameter (absolute error variance) is typically defined in generalizability theory as the expected value over the population of the conditional absolute error variances, the mean over persons of the estimated conditional absolute error variances can be viewed as unbiased with respect to person sampling.[3]

[3] In theory, absolute error variance could be defined in a different manner (see Cronbach et al., 1972, pp. 79, 82), but the alternative definition of absolute error variance would make it equivalent to relative error variance.

Note also that

$$\begin{aligned} S^2(\overline{X}_p)[1-\hat{\rho}_{21}] &= S^2(\overline{X}_p) - \left[\frac{n_i S^2(\overline{X}_p)}{n_i - 1} - \frac{\overline{X}(1-\overline{X})}{n_i - 1}\right] \\ &= \frac{\overline{X}(1-\overline{X}) - S^2(\overline{X}_p)}{n_i - 1} \\ &= \frac{\overline{X}(1-\overline{X})}{n_i - 1} - \frac{1}{n_i - 1}\left[\frac{\sum_p \overline{X}_p^2}{n_p} - \overline{X}^2\right] \\ &= \frac{n_p\overline{X}(1-\overline{X}) - \sum_p \overline{X}_p^2 + n_p\overline{X}^2}{n_p(n_i - 1)} \\ &= \frac{\sum_p \overline{X}_p - \sum_p \overline{X}_p^2}{n_p(n_i - 1)}, \end{aligned} \tag{16}$$

which is equal to Equations 14 and 15. In short,

$$\overline{\hat{\sigma}^2}(\Delta_p) = \hat{\sigma}^2(\Delta) = S^2(\overline{X}_p)[1-\hat{\rho}_{21}], \tag{17}$$

which provides a curious mixture of biased and unbiased estimates. Specifically, $\overline{\hat{\sigma}^2}(\Delta_p) = \hat{\sigma}^2(\Delta)$ is unbiased with respect to item sampling, while $S^2(\overline{X}_p)$ is biased with respect to person sampling.

The introduction to this technical note made reference to the fact that Keats' (1957) corrected conditional error variance for an individual is

$$\hat{\sigma}^2_{Keats}(\Delta_p) = \frac{\overline{X}_p(1-\overline{X}_p)}{n_i - 1}\left[\frac{1-\hat{\rho}_{20}}{1-\hat{\rho}_{21}}\right]. \tag{18}$$

Keats motivation for suggesting his correction factor was that the average over persons of Equation 18 would be the error variance in KR–20, namely, $\hat{\sigma}^2(\delta)$. That is not strictly true. Rather, with the results reported here it is easy to show that the average is $[(n_p-1)/n_p]\hat{\sigma}^2(\delta)$.

KR–21, KR–20, and $\widehat{\boldsymbol{\Phi}}$

Let $\hat{\rho}_{20}$ denote the KR–20 reliability coefficient. In terms of person mean scores (over items), a typical expression for KR–20 is

$$\hat{\rho}_{20} = \frac{n_i}{n_i - 1}\left[1 - \frac{\sum \overline{X}_i(1-\overline{X}_i)}{n_i^2\, S^2(\overline{X}_p)}\right], \tag{19}$$

where $\overline{X}_i$ is the difficulty level of item i. An equivalent expression is

$$\hat{\rho}_{20} = \frac{\dfrac{n_i}{n_i - 1}\left[S^2(\overline{X}_p) - \dfrac{\sum \overline{X}_i(1-\overline{X}_i)}{n_i^2}\right]}{S^2(\overline{X}_p)}. \tag{20}$$

It is evident that these expressions for KR–20 use a biased estimate (for person sampling) of the observed score variance for persons. Also, the numerator of Equation 20 is a biased estimate (for person sampling) of universe score variance, $\sigma^2(p)$. Of course, we could express Equation 20 in terms of unbiased estimates (for person sampling) by multiplying both the numerator and the denominator by $n_p/(n_p-1)$, which would not alter the value of the coefficient.

As discussed by Brennan (2001), $\hat{\rho}_{20}$ can be expressed in terms of variance component estimates as follows:

$$\hat{\rho}_{20} = \frac{\hat{\sigma}^2(p)}{\hat{\sigma}^2(p) + \hat{\sigma}^2(\delta)}. \tag{21}$$

It is important to note that this expression for $\hat{\rho}_{20}$ uses unbiased estimates (with respect to person sampling) for both the numerator (estimated universe score variance) and denominator (observed score variance).

The denominator is sometimes denoted

$$\hat{\sigma}^2(\overline{X}_p) = \hat{\sigma}^2(p) + \hat{\sigma}^2(\delta), \tag{22}$$

where

$$\hat{\sigma}^2(\overline{X}_p) = \frac{n_p}{n_p-1} S^2(\overline{X}_p) = \frac{n_p}{n_p-1}\left[\frac{n_p-1}{n_p\, n_i}\right] MS(p) = \frac{MS(p)}{n_i}, \tag{23}$$

which clearly means that the denominator is unbiased with respect to person sampling. To prove the equality in Equation 22 note that $\hat{\sigma}^2(p) = [MS(p) - MS(pi)]/n_i$, and $\hat{\sigma}^2(\delta) = MS(pi)/n_i$. It follows that

$$\hat{\sigma}^2(p) + \hat{\sigma}^2(\delta) = \frac{MS(p) - MS(pi)}{n_i} + \frac{MS(pi)}{n_i} = \frac{MS(p)}{n_i}, \tag{24}$$

which is identical to Equation 23.

