Statistical Inference for *G*-indices of Agreement

Douglas G. Bonett

University of California, Santa Cruz

Journal of Educational and Behavioral Statistics

2022

Statistical Inference for *G*-indices of Agreement

Abstract

The limitations of Cohen's kappa are reviewed and an alternative *G*-index is recommended for assessing nominal scale agreement. Maximum likelihood estimates, standard errors, and confidence intervals for a two-rater *G*-index are derived for one-group and two-group designs. A new *G*-index of agreement for multi-rater designs is proposed. Statistical inference methods for some important special cases of the multi-rater design also are derived. *G*-index meta-analysis methods are proposed and can be used to combine and compare agreement across two or more populations. Closed-form sample size formulas to achieve desired confidence interval precision are proposed for two-rater and multi-rater designs. R functions are given for all results.

Keywords: *nominal agreement, multi-rater agreement, interval estimation, meta-analysis, sample size planning*

1. Introduction

The assignment of mutually exclusive nominal ratings to a set of objects by different raters is a very common activity in many fields of research. For example, suppose two deliberately selected raters assign every object in a population of objects to one and only one of *r* clearly defined nominal categories. The ratings can then be summarized in an *r* x *r* contingency table of population proportions where π_{ij} is the population proportion of objects assigned to category *i* (*i* = 1 to *r*) by Rater 1 and category *j (j* = 1 to *r)* by Rater 2, π_{i+} is the population proportion of objects assigned to category *i* (*i* = 1 to *r*) by Rater 1, and π_{+i} is the population proportion of objects assigned to category *i* (*i* = 1 to *r*) by Rater 2.

If the population of objects is large or if the rating process is time consuming or costly, it may be necessary to obtain a random sample of *n* objects from the population and have each rater classify the sample of *n* objects. When two raters each classify a random sample of *n* objects, the resulting classifications can be summarized into an *r* x *r* table of observed frequency counts where f_{ij} is the number of sample objects assigned to category *i* (*i* = 1 to *r*) by Rater 1 and category *j* (*j* = 1 to *r*) by Rater 2, f_{i+} is the number of sample objects assigned to category *i* (*i* = 1 to *r*) by Rater 1, and f_{+i} is the number of sample objects assigned to category *i* (*i* = 1 to *r*) by Rater 2. The observed frequency counts can be used to estimate the population proportions in the *r* x *r* contingency table. Specifically, the maximum likelihood estimates of π_{ij}, π_{i+}, and π_{+j} are f_{ij}/n, f_{i+}/n, and f_{+i}/n, respectively.

In the two-rater design, one measure of agreement is the proportion of objects that are assigned to the same nominal category by both raters. This proportion is $\pi_A = \sum_{i=1}^{r} \pi_{ii}$ and its maximum likelihood estimate is $\hat{\pi}_A = \sum_{i=1}^{r} f_{ii}/n$.

2. Chance-corrected Measures of Agreement

As a measure of agreement, π_A exaggerates the degree of agreement between two raters because π_{ii} can be greater than 0 even if both raters classify the objects in a purely random manner. A general chance-corrected measure of agreement (Scott, 1955) is defined as

$$\kappa = \frac{\pi_A - \pi_R}{1 - \pi_R} \quad (1)$$

where π_R is the proportion of agreements that would be expected if the two raters had assigned the classifications independently and at random. The maximum value of κ is 1, representing perfect agreement, and the minimum value of κ is $-\pi_R/(1 - \pi_R)$.

Different definitions of π_R define different chance-corrected measures of agreement. Scott (1955) sets $\pi_R = \sum_{i=1}^{r} \pi_i{}^2$, Bennett, Alpert and Goldstein (1954) set $\pi_R = 1/r$, and Gwet (2008) sets $\pi_R = \sum_{i=1}^{r} \pi_i(1 - \pi_i)/(r - 1)$ where $\pi_i = (\pi_{+i} + \pi_{i+})/2$. The chance-corrected measure of agreement proposed by Krippendorff (1970) also sets $\pi_R = \sum_{i=1}^{r} \pi_i{}^2$ but makes a small-sample adjustment to the estimate of π_A. For $r = 2$, Scott's measure of agreement is also called intraclass kappa (Shoukri, 2011). The most well-known and widely used chance-corrected measure of agreement in the social sciences is Cohen's kappa (Cohen, 1960) where $\pi_R = \sum_{i=1}^{r} \pi_{+i}\pi_{i+}$.

The population value of Cohen's kappa is denoted here as κ_c. A maximum likelihood estimate of κ_c, denoted as $\hat{\kappa}_c$, is obtained by replacing π_A with $\hat{\pi}_A = \sum_{i=1}^{r} \frac{f_{ii}}{n}$ and replacing π_R with $\hat{\pi}_R = \frac{\sum_{i=1}^{r} f_{+i} f_{i+}}{n^2}$ in Equation 1. An approximate large-sample standard error of $\hat{\kappa}_c$ was derived by Fleiss, Cohen, and Everitt (1969) which can be used to construct a Wald confidence interval for κ_c.

SAS and Stata compute the Wald confidence interval for κ_c. The `cohen.kappa` function in the `psych` R package also computes the Wald confidence interval for κ_c. The Wald confidence

interval for Cohen's kappa (κ_C) is known to have poor performance characteristics (Flack, 1987; Blackman & Koval, 2000). Although several alternative confidence interval estimation methods for κ_c have been proposed (see Lee & Tu, 1994), bootstrap confidence intervals for κ_c appear to be the best option (Klar et al., 2002). Lee and Tu (1994) found that a sample of at least n = 100 objects must be rated before the Wald confidence interval for κ_c can be expected to perform properly. The simulation results of Klar et al. (2002) suggest that the bootstrap confidence interval for κ_c should not be used for $n < 35$.

Cohen's kappa is widely used but has several limitations. Feinstein and Cicchetti (1990) explain how $\hat{\kappa}_c$ can have a very low value even when $\hat{\pi}_A$ is close to 1. For example, with f_{11} = 97, f_{12} = 0, f_{21} = 2, and f_{22} = 1, there is near perfect agreement between the two raters but $\hat{\kappa}_c$ = .492. With f_{11} = 80, f_{12} = 20, f_{21} = 0, and f_{22} = 0 there is high agreement between the two raters but $\hat{\kappa}_c$ = 0. Warrens (2010) shows that $\hat{\kappa}_c$ is paradoxically larger when $\pi_{i+} \neq \pi_{+i}$. For example, $\hat{\kappa}_c$ = .40 with f_{11} = 35, f_{12} = 15, f_{21} = 15, and f_{22} = 35 where $\pi_{i+} = \pi_{+i}$, but $\hat{\kappa}_c$ paradoxically increases to .45 with f_{11} = 35, f_{12} = 0, f_{21} = 30, and f_{22} = 35 where $\hat{\pi}_A$ is unchanged but $\pi_{i+} \neq \pi_{+i}$. Maclure and Willett (1987) argue that $\hat{\kappa}_c$ should not be use with $r > 2$.

Cohen's definition of $\pi_R = \sum_{i=1}^{r} \pi_{+i}\pi_{i+}$ is perhaps its most serious weakness as a measure of nominal agreement. This definition implies that Rater 1 would assign objects to category *i* with probability π_{+i} and Rater 2 would assign subjects to category *i* with probability π_{i+} if both raters were simply guessing. However, $\pi_R = \sum_{i=1}^{r} \pi_{+i}\pi_{i+}$ is estimated using sample data in which raters typically are not guessing. For example, if r = 3 and Rater 1 classifies 20% of the objects in category 1, 45% in category 2, and 35% in category 3, Cohen's kappa assumes that these exact same marginal proportions would be obtained if Rater 1 was simply guessing. Green (1981) argued that $\pi_R = \sum_{i=1}^{r} \pi_{+i}\pi_{i+}$ could be justified in the arguably unrealistic situation where a rater would

"guess at a rate equivalent to the proportion of time they have determined the presence of a characteristic when they were not guessing". Except for the chance-corrected measure of agreement proposed by Bennett, Alpert and Goldstein (1954), all of the other measures of agreement also use an estimate of π_R in situations where raters are typically not guessing. Furthermore, estimating $\pi_R = \sum_{i=1}^{r} \pi_{+i}\pi_{i+}$ introduces another source of sampling variability into $\hat{\kappa}_c$ which degrades the small-sample performance of the Wald confidence interval for κ_c.

For r = 2, Byrt, Bishop, and Carlin (1993) define prevalence as $\pi_{11} - \pi_{22}$ and bias as $\pi_{i+} - \pi_{+i}$ and then show that a bias adjustment to κ_c is equal to the intraclass kappa. They also show that a bias adjustment combined with a prevalence adjustment to κ_c is equal to the Bennett-Alpert-Goldstein index. For r = 2, Blackman and Koval (1993) show that intraclass kappa is a large-sample approximation to an intraclass reliability of a single rater from a one-way ANOVA, while Cohen's kappa is a large-sample approximation to an intraclass reliability of a single rater from a two-way ANOVA. In an interrater reliability study, the intraclass reliability coefficient describes the reliability of a single rater assuming parallel measurements (McDonald, 1999). Parallel measurements are assumed to be homoscedastic and this assumption is violated in the case of r = 2 if $\pi_{i+}(1 - \pi_{i+}) \neq \pi_{+i}(1 - \pi_{+i})$. Although Cohen's kappa is arguably an inappropriate measure of interrater agreement, it is an appropriate measure of interrater reliability in the special case of r = 2 and homoscedasticity.

Some of the controversy and debate regarding interrater agreement stems from a failure to distinguish between interrater agreement and interrater reliability (Kottner & Stiener, 2011). One of the claimed limitations of κ_c is the attenuation that occurs in a 2 x 2 table when π_{+1} or π_{1+} is close to 1 or 0. Although the attenuation of $\hat{\kappa}_c$ is an inappropriate characteristic for a measure of interrater agreement, it is a perfectly appropriate characteristic for a measure of interrater reliability

because reliability cannot be large if there is little variability in the ratings, and the variability of the dichotomous ratings will be small when $\hat{\pi}_{+1}$ or $\hat{\pi}_{1+}$ is close to 1 or 0. The problem of assessing interrater agreement rather than interrater reliability is addressed here.

3. *G*-index for Two Raters

As noted above, the chance-corrected measure of agreement proposed by Bennett, Alpert and Goldstein (1954) sets $\pi_R = 1/r$. This value for π_R assumes two independent raters, if they were simply guessing, would select one of the r categories with probability $1/r$ so that the joint probabilities in the r x r contingency table under random and independent ratings is $1/r^2$. The sum of the r joint agreement probabilities gives $\pi_R = 1/r$. This definition of π_R is a sensible assumption for random nominal scale classifications and is consistent with signal detection theory (Wickens, 2002, p. 95) when choosing among r alternatives under a pure noise condition. Setting $\pi_R = 1/r$ is also consistent with the conceptualization of chance agreement proposed by Lawlis and Lu (1972), Maxwell (1977), and Grove et al. (1981). Hayes and Krippendorff (2007) criticize using $\pi_R = 1/r$. They argue that any category that is unused by both raters will "inflate" the Bennet-Alpert-Goldstein coefficient but not Cohen's kappa. However, a category that is unused by both raters indicates perfect agreement for that category and it is appropriate that the Bennet-Alpert-Goldstein coefficient reflects this agreement.

The population chance-corrected measure of agreement proposed by Bennett, Alpert and Goldstein (1954) and Brennan and Prediger (1981), and referred to by Holley and Guilford (1964) as a *G*-index of agreement, will be denoted here as κ_G. Setting $\pi_R = 1/r$ in Equation 1 gives

$$\kappa_G = \frac{\pi_A - 1/r}{1 - 1/r} = (r\pi_A - 1)/(r - 1) \tag{2}$$

and has a possible range of $-1/(r-1)$ to 1. The maximum likelihood estimate of κ_G is

$$\hat{\kappa}_G = (r\hat{\pi}_A - 1)/(r - 1) \tag{3}$$

where $\hat{\pi}_A = \sum_{i=1}^{r} f_{ii}/n$ is the unbiased maximum likelihood estimate of π_A. The estimator of κ_G is a unbiased maximum likelihood estimator because it is a linear function of the unbiased maximum likelihood estimator of π_A. An approximate standard error for $\hat{\kappa}_G$ given below

$$SE(\hat{\kappa}_G) = [r/(r-1)]\sqrt{\hat{\pi}_A(1-\hat{\pi}_A)/n} \quad (4)$$

which is a linear function of the standard error of $\hat{\pi}_A$. In the above example where f_{11} = 97, f_{12} = 0, f_{21} = 2, and f_{22} = 1, the estimate of κ_G (Equation 3) is .96 and appropriately describes the near perfect agreement between the two raters.

The following 100(1 – α)% adjusted Wald confidence interval for κ_G is proposed here

$$[r/(r-1)]\left[\hat{\pi}_A^* \pm z_{\alpha/2}\sqrt{\hat{\pi}_A^*(1-\hat{\pi}_A^*)/(n+4)}\right] - 1/(r-1) \quad (5)$$

where $\hat{\pi}_A^* = (f_A + 2)/(n+4)$ and $f_A = \sum_{i=1}^{r} f_{ii}$. The `ci.qrater` R function in the online Supplementary Materials computes Equations 3, 4 and 5.

Although a point estimate of κ_G was proposed decades ago (Bennett, Alpert, & Goldstein, 1954), little progress has been made in terms of statistical inference for κ_G. For the special case of r = 2, a standard error of $\hat{\kappa}_G$ is given in Shoukri (2011) as

$$(1 - \hat{\kappa}_G^2)/n \quad (6)$$

which can be algebraically re-expressed in the form of Equation 4. For r = 2, this standard error could be used to compute the following 100(1 – α)% Wald confidence interval for κ_G

$$\hat{\kappa}_G \pm z_{\alpha/2} SE(\hat{\kappa}_G). \quad (7)$$

It can be shown that Equation 7 can be expressed in the form of Equation 5 where the interval estimate in brackets is the traditional Wald confidence interval for a population proportion. Agresti and Coull (1998) showed that the traditional Wald confidence interval has poor performance characteristics under realistic conditions and they proposed an adjusted Wald confidence interval.

With sample sizes as small as n = 10, the 95% Agresti-Coull confidence interval had an average coverage probability close to .95 and a worst-case coverage probability no less than .92 across the entire range of possible population proportion values (Agresti & Coull, 1998). The interval estimate in brackets of Equation 7 is the Agresti-Coull confidence interval and hence Equation 5 inherits all of its performance characteristics because κ_G is a linear function of π_A. The "exact" Clopper-Pearson confidence interval for π_A also could be used in place of the Agresti-Coull confidence interval in Equation 5 (see Conclusion section). Note that Cohen's kappa cannot be expressed solely in terms of π_A and hence the Agresti-Coull or Clopper-Pearson confidence intervals cannot be used to obtain a confidence interval for Cohen's kappa.

4. Comparing Agreement in Two-group Designs

Assessing interrater agreement from two independent groups can answer a wide variety of interesting research questions. The two-group design can be experimental or nonexperimental. A two-group nonexperimental design could consist of two types of randomly sampled objects (e.g., male vs female students) that are classified into r categories by the same two raters. In a two-group experimental design, a single sample of objects is randomly divided into two groups and each group could be rated by different types of raters (e.g., expert vs novice) or under differing rating conditions (e.g., complete vs incomplete case files). Methods for comparing two or more independent intraclass kappa values have been proposed by Donner, Eliasziw, and Klar (1996) and Donner and Zou (2002).

Let κ_{Gj} denote the population value of the G-index that will be estimated from subpopulation j in a two-group nonexperimental design or from condition j in a two-group experimental design. It is easy to show that the difference $\kappa_{G1} - \kappa_{G2}$ can be expressed as $[r/(r-1)](\pi_{A1} - \pi_{A2})$ where π_{Aj} is the population proportion of agreements that will be estimated

in group j. The following 100(1 – α)% adjusted Wald confidence interval for $\kappa_{G1} - \kappa_{G2}$ is proposed here

$$[r/(r-1)]\left[\hat{\pi}^*_{A1} - \hat{\pi}^*_{A2} \pm z_{\alpha/2}\sqrt{\frac{\hat{\pi}^*_{A1}(1-\hat{\pi}^*_{A1})}{n_1+2} + \frac{\hat{\pi}^*_{A2}(1-\hat{\pi}^*_{A2})}{n_2+2}}\right] \quad (8)$$

where $\hat{\pi}^*_{Aj} = (f_{Aj}+1)/(n_j+2)$. Note that the confidence interval in brackets is the adjusted Wald confidence interval for a difference between two independent proportions developed by Agresti and Caffo (2000) which has been shown to have excellent small-sample properties under a wide range of conditions. With sample sizes as small as 10 per group, the 95% Agresti-Caffo confidence interval had an average coverage probability close to .95 and a worst-case coverage probability no less than .92 in the 10,000 conditions they considered (Agresti & Caffo, 2000). The `ci.diff` R function in the online Supplementary Materials computes Equation 6.

5. Multi-study Designs

A small sample of objects might be necessary if the ratings are costly or time consuming. However, a confidence interval for κ_G might be uselessly wide if the sample size is too small. One way to obtain a more accurate estimate of κ_G is to statistically combine agreement estimates from two or more independent studies. Combining parameter estimates from two or more studies is called a meta-analysis, and Vacha-Haase (1998) referred to a meta-analysis of reliability estimates as reliability generalization. Bonett (2010) developed statistical methods for combining and comparing Cronbach alpha reliability coefficients from two or more studies. The logic and rationale for reliability generalization also applies to interrater agreement with the goal of obtaining a more precise and generalizable estimate of agreement and also to assess the degree to which an agreement index might vary across different types of raters, different rating conditions, or different types of rated objects.