Thus, $\hat{\rho}_{20}$ can be expressed as

$$\hat{\rho}_{20} = \frac{\hat{\sigma}^2(\overline{X}_p) - \hat{\sigma}^2(\delta)}{\hat{\sigma}^2(\overline{X}_p)}. \tag{25}$$

In a similar manner, $\hat{\rho}_{21}$ can be expressed as

$$\hat{\rho}_{21} = \frac{S^2(\overline{X}_p) - \hat{\sigma}^2(\Delta)}{S^2(\overline{X}_p)}. \tag{26}$$

Even though Equation 26 appears to have the form of a reliability coefficient (i.e., a ratio of true score variance to observed score variance), it is based on a curious mixture of unbiased and biased estimates—namely, $S^2(\overline{X}_p)$ is biased with respect to person sampling, while $\hat{\sigma}^2(\Delta)$ is unbiased with respect to both person and item sampling. This results in true score variance being estimated as $\hat{\sigma}^2(p) - \hat{\sigma}^2(\overline{X})$, rather than $\hat{\sigma}^2(p)$, as discussed in the proof of Equation 2.

It is clear from Equations 2 and 3 that when $n_p \to \infty$,

$$\hat{\rho}_{21} = \frac{\hat{\sigma}^2(p) - \dfrac{\hat{\sigma}^2(i)}{n_i}}{\left[\hat{\sigma}^2(p) - \dfrac{\hat{\sigma}^2(i)}{n_i}\right] + \left[\dfrac{\hat{\sigma}^2(i)}{n_i} + \dfrac{\hat{\sigma}^2(pi)}{n_i}\right]} \tag{27}$$

$$= \frac{\hat{\sigma}^2(p) - \hat{\sigma}^2(i)/n_i}{\hat{\sigma}^2(p) + \hat{\sigma}^2(\delta)}. \tag{28}$$

Equation 28 may give the impression that the error variance in $\hat{\rho}_{21}$ is relative error variance. However, a reliability-like coefficient always has the form of universe score variance divided by itself plus error variance. Equation 28 does not have this form, whereas Equation 27 does, and Equation 27 clearly indicates that the error variance in $\hat{\rho}_{21}$ is $\hat{\sigma}^2(\Delta)$. Note also that when $n_p \to \infty$, if $\hat{\sigma}^2(i)$ is small and/or n_i is large, then $\hat{\rho}_{21}$ will be close to $\hat{\rho}_{20}$.

$\hat{\rho}_{21}$ was developed by Kuder and Richardson (1937) for the case in which items are all equally difficult. When items are not all equally difficult, however, Equation 27 indicates that item variability decreases the estimate of universe score variance and increases the estimate of error variance. Except for computational ease, there is not much reason to recommend using $\hat{\rho}_{21}$ when items are unequally difficult, except for situations in which one wants $\widehat{\Phi}(\lambda = \overline{X}) = \hat{\rho}_{21}$.

The estimated dependability index $\widehat{\Phi}$ is

$$\widehat{\Phi} = \frac{\hat{\sigma}^2(p)}{\hat{\sigma}^2(p) + \hat{\sigma}^2(\Delta)},$$

which equals $\hat{\rho}_{21}$ in Equation 2 when $\hat{\sigma}^2(\overline{X}) = 0$ (see Equation 3). With real data, often $\hat{\sigma}^2(\overline{X})$ is quite small. In such cases, $\hat{\rho}_{21}$ tends to be close to $\widehat{\Phi}$. However, when variance components are positive, strictly speaking $\hat{\sigma}^2(\overline{X}) = 0$ only when $n_p \to \infty$ and $n_i \to \infty$. In short, the fact that $\hat{\rho}_{21}$ and $\widehat{\Phi}$ are usually close to equal should not be taken as an indication that they have the same theoretical foundation, although they do share the same estimated error variance, $\hat{\sigma}^2(\Delta)$.

References

Brennan, R. L. (2001). *Generalizability theory.* New York: Springer-Verlag.

Brennan, R. L., & Kane, M. T. (1977a). An index of dependability for mastery tests. *Journal of Educational Measurement*, *14*, 277–289.

Brennan, R. L., & Kane, M. T. (1977b). Signal/noise ratios for domain-referenced tests. *Psychometrika*, *42*, 609–625.

Buros, O. K. (1963). *Summary of definitions, basic concepts, and formulas on the schematization of test reliability.* Unpublished manuscript originally available from Gryphon Press, Highland Park, NJ.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 292–334.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles.* New York: Wiley.

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.) (pp. 105–146). New York: American Council on Education and Macmillan.

Huynh, H. (1976). On the reliability of decisions in domain-referenced testing. *Journal of Educational Measurement, 13*, 253–264.

Keats, J. A. (1957). Estimation of error variances of test scores. *Psychometrika, 22*, 29–41.

Kuder, G. F. & Richardson, M. W. (1937). The theory of estimation of test reliability. *Psychometrika*, *2*, 151–160.

Lord, F. M. (1955). Estimating test reliability. *Educational and Psychological Measurement*, *15*, 325–336.

Lord, F. M. (1957). Do tests of the same length have the same standard error of measurement? *Educational and Psychological Measurement*, *17*, 510–521.