Sun (2011) showed how the classical fixed-effect and random-effects meta-analysis methods (see Borenstein et al., 2009) can be used to obtain a confidence interval for an average of two or more population Cohen kappa values. The varying-coefficient meta-analysis method (Bonett & Price, 2015) is used here to obtain a confidence interval for an average of *G*-index values from multiple populations. The varying-coefficient model does not require the unrealistic assumptions of the classical fixed-effect and random-effects meta-analysis methods. Unlike the traditional fixed-effect model, the varying-coefficient model does not assume effect-size homogeneity, and unlike the random-effects model the varying-coefficient model does not assume randomly selected studies or effect size standard errors that are uncorrelated with the effect sizes (see Bonett & Price, 2015 for more details).

Let κ_{Gj} represent the population *G*-index value that has been estimated in study *j* (j = 1 to m).The following 100(1 – α)% adjusted Wald confidence interval for $\sum_{j=1}^{m}\kappa_{Gj}/m$ is proposed here

$$\sum_{j=1}^{m}\hat{\kappa}_{Gj}^{*}/m \pm z_{\alpha/2}\sqrt{\sum_{j=1}^{m}SE(\hat{\kappa}_{Gj}^{*})^{2}/m^{2}} \qquad (9)$$

where $\hat{\kappa}_{Gj}^{*} = (r\hat{\pi}_{Aj}^{*} - 1)/(r-1)$, $SE(\hat{\kappa}_{Gj}^{*}) = [r/(r-1)]\sqrt{\frac{\hat{\pi}_{Aj}^{*}(1-\hat{\pi}_{Aj}^{*})}{n_j + 4/m}}$, and $\hat{\pi}_{Aj}^{*} = (f_{Aj} + 2/m)/(n_j + 4/m)$. Note that $\sum_{j=1}^{m}\kappa_{Gj}/m = [r/(r-1)]\sum_{j=1}^{m}\pi_{Aj}/m - m/(r-1)$ and hence the endpoints of Equation 9 are linear functions of the endpoints of the adjusted Wald confidence interval proposed by Price and Bonett (2004) for a linear function of independent proportions. The Price-Bonett confidence interval is a generalization of the Agesti-Coull confidence interval and has been shown to have excellent small-sample properties under a wide range of conditions. In meta-analysis applications with m = 5, 15, or 30 studies and sample sizes as small as 10 per group, the 95% Price-Bonett confidence interval had an average coverage probability close to .95 and a worst-case

coverage probability no less than .938 in the 13,500 conditions they considered (Price & Bonett, 2004). The `ci.meta` R function in the online Supplementary Materials computes Equation 9 using the sample sizes and number of agreements from each study.

Important differences in κ_{Gj} values across the *m* populations could be due to differences in rating conditions, rater characteristics, or characteristics of the rated objects. A linear contrast of κ_{Gj} values can be expressed as $\sum_{j=1}^{m} h_j \kappa_{Gj}$ where the h_j values are specified by the researcher and $\sum_{j=1}^{m} h_j = 0$. For example, in a meta-analysis of $m = 5$ comparable interrater agreement studies where the first three studies rated the behavior of high school students and the last two studies rated the behavior of college students, the researcher might want to estimate $(\kappa_{G1} + \kappa_{G2} + \kappa_{G3})/3 - (\kappa_{G4} + \kappa_{G5})/2$. This linear contrast can be specified with contrast coefficients $h_1 = 1/3$, $h_2 = 1/3$, $h_3 = 1/3$, $h_4 = -1/2$, and $h_5 = -1/2$.

The following 100(1 – α)% adjusted Wald confidence interval for $\sum_{j=1}^{m} h_j \kappa_{Gj}$ is proposed here

$$\sum_{j=1}^{m} h_j \hat{\kappa}_{Gj}^{*} \pm z_{\alpha/2} \sqrt{\sum_{j=1}^{m} h_j^2 SE(\hat{\kappa}_{Gj}^{*})^2} \tag{10}$$

where $\hat{\kappa}_{Gj}^{*}$ and $SE(\hat{\kappa}_{Gj}^{*})$ are defined above with *m* set equal to the number of non-zero h_j values as recommended by Price and Bonett (2004). Note that $\sum_{j=1}^{m} h_j \kappa_{Gj} = [r/(r-1)] \sum_{j=1}^{m} h_j \pi_{Aj}$ and hence the endpoints of Equation 10 are linear functions of the endpoints of the adjusted Wald confidence interval proposed by Price and Bonett (2004) for a linear function of independent proportions. The Price-Bonett confidence interval has been shown to have excellent performance characteristics. With sample sizes as small as 20 per group, the 95% Price-Bonett confidence interval had an average coverage probability close to .95 and a worst-case coverage probability no less than .920 across 10,000 population proportion values and several different types of linear contrasts (Price &

Bonett, 2004). The `ci.contrast` R function in the online Supplementary Materials computes Equation 10.

6. Multi-rater Designs

If q different and deliberately selected raters each classify a random sample of objects into r categories, the results can be summarized in a r^q contingency table. A *G*-index of agreement for any two raters can be computed by collapsing the r^q table into an r x r table for the two raters of interest and then applying Equations 3, 4, and 5.

A *G*-index of agreement among all q raters is defined here as

$$\kappa_{G(q)} = \frac{\pi_A - \pi_R}{1 - \pi_R} = \frac{\pi_A - 1/r^{q-1}}{1 - 1/r^{q-1}} = (r^{q-1}\pi_A - 1)/(r^{q-1} - 1) \quad (11)$$

where $\pi_A = \sum_{i=1}^{r} \pi_{ii\ldots i}$ is the probability of unanimous agreement among the q raters and $\pi_{ii\ldots i}$ is the probability of unanimous agreement among the q raters for one of the r categories. Assuming random and independent ratings, the joint probabilities in the r^q contingency table is $1/r^q$ so that the probability of random agreement in any of the r categories is $\pi_R = \sum_{i=1}^{r} 1/r^q = 1/r^{q-1}$. Note that π_A satisfies the definition of agreement among multiple raters given by Hubert (1977).

The maximum likelihood estimate of $\kappa_{G(q)}$ is

$$\hat{\kappa}_{G(q)} = (r^{q-1}\hat{\pi}_A - 1)/(r^{q-1} - 1) \quad (12)$$

where $\hat{\pi}_A = \sum_{i=1}^{r} f_{ii\ldots i}/n$ is the maximum likelihood estimate of π_A. The approximate standard error for $\hat{\kappa}_{G(q)}$ given below is a function of the variance of $\hat{\pi}_A$.

$$SE(\hat{\kappa}_{G(q)}) = [r^{q-1}/(r^{q-1} - 1)]\sqrt{\hat{\pi}_A(1 - \hat{\pi}_A)/n} \quad (13)$$

The following 100(1 – α)% adjusted Wald confidence interval for $\kappa_{G(q)}$ is proposed here

$$[r^{q-1}/(r^{q-1} - 1)]\left[\hat{\pi}_A^* \pm z_{\alpha/2}\sqrt{\hat{\pi}_A^*(1 - \hat{\pi}_A^*)/(n+4)}\right] - 1/(r^{q-1} - 1) \quad (14)$$

where $\hat{\pi}_A^* = (f_A + 2)/(n + 4)$ and $f_A = \sum_{i=1}^{r} f_{ii\ldots i}$. Like Equation 5, Equation 14 uses the Agresti-Coull adjusted Wald confidence interval for a single population proportion. The `ci.qrater` R function in the online Supplementary Materials computes Equation 14.

6.1 Three-rater Design (r = 2)

Consider a $q = 3$ rater design with $r = 2$. The three pairwise *G*-indices are denoted as $\kappa_{G\{1,2\}}$, $\kappa_{G\{1,3\}}$, and $\kappa_{G\{2,3\}}$. The average of all pairwise measures of agreement is an alternative multi-rater measure of agreement proposed by Hubert (1977). It can be shown for the special case of $q = 3$, using straightforward but tedious algebra, that $\kappa_{G(3)} = (\kappa_{G\{1,2\}} + \kappa_{G\{1,3\}} + \kappa_{G\{2,3\}})/3$.

In some applications it will be informative to compare $\kappa_{G\{1,2\}}$, $\kappa_{G\{1,3\}}$, and $\kappa_{G\{2,3\}}$. For example, if Raters 1 and 2 are novices and Rater 3 is an expert, confidence intervals for $\kappa_{G\{1,2\}} - \kappa_{G\{1,3\}}$, $\kappa_{G\{1,2\}} - \kappa_{G\{2,3\}}$, and $\kappa_{G\{1,3\}} - \kappa_{G\{2,3\}}$ will provide information regarding the direction and the magnitude of these pairwise differences. The three *G*-indices in a 3-rater design with $r = 2$ can be expressed as

$$\kappa_{G\{1,2\}} = 2(\pi_{111} + \pi_{112} + \pi_{221} + \pi_{222}) - 1 \tag{15}$$

$$\kappa_{G\{1,3\}} = 2(\pi_{111} + \pi_{121} + \pi_{212} + \pi_{222}) - 1 \tag{16}$$

$$\kappa_{G\{2,3\}} = 2(\pi_{111} + \pi_{211} + \pi_{122} + \pi_{222}) - 1 \tag{17}$$

and it follows that the pairwise differences in these *G*-indices can be expressed as

$$\kappa_{G\{1,2\}} - \kappa_{G\{1,3\}} = 2(\pi_{112} + \pi_{221}) - 2(\pi_{121} + \pi_{212}) \tag{18}$$

$$\kappa_{G\{1,2\}} - \kappa_{G\{2,3\}} = 2(\pi_{112} + \pi_{221}) - 2(\pi_{211} + \pi_{122}) \tag{19}$$

$$\kappa_{G\{1,3\}} - \kappa_{G\{2,3\}} = 2(\pi_{121} + \pi_{212}) - 2(\pi_{211} + \pi_{122}) \tag{20}$$

Maximum likelihood estimates of $\kappa_{G\{1,2\}}$, $\kappa_{G\{1,3\}}$, and $\kappa_{G\{2,3\}}$ and the pairwise differences in

G-indices are obtained by replacing π_{ijk} in Equations 15-20 with the maximum likelihood estimate $\hat{\pi}_{ijk} = f_{ijk}/n$.

An approximate standard error of each pairwise difference is derived by first collapsing the 2^3 table of multinomial proportions into three mutually exclusive categories. For example, from Equation 18 the three categories needed to assess $\kappa_{G\{1,2\}} - \kappa_{G\{1,3\}}$ have probabilities of $\pi_1 = \pi_{112} + \pi_{221}$, $\pi_2 = \pi_{121} + \pi_{212}$, and $\pi_3 = 1 - (\pi_1 + \pi_2)$. Using the variances and covariances of a multinomial distribution (Bishop, Finberg, & Holland, 1976, p. 442), an approximate standard error of $\hat{\kappa}_{G\{1,2\}} - \hat{\kappa}_{G\{1,3\}}$ is

$$SE(\hat{\kappa}_{G\{1,2\}} - \hat{\kappa}_{G\{1,3\}}) = \sqrt{4\{\hat{\pi}_1 + \hat{\pi}_2 - (\hat{\pi}_1 - \hat{\pi}_2)^2\}/n} \quad (21)$$

where $\hat{\pi}_1 = (f_{112} + f_{221})/n$ and $\hat{\pi}_2 = (f_{121} + f_{212})/n$. An approximate standard error of $\hat{\kappa}_{G\{1,2\}} - \hat{\kappa}_{G\{2,3\}}$ is given by Equation 21 with $\hat{\pi}_1 = (f_{112} + f_{221})/n$ and $\hat{\pi}_2 = (f_{211} + f_{122})/n$, and an approximate standard error of $\hat{\kappa}_{G\{1,3\}} - \hat{\kappa}_{G\{2,3\}}$ is given by Equation 21 with $\hat{\pi}_1 = (f_{121} + f_{212})/n$ and $\hat{\pi}_2 = (f_{211} + f_{122})/n$.

A 100(1 – α)% adjusted Wald confidence interval for $\kappa_{G\{1,2\}} - \kappa_{G\{1,3\}}$ is

$$2[\hat{\pi}_1^* - \hat{\pi}_2^* \pm z_{\alpha/2}\sqrt{\{\hat{\pi}_1^* + \hat{\pi}_2^* - (\hat{\pi}_1^* - \hat{\pi}_2^*)^2\}/(n+2)}] \quad (22)$$

where $\hat{\pi}_1^* = (f_{112} + f_{221} + 1)/(n+2)$ and $\hat{\pi}_2^* = (f_{121} + f_{212} + 1)/(n+2)$. The confidence interval for $\kappa_{G\{1,2\}} - \kappa_{G\{2,3\}}$ is given by Equation 22 with $\hat{\pi}_1^* = (f_{112} + f_{221} + 1)/(n+2)$ and $\hat{\pi}_2^* = (f_{121} + f_{212} + 1)/(n+2)$, and the confidence interval for $\kappa_{G\{1,2\}} - \kappa_{G\{2,3\}}$ is given by Equation 22 with $\hat{\pi}_1^* = (f_{121} + f_{212} + 1)/(n+2)$ and $\hat{\pi}_2^* = (f_{211} + f_{122} + 1)/(n+2)$. The adjusted Wald confidence interval in brackets of Equation 22 was developed by Bonett and Price (2012) and was shown to have excellent small-sample properties under a wide range of conditions. With sample sizes as small as *n* = 15, the 95% Bonett-Price confidence interval had an average

coverage probability that was slightly greater than .95 and a worst-case coverage probability no less than .910 in the 25,000 conditions they considered (Bonett & Price, 2012). The `agree.3rater` R function in the online Supplementary Materials computes Equation 22 for all three pairs of raters and also computes Equation 14 for $\kappa_{G(3)}$.

6.2 Four-rater Design (r = 2)

The case of $q = 4$ raters with $r = 2$ has received special attention in the literature (Donner et al, 2000; McKenzie et al., 1996; Williamson & Manatunga, 1997; Banerjee, Capozzoli, McSweeney & Sinha, 1999) because some of the research questions that can be answered using the two-group design described previously might be answered more economically using a one-group design with four raters. The four rater design can be used to compare the agreement between two different types of raters such as two novice raters and two expert raters or two male raters and two female raters.

A *G*-index for any two raters in a 4-rater design can be computed by collapsing the 2^4 table into a 2 x 2 table for any two raters of interest and then applying Equations 3, 4, and 5. The focus here will be the comparison of agreement between the first two raters ($\kappa_{G\{1,2\}}$) with the last two raters ($\kappa_{G\{3,4\}}$). These two *G*-indices in a 4-rater design can be expressed as

$$\kappa_{G\{1,2\}} = 2(\pi_{1111} + \pi_{1121} + \pi_{1112} + \pi_{1122} + \pi_{2211} + \pi_{2221} + \pi_{2212} + \pi_{2222}) - 1 \quad (23)$$

$$\kappa_{G\{3,4\}} = 2(\pi_{1111} + \pi_{1211} + \pi_{2111} + \pi_{2211} + \pi_{1122} + \pi_{1222} + \pi_{2122} + \pi_{2222}) - 1 \quad (24)$$

and it follows that the difference between these two *G*-indices can be expressed as

$$\kappa_{G\{1,2\}} - \kappa_{G\{3,4\}} = 2(\pi_1 - \pi_2) \quad (25)$$

where $\pi_1 = \pi_{1121} + \pi_{1112} + \pi_{2221} + \pi_{2212}$ and $\pi_2 = \pi_{1211} + \pi_{2111} + \pi_{1222} + \pi_{2122}$. The maximum likelihood estimate of $\kappa_{G\{1,2\}} - \kappa_{G\{3,4\}}$ is obtained by replacing the population proportions in Equations 23 and 24 with their maximum likelihood estimates.

Applying the same approach used to derive Equation 22 gives the following approximate standard error of $\hat{\kappa}_{G\{1,2\}} - \hat{\kappa}_{G\{3,4\}}$

$$SE(\hat{\kappa}_{G\{1,2\}} - \hat{\kappa}_{G\{3,4\}}) = \sqrt{4\{\hat{\pi}_1 + \hat{\pi}_2 - (\hat{\pi}_1 - \hat{\pi}_2)^2\}/n} \quad (26)$$

and the following 100(1 – α)% adjusted Wald confidence interval for $\kappa_{G\{1,2\}} - \kappa_{G\{3,4\}}$

$$2[\hat{\pi}_2^* - \hat{\pi}_2^* \pm z_{\alpha/2}\sqrt{\{\hat{\pi}_1^* + \hat{\pi}_2^* - (\hat{\pi}_1^* - \hat{\pi}_2^*)^2\}/(n+2)}] \quad (27)$$

where $\hat{\pi}_1$ and $\hat{\pi}_2$ are maximum-likelihood estimates, $\hat{\pi}_1^* = (f_{1121} + f_{1112} + f_{2221} + f_{2212} + 1)/(n+2)$ and $\hat{\pi}_2^* = (f_{1211} + f_{2111} + f_{1222} + f_{2122} + 1)/(n+2)$. Note that the adjusted Wald confidence interval in brackets is a Bonett-Price confidence described in Equation 22. The `ci.4rater` R function in the online Supplementary Materials computes the maximum likelihood estimate and confidence interval for $\kappa_{G\{1,2\}} - \kappa_{G\{3,4\}}$ requiring only the sample size, $f_1 = f_{1121} + f_{1112} + f_{2221} + f_{2212}$, and $f_2 = f_{1211} + f_{2111} + f_{1222} + f_{2122}$ as input.

The above results for q = 3 and q = 4 raters can be applied to designs with $q \geq 4$ raters by collapsing a 2^q table into a 2^3 table if the comparison involves three raters or a 2^4 table if the comparison involves four raters. For example, with q = 5, an estimate and confidence interval for $\kappa_{G\{1,3\}} - \kappa_{G\{3,4\}}$ is computed from a 2^3 table for Raters 1, 3, and 4, and an estimate and confidence interval for $\kappa_{G\{1,3\}} - \kappa_{G\{4,5\}}$ is computed from a 2^4 table for Raters 1, 3, 4, and 5.

7. Benchmark *G*-index Values

When reporting the numerical results for a measure of agreement, it is common to also provide a verbal description of the strength of agreement. Altman (1991), Fleiss, Levin, and Paik (2003), and Landis and Koch (1977) have each suggested their own benchmark verbal descriptions for different point estimates of Cohen's kappa. Landis and Koch (1977) suggested that a Cohen kappa value within one of the following six ranges < 0, 0 to .20, .21 to .40, .41 to .60, .61 to .80, or .81 to 1.0 represents a "poor", "slight", "fair", "moderate", "substantial", or "almost perfect"

agreement, respectively. The Landis-Koch benchmarks are frequently used to describe a point estimate of an agreement index. This practice is misleading because point estimates contain sampling error of unknown magnitude and direction. Benchmark descriptions should be applied to an interval estimate rather than a point estimate. Although the Landis-Koch benchmark scale is appealing because of its six very specific descriptive categories, a confidence interval is likely to cover two or more of these categories unless *n* is large. Fleiss, Levin, and Paik (2003) proposed a three category scale for Cohen's kappa where values below .4 represent "poor" agreement, values between .40 and .75 represent "good" agreement, and values greater than .75 represent "excellent" agreement. Some researchers might find the Fleiss scale to be too crude for their purposes.

A four category benchmark scale is proposed here for the *G*-index. A *G*-index value within one of the four ranges < .25, .25 to .50, .51 to .75, and .76 to 1.0 could be described as "poor", "fair", "good", and "excellent", respectively. A confidence interval is less likely to include more than two descriptive categories in a four category scale than a six category scale. For example, a confidence interval of [.581, .824] would be described as "moderate, substantial, or almost perfect agreement" using the Landis-Koch scale and would be described as "good or excellent agreement" using the proposed scale.

8. Sample Size Planning

Sample size planning is one of the most important components in the design of an interrater agreement study. If the number of objects sampled is too small, the confidence interval for the population agreement index could be uselessly wide. Several methods to approximate the sample size requirement when assessing Cohen's kappa have been proposed (Bujang & Baharun, 2017; Cantor, 1996; Donner & Eliasziw, 1992; Flack et al., 1988). The available sample size methods for Cohen's kappa are of limited value because they require assumptions about the classification probabilities of each rater in addition to the value for κ_c. Closed-form sample size formulas are

derived here that approximate the required sample size to obtain $100(1-\alpha)\%$ confidence intervals for κ_G, $\kappa_{G1} - \kappa_{G2}$, and $\kappa_{G(q)}$ with desired precision. These closed-form formulas are particularly useful because they do not require assumptions about the classification probabilities of each rater.

Larger sample sizes give narrower confidence intervals and it is possible to approximate the sample size that will give the desired width (w) of a confidence interval for a specified level of confidence. The sample size needed to obtain a $100(1-\alpha)\%$ confidence interval for κ_G (Equation 5) having a desired width (upper limit minus lower limit) equal to w is approximately

$$n' = 4(\tilde{\kappa}_G + \frac{1}{r-1})(1 - \tilde{\kappa}_G)(z_{\alpha/2}/w)^2 \quad (28)$$

where $\tilde{\kappa}_G$ is a planning value of κ_G. Equation 28 was derived from Equation 4 where $2z_{\alpha/2}SE(\hat{\kappa}_G)$ is the approximate width of Equation 5. Setting this width to w, solving for n, and then replacing $\hat{\kappa}_G$ with $\tilde{\kappa}_G$ gives Equation 28. A planning value of κ_G is obtained from expert opinion, pilot studies, or previously published research. Setting $\tilde{\kappa}_G = (r-2)/[2(r-1)]$ maximizes Equation 28 and is useful in applications where no prior information about κ_G is available.

The width of Equation 5 tends to be greater than $2z_{\alpha/2}SE(\hat{\kappa}_G)$, especially in small samples or if $\hat{\pi}_A$ is close to 0 or 1, and hence n' tends to understate the required sample size. Following the general approach of Bonett and Wright (2000), a more accurate sample size approximation is

$$n = n'(w_0/w)^2 \quad (29)$$

where w_0 is the width of Equation 5 computed using $n = n'$ and the value of $\hat{\pi}_A^*$ implied by $\tilde{\kappa}_G$.The `size.ci.qrater` R function in the online Supplementary Materials computes Equations 28 and then adjusts the result using Equation 29.

The sample size per group needed to obtain a $100(1-\alpha)\%$ confidence interval for $\kappa_{G1} - \kappa_{G2}$ (Equation 8) in a two-group design having a desired width of w is approximately

$$n' = 4v(z_{\alpha/2}/w)^2 \tag{30}$$

where $v = (\tilde{\kappa}_{G1} + \frac{1}{r-1})(1 - \tilde{\kappa}_{G1}) + (\tilde{\kappa}_{G2} + \frac{1}{r-1})(1 - \tilde{\kappa}_{G2})$ and $\tilde{\kappa}_{Gj}$ is a planning value of κ_{Gj}. If no prior information is available, $\tilde{\kappa}_{G1}$ and $\tilde{\kappa}_{G2}$ can be set to (r – 2)/[2(r – 1)] which maximizes Equation 30. Equation 30 was derived by setting $2z_{\alpha/2}\sqrt{SE(\hat{\kappa}_{G1})^2 + SE(\hat{\kappa}_{G2})^2}$ equal to w, solving for n, and replacing $\hat{\kappa}_{Gj}$ with $\tilde{\kappa}_{Gj}$.

The width of Equation 8 tends to be greater than $2z_{\alpha/2}\sqrt{SE(\hat{\kappa}_{G1})^2 + SE(\hat{\kappa}_{G2})^2}$, especially in small samples or if either $\hat{\pi}_{A1}$ or $\hat{\pi}_{A2}$ is close to 0 or 1, and hence n' tends to understate the required sample size per group. The `size.ci.diff` R function in the online Supplementary Materials computes Equations 30 and then adjusts the result using Equation 29.

The sample size needed to obtain a $100(1-\alpha)\%$ confidence interval for $\kappa_{G(q)}$ (Equation 14) having a desired width of w is approximately

$$n' = 4(\tilde{\kappa}_{G(q)} + \frac{1}{r^{q-1} - 1})(1 - \tilde{\kappa}_{G(q)})(z_{\alpha/2}/w)^2 \tag{31}$$

where $\tilde{\kappa}_{G(q)}$ is a planning value of $\kappa_{G(q)}$. Equation 31 was derived from Equation 13 where $2z_{\alpha/2}SE(\hat{\kappa}_{G(q)})$ is the approximate width of Equation 14. Setting this width to w, solving for n, and then replacing $\hat{\kappa}_{G(q)}$ with $\tilde{\kappa}_{G(q)}$ gives Equation 31. Setting $\tilde{\kappa}_{G(q)}$ = (r^{q-1} – 2)/[2(r^{q-1} – 1)] maximizes Equation 31 and is useful in applications where no prior information about $\kappa_{G(q)}$ is available.

The width of Equation 14 tends to be greater than $2z_{\alpha/2}SE(\hat{\kappa}_{G(q)})$, especially in small samples or if $\hat{\pi}_A$ is close to 0 or 1, and hence n' tends to understate the required sample size. The `size.ci.qrater` R function in the online Supplementary Materials computes Equation 31 and then adjusts the result using Equation 29.

9.0 Illustrative Examples

The online Supplementary Materials contains R functions that will compute confidence intervals for: 1) κ_G (Equation 5), 2) $\kappa_{G1} - \kappa_{G2}$ in two-group designs (Equation 8), 3) $\sum_{j=1}^{m} \kappa_{Gj} \,/m$ in meta-analysis applications (Equation 9), 4) a general linear contrast of κ_{Gj} in multiple group designs (Equation 10), 5) $\kappa_{G(q)}$ in multi-rater designs (Equation 14), and 6) pairwise comparisons in 3-rater (Equation 22) and 4-rater designs (Equation 27). The online Supplementary Materials also contain R functions that will compute the sample size requirements to: 1) estimate κ_G with desired precision (Equation 28), 2) estimate $\kappa_{G1} - \kappa_{G2}$ with desired precision (Equation 30), and 3) estimate $\kappa_{G(q)}$ with desired precision (Equation 31). The following examples are adapted from the author's statistical consulting files.

9.1 Two Raters in One-Group Design

Two research assistants classified a random sample of n = 90 open-ended questionnaire responses into r = 3 predetermined categories and 82 of the 90 responses were classified into the same categories by both research assistants. The command `ci.qrater(.05, 90, 82, 3)` computes Equations 3-5 and returns a point estimate for κ_G of .867 with a 95% confidence interval of [.747, .934].

9.2 Two Raters in Two-group Design

Two school psychologists evaluated a random sample of n_1 = 75 case files for boys and a random sample of n_2 = 60 case files for girls. Each psychologist rated each child as having or not having (r = 2) ADHD symptoms. The psychologist ratings agreed in 70 of the 75 cases for the boys and in 45 of the 60 cases for the girls. The command `ci.diff(.05, 75, 70, 60, 45, 2)` computes Equation 8 and returns a 95% confidence interval for the difference in agreement for boys and girls of [.112, .609]. This function also returns point estimates of .867 and .500 with 95%

confidence intervals of [.697, .948] and [.252, .685] for boys and girls, respectively. In this example where the two ratings are dichotomous, Cohen's kappa and intraclass kappa are questionable measures of interrater agreement but they are valid measures of interrater reliability. To obtain point and interval estimates of Cohen's kappa and intraclass kappa, the frequency counts in the two 2x2 contingency tables for boys and girls are required. Suppose the frequency counts for boys are f_{11} = 65, f_{12} = 4, f_{21} = 1, and f_{22} = 5. The point estimates are .631 and .630, and the 95% Wald confidence intervals are [.336, .926] and [.331, .928] for Cohen's kappa and intraclass kappa, respectively. Suppose the frequency counts for girls are f_{11} = 35, f_{12} = 8, f_{21} = 7, and f_{22} = 10. The point estimates are .395 and .395, and the 95% Wald confidence intervals are [.142, .649] and [.141, .649] and for Cohen's kappa and intraclass kappa, respectively.

9.3 Meta-analysis of Two-rater Studies

Suppose two published studies used two raters to assess interrater agreement for the absence or presence (r = 2) of gender stereotype behavior in educational children's TV shows. The first published study used a random sample of n_1 = 50 episodes and reported agreement in 41 of the 50 episodes. The second published study used a random sample of n_2 = 70 episodes and reported agreement in 58 of the 70 episodes. The three commands `f = c(41, 58), n = c(50, 70),` and `ci.meta(.05, f, n, 2)` computes Equation 9 and returns a point estimate for $(\kappa_{G1} + \kappa_{G2})/2$ of .648 with a 95% confidence interval of [.488, .766].

9.4 Linear Contrast in a Three-study Design

Suppose the above meta-analysis included a third published study that sampled episodes of non-educational children's TV shows. The third study used a random sample of n_3 = 90 episodes and reported agreement in 85 of the 90 episodes. The four commands `f = c(41, 58, 85), n = c(50, 70, 90), h = c(-.5, -.5, 1),` and `ci.contrast(.05, f, n, h, 2)`

computes Equation 10 and returns a point estimate for $\kappa_{G3} - (\kappa_{G1} + \kappa_{G2})/2$ of .240 with a 95% confidence interval of [.071, .412]. This result suggests that interrater agreement is greater for non-educational than educational TV shows.

9.5 Four-rater Design

Four parole officers (q = 4) evaluated the parole application files of 100 prisoners and there was a unanimous grant or deny (r = 2) agreement in 87 of the 100 files. The command `ci.qrater(.05, 100, 87, 4, 2)` computes Equations 12-14 and returns a point estimate for $\kappa_{G(4)}$ of .851 with a 95% confidence interval of [.758, .912].

9.6 Three-rater Design

A school psychologist (Rater 1), a teacher (Rater 2), and a principal (Rater 3) evaluated a sample of 300 high school students with disciplinary problems and gave a suspension or a non-suspension recommendation (r = 2) for each student. The `ci.3rater` function requires a vector of the eight frequency counts in the 2^3 contingency table (see comments in the `ci.3rater` function about how to order the frequencies). In this example suppose the frequency counts are 100, 6, 4, 40, 20, 1, 9, and 120. The two commands `f = c(100, 6, 4, 40, 20, 1, 9, 120)` and `ci.3rater(.05, f)` computes Equations 21 and 22 and returns 95% confidence intervals for $\kappa_{G\{1,2\}} - \kappa_{G\{1,3\}}$, $\kappa_{G\{1,2\}} - \kappa_{G\{2,3\}}$, and $\kappa_{G\{1,3\}} - \kappa_{G\{2,3\}}$ of [.006, .127], [-.407, -.189], and [-.462, -.266], respectively. These results indicate that the agreement between the school psychologist and teacher is greater than the agreement between the school psychologist and principal. The results also indicate that the agreement between the teacher and the principal is greater than the agreement between the school psychologist and teacher as well as the agreement between the school psychologist and principal.

9.7 Comparison of Two-rater Agreement in Four-rater Design

Two graduate students (Raters 1 and 2) and two undergraduate students (Raters 3 and 4) were trained to code open-ended responses for the absence or presence (r = 2) of a particular ideological theme in newspaper articles. Suppose these four raters coded 300 articles and frequency counts of f_1 = 78 and f_2 = 52 are extracted from the 2^4 table (see Equation for 27 for definition of f_1 and f_2). The command `ci.4rater(.05, 300, 78, 52)` computes Equations 26 and 27 and returns a point estimate $\kappa_{G\{1,2\}} - \kappa_{G\{3,4\}}$ of .173 with a 95% confidence interval of [.024, .320].

9.8 Sample Size Requirements for Two-rater and Three-rater Designs

A proposed study will use two master teachers (q = 2) to provide dichotomous ratings (meets expectations or needs improvement) for a sample of student teachers based on classroom observations. Setting $\tilde{\kappa}_G$ = .90, r = 2, α = .05, and a desired confidence interval width of .25, `size.qrater(.05, .9, 2, .25, 2)` computes Equations 28 and 29 and returns a sample size requirement of 71 student teachers to be rated by two master teachers. The researcher is also considering using three master teachers to rate each student. Setting $\tilde{\kappa}_G$ = .90, r = 2, α = .05, and a desired confidence interval width of .25, `size.qrater(.05, .9, 2, .25, 3)` computes Equation 31 and 29 and returns a sample size requirement of 42 student teachers to be rated by three master teachers.

9.9 Sample Size Requirement for Two-group Design

The interrater reliability for two expert raters will be compared with the interrater reliability of two novice raters. The expert raters will classify one random sample of newspaper articles regarding educational reform into three different categories. The novice raters will perform the same task using another random sample of newspaper articles. Setting $\tilde{\kappa}_{G1}$ = .80, $\tilde{\kappa}_{G2}$ = .7, r = 3,

α = .05, and a desired confidence interval width of .30, `size.diff(.05, .8, .7, 3, .3)` computes Equation 30 and 29 returns a sample size requirement of 107 per group. The two expert raters should evaluate one random sample of 107 newspaper articles and the two novice raters should evaluate a second random sample of 107 newspapers articles.

10. Conclusion

The *G*-index of agreement is an attractive alternative to Cohen's kappa for the assessment of nominal scale agreement. The new confidence interval and sample size methods presented here parallel the methods developed for Cohen's kappa over the last 50 years. The R functions in the online Supplementary Materials can be used to apply the new methods presented here for the *G*-index of agreement so that researchers will now be able to perform the same types of inferential and sample size analyses that are currently available for Cohen's kappa.

All of the proposed confidence intervals are based on adjusted Wald confidence intervals which have been shown to have excellent performance characteristics in terms of expected coverage probability, worst-case coverage probability, and expected confidence interval width (Agresti & Coull, 1998; Agresti & Caffo, 2000; Bonett & Price, 2012; Price & Bonett, 2004). Newcombe (2013) describes alternatives to the adjusted Wald confidence intervals that have different performance characteristics that might be more desirable in certain applications. For example, the Clopper-Pearson confidence interval for a single population proportion tends to be substantially wider that the adjusted Wald interval but it has a worst-case coverage probability that is guaranteed to be no less than 1 – α. If worst-case coverage probability is the primary concern, then the adjusted Wald confidence intervals (the terms in brackets) in Equations 5 and 14 could be replaced with Clopper-Pearson confidence intervals. If any current or newly-developed confidence interval for a single proportion, a difference of independent proportions, a difference of paired

proportions, or a linear function of proportions is considered to be more appropriate than an adjusted Wald confidence, then that confidence interval can be used in place of the adjusted Wald intervals used here.

The results in sections 3, 4, 5, and 6 are general for $r \geq 2$, but the results in sections 6.1 and 6.2 are limited to $r = 2$. Future research could extend the results in sections 6.1 and 6.2 to the general case of $r \geq 2$.

References

Agresti, A., & Coull, B. A. (1998). Approximate is better than "exact" for interval estimation of binomial proportions. *The American Statistician, 52,* 119-126.

Agresti, A., & Caffo, B. (2000). Simple and effective confidence intervals for proportions and differences of proportions result from adding two successes and two failures. *The American Statistician, 54,* 280-288.

Altman, D. G. (1991). *Practical statistics for medical research*. Boca Raton: Chapman and Hall

Banerjee, M., Capozzoli, M., McSweeney, L., & Sinha, D. (1999). Beyond kappa: A review of interrater agreement measures. *Canadian journal of statistics, 27,* 3-23.

Bennett, E. M., Alpert, R., & Goldstein, A. C. (1954). Communications through limited response questioning. *Public Opinion Quarterly, 18,* 303-308.

Bishop, Y. M. M., Fienberg, S. E., & Holland, P. W. (1975). *Discrete multivariate analysis: Theory and Practice*. Cambridge: MIT Press,

Blackman, N. J-M., & Koval, J. J. (1993). Estimating rater agreement in 2x2 tables: Correction for chance and intraclass correlation. *Applied Psychological Measurement, 17,* 211-223.

Blackman, N. J-M., & Koval, J. J. (2000). Interval estimation for Cohen's kappa as a measure of agreement. *Statistics in Medicine, 19,* 723-741.

Bonett, D. G. (2010). Varying coefficient meta-analytic methods for alpha reliability. *Psychological Methods, 15,* 368-385.

Bonett, D. G., & Price, R. M. (2012). Adjusted Wald interval for a difference of binomial proportions based on paired data. *Journal of Educational and Behavioral Statistics, 37,* 479-488.

Bonett, D. G. & Wright, T. A. (2000). Sample size requirements for estimating Pearson, Kendall, and Spearman correlations. *Psychometrika, 65,* 23-28.

Bonett, D. G. & Price, R. M. (2015). Varying coefficient meta-analysis methods for odds ratios and risk ratios. *Psychological Methods, 20,* 394-406.

Borenstein, M., Hedges, L. V., Higgins, J. P. T, & Rothstein, H. R. (2009). *Introduction to meta-analysis*. New York: Wiley.

Brennan, R. L., & Prediger, D. J. (1981). Coefficient kappa: Some uses, misuses, and alternatives. *Educational and Psychological Measurement, 41*, 687-699.

Bujang, M. A., & Baharum, N. (2017). Guidelines of the minimum sample size requirements for Cohen's kappa. *Biostatistics, 14*, e12267-1.

Byrt, T., Bishop, J, & Carlin, J. B. (1993). Bias, prevalence and kappa. *Journal of Clinical Epidemiology, 46,* 423- 429.

Cantor, A. B. (1996). Sample-size calculations for Cohen's kappa. *Psychological Methods, 1,* 150-153.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20,* 37-46.

Donner, A., Eliasaziw, M., & Klar, N. (1996). Testing the homogeneity of kappa statistics. *Biometrics, 52,* 176-183.

Donner, A., & Eliasziw, M. (1992). A goodness-of-fit approach to inference procedures for the kappa statistic: Confidence interval construction, significance-testing and sample size estimation. *Statistics in Medicine, 11,* 1511-1519.

Donner, A., Shoukri, M. M., Klar, N., & Bartfay, E. (2000). Testing the equality of two dependent kappa statistics. *Statistics in Medicine, 19,* 373-387.

Donner, A., & Zou, G. (2002). Interval estimation for a difference between intraclass kappa statistics. *Biometrics, 58*, 209-215.

Feinstein, A. R., & Cicchetti, D. V. (1990). High agreement but low kappa: I. The problem of two paradoxes. *Journal of Clinical Epidemiology, 43*, 543-549.

Flack, V. F. (1987). Confidence intervals for the interrater agreement measure kappa. *Communications in Statistics – Theory and Methods, 16*, 953-968.

Flack, V. F, Afifi, A. A., Lachenbruch, P. A., & Schouten, H. J. A. (1988). Sample size determination for the two rater kappa statistic. *Psychometrika, 53*, 321-325.

Fleiss, J. L. (1975). Measuring agreement between two judges on the presence or absence of a trait. *Biometrics, 31*, 651-659.

Fleiss, J. L., Cohen, J., & Everitt, B. S. (1969). Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin, 72*, 323-327

Fleiss, J. L., Levin, B., & Paik, M. C. (2003). *Statistical methods for rates and proportions, 3rd ed.* New York: Wiley.

Green, S. B. (1981). A comparison of three indexes of agreement between observers: Proportion of agreement, G-index, and kappa. *Educational and Psychological Measurement, 41*, 1069-1072.

Grove, W. M., Andreases, N.C., McDonald-Scott, P. Keller, M. B., & Shapiro, R. W. (1981). Reliability studies of psychiatric diagnosis. *Archives of General Psychiatry, 38*, 408-413.

Gwet, K. L. (2008). Computing inter-rater reliability and its variance in the presence of high agreement. *British journal of Mathematical and Statistical Psychology, 61*, 29-48.

Hayes, A. F., & Krippendorff, K. (2007). Answering the call for a standard reliability measures for coding data. *Communications Methods and Measures, 1*, 77-89.

Holley, W., & Guilford, J. P. (1964). A note on the G-index of agreement. *Educational and Psychological Measurement, 24,* 749-754.

Hubert, L. (1977). Kappa revisited. *Psychological Bulletin, 48,* 289-297.

Klar, N., Lipsitz, S. R., Parzen, M., & Leong, T. (2002). Exact confidence interval for κ in small samples. *Journal of the Royal Statistical Society: Series D, 51,* 467-478.

Kottner, J., & Streiner, D. L. (2011). The difference between reliability and agreement. *Journal of Clinical Epidemiology, 64,* 701-207.

Krippendorff, K. (1970). Estimating the reliability, systematic error, and random error of interval data. *Educational and Psychological measurement, 30,* 61-70.

Krippendorff, K. (2004). Reliability in content analysis: Some common misconceptions and recommendations. *Human Communication Research, 30,* 411-433.

Landis, J. R. & Koch, G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33,* 159-174.

Lawlis, G. F. & Lu, E. (1972). Judgements of counseling process: Reliability, agreement, and error. *Psychological Bulletin, 78,* 17-20.

Lee, J. J., & Tu, Z. N. (1994). A better confidence interval for kappa (κ) on measuring agreement between two raters with binary outcomes. *Journal of Computational and Graphical Statistics, 3,* 301-321.

Maclure, M., & Willett, W. C. (1987). Misinterpretation and misuse of the kappa statistic. *American Journal of Epidemiology, 126,* 161-169.

Maxwell, A. E. (1977). Coefficient of agreement between observers and their interpretations. *British Journal of Psychiatry, 130,* 79-83.

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Lawrence Earlbaum.

McKenzie, D. P., MacKinnon, A. J., Péladeau, N., Onghena, P. C., Clarke, D. M. Harringan, S., & McGorry, P. D. (1996). Comparing correlated kappas by resampling: Is one level of agreement significantly different from another? *Journal of Psychiatry Research, 30*, 483-492.

Newcombe, R. G. (2013). *Confidence intervals for proportions and related measures of effect size*. Boca Raton: CRC Press.

Price, R. M., & Bonett, D. G. (2004). Improved confidence interval for a linear function of binomial proportions. *Computational Statistics & Data Analysis, 45*, 449-456.

Scott, W. A. (1955). Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quarterly, 19*, 321-325.

Shoukri, M. M. (2011). *Measures of interobserver agreement and reliability, 2nd ed*. Boca Raton: CRC Press.

Sun, S. (2011). Meta-analysis of Cohen's kappa. *Health Services and Outcomes Research Methodology, 11***,** 145–163

Vacha-Haase, T. (1998). Reliability generalization: Exploring variance in measurement error affecting score reliability. *Educational and Psychological Measurement, 58*, 6-20.

Warrens, M. J. (2010). A formal proof of a paradox associated with Cohen's kappa. *Journal of Classification, 27*, 322-332.

Wickens, T. D, (2002). *Elementary signal detection theory*. New York, Oxford University Press.

Williamson, J. M., & Manatunga, A. K. (1997). The Consultant's Forum: Assessing interrater agreement from dependent data. *Biometrics, 53*, 707-714.

Supplementary Material: R functions

```
ci.qrater <- function(alpha, n, f, q, r) {
 # Computes adjusted Wald confidence interval for G with
 # q = 2 raters and r-category ratings or for G(q) with
 # q > 2 raters and r-category ratings.
 # Arguments:
 #   alpha:  alpha level for 1-alpha confidence
 #   n:      sample size
 #   f:      number of objects rated in unanimous agreement
 #           (f11...1 + f22...2 + ... + frr...r)
 #   q:      number of raters
 #   r:      number of rating categories
 # Returns:
 #   ML estimate, SE, and adjusted Wald confidence interval for G or G(q)
 z <- qnorm(1 - alpha/2)
 a <- r^(q - 1)
 p.ml <- f/n
 p <- (f + 2)/(n + 4)
 G.ml <- a*p.ml/(a - 1) - 1/(a - 1)
 se.G.ml <- a*sqrt(p.ml*(1 - p.ml)/n)/(a - 1)
 se <- sqrt(p*(1 - p)/(n + 4))
 LL <- a*(p - z*se)/(a - 1) - 1/(a - 1)
 UL <- a*(p + z*se)/(a - 1) - 1/(a - 1)
 out <- t(c(G.ml, se.G.ml, LL, UL))
 colnames(out) <- c("Estimate", "SE", "LL", "UL")
 return(out)
}
```

Example 1 (two raters and three categories)

```
ci.qrater(.05, 90, 82, 2, 3)
      Estimate         SE        LL        UL
[1,] 0.8666667 0.04499657 0.7469308 0.9339203
```

Example 2 (four raters and two categories)

```
ci.qrater(.05, 100, 87, 4, 2)
      Estimate         SE       LL        UL
[1,] 0.8514286 0.03843468 0.757998 0.9123317
```

```
ci.diff <- function(alpha, n1, f1, n2, f2, r) {
 # Computes confidence interval for a difference in G-index values for two
 # raters and r-category ratings estimated from two independent samples.
 # Arguments:
 #   alpha:  alpha level for 1-alpha confidence
 #   n1:     sample size for group 1
 #   f1:     number of objects rated in agreement in group 1
 #   n2:     sample size for group 2
 #   f2:     number of objects rated in agreement in group 2
 #   r:      number of rating categories
 # Returns:
 #   ML estimate and adjusted Wald confidence intervals
 z <- qnorm(1 - alpha/2)
 a <- r/(r - 1)
 p1.ml <- f1/n1;    p1 <- (f1 + 2)/(n1 + 4)
 G1 <- a*p1.ml - 1/(r - 1)
 se1 <- sqrt(p1*(1 - p1)/(n1 + 4))
 LL1 <- a*(p1 - z*se1) - 1/(r - 1);   UL1 <- a*(p1 + z*se1) - 1/(r - 1)
 p2.ml <- f2/n2;    p2 <- (f2 + 2)/(n2 + 4)
 G2 <- a*p2.ml - 1/(r - 1)
 se2 <- sqrt(p2*(1 - p2)/(n2 + 4))
 LL2 <- a*(p2 - z*se2) - 1/(r - 1);   UL2 <- a*(p2 + z*se2) - 1/(r - 1)
 p1.d <- (f1 + 1)/(n1 + 2);  p2.d <- (f2 + 1)/(n2 + 2)
```

```
 se.d <- sqrt(p1.d*(1 - p1.d)/(n1 + 2) + p2.d*(1 - p2.d)/(n2 + 2))
 LL3 <- a*(p1.d - p2.d - z*se.d);     UL3 <- a*(p1.d - p2.d + z*se.d)
 out1 <- t(c(G1, LL1, UL1))
 out2 <- t(c(G2, LL2, UL2))
 out3 <- t(c(G1 - G2, LL3, UL3))
 out <- rbind(out1, out2, out3)
 colnames(out) <- c("Estimate", "LL", "UL")
 rownames(out) <- c("G1", "G2", "G1 - G2")
 return(out)
}
```

Example

```
ci.diff(.05, 75, 70, 60, 45, 2)
         Estimate        LL        UL
G1      0.8666667 0.6974555 0.9481141
G2      0.5000000 0.2523379 0.6851621
G1 - G2 0.3666667 0.1117076 0.6088621
```

```
ci.3rater <- function(alpha, f) {
 # Computes ML estimates and adjusted Wald confidence intervals for G{1,2}, G{1,3}
 # G{2,3}, G{1,2}-G{1,3}, G{1,2}-G{2,3}, G{1,3}-G{2,3}, and G(3) for three
 # dichotomous ratings.
 # Arguments;
 #   alpha;  alpha level for 1-alpha confidence
 #   n;      sample size
 #   f;      vector of frequency counts from 2x2x2 table
 #           f = [f111, f112, f121, f122, f211, f212, f221, f222]
 #           first subscript represents rating of rater 1
 #           second subscript represents rating of rater 2
 #           third subscript represent rating of rater 3
 # Returns;
 #   ML estimates and adjusted Wald confidence intervals for G-indices
 z <- qnorm(1 - alpha/2)
 f111 <- f[1]; f112 <- f[2]; f121 <- f[3]; f122 <- f[4]
 f211 <- f[5]; f212 <- f[6]; f221 <- f[7]; f222 <- f[8]
 n <- sum(f);
 p12.ml <- (f111 + f112 + f221 + f222)/n;     p12 <- (f111 + f112 + f221 + f222 + 2)/(n + 4)
 p13.ml <- (f111 + f121 + f212 + f222)/n;     p13 <- (f111 + f121 + f212 + f222 + 2)/(n + 4)
 p23.ml <- (f111 + f211 + f122 + f222)/n;     p23 <- (f111 + f211 + f122 + f222 + 2)/(n + 4)
 G12.ml <- 2*p12.ml - 1;                      G12 <- 2*p12 - 1
 G13.ml <- 2*p13.ml - 1;                      G13 <- 2*p13 - 1
 G23.ml <- 2*p23.ml - 1;                      G23 <- 2*p23 - 1
 se.G12 <- sqrt(p12*(1 - p12)/(n + 4))
 se.G13 <- sqrt(p13*(1 - p13)/(n + 4))
 se.G23 <- sqrt(p23*(1 - p23)/(n + 4))
 p1.ml <- (f112 + f221)/n;   p1 <- (f112 + f221 + 1)/(n + 2)
 p2.ml <- (f121 + f212)/n;   p2 <- (f121 + f212 + 1)/(n + 2)
 p3.ml <- (f211 + f122)/n;   p3 <- (f211 + f122 + 1)/(n + 2)
 G12_13.ml <- 2*(p1.ml - p2.ml);            G12_13 <- 2*(p1 - p2)
 G12_23.ml <- 2*(p1.ml - p3.ml);            G12_23 <- 2*(p1 - p3)
 G13_23.ml <- 2*(p2.ml - p3.ml);            G13_23 <- 2*(p2 - p3)
 se.G12_13 <- sqrt((p1 + p2 - (p1 - p2)^2)/(n + 2))
 se.G12_23 <- sqrt((p1 + p3 - (p1 - p3)^2)/(n + 2))
 se.G13_23 <- sqrt((p2 + p3 - (p2 - p3)^2)/(n + 2))
 p123.ml <- (f111 + f222)/n;                p123 <- (f111 + f222 + 2)/(n + 4)
 G3.ml <- (4*p123.ml - 1)/3;                G3 <- (4*p123 - 1)/3
 se.G3 <- sqrt(p123*(1 - p123)/(n + 4))
 LL.G12 <- 2*(p12 - z*se.G12) - 1;          UL.G12 <- 2*(p12 + z*se.G12) - 1
 LL.G13 <- 2*(p13 - z*se.G13) - 1;          UL.G13 <- 2*(p13 + z*se.G13) - 1
 LL.G23 <- 2*(p23 - z*se.G23) - 1;          UL.G23 <- 2*(p23 + z*se.G23) - 1
 LL.G12_13 <- 2*(p1 - p2 - z*se.G12_13);    UL.G12_13 <- 2*(p1 - p2 + z*se.G12_13)
 LL.G12_23 <- 2*(p1 - p3 - z*se.G12_23);    UL.G12_23 <- 2*(p1 - p3 + z*se.G12_23)
 LL.G13_23 <- 2*(p2 - p3 - z*se.G13_23);    UL.G13_23 <- 2*(p2 - p3 + z*se.G13_23)
 LL.G3 <- (4/3)*(p123 - z*se.G3) - 1/3;     UL.G3 <- (4/3)*(p123 + z*se.G3) - 1/3
 out1 <- t(c(G12.ml, LL.G12, UL.G12))
 out2 <- t(c(G13.ml, LL.G13, UL.G13))
```

```
 out3 <- t(c(G23.ml, LL.G23, UL.G23))
 out4 <- t(c(G12_13.ml, LL.G12_13, UL.G12_13))
 out5 <- t(c(G12_23.ml, LL.G12_23, UL.G12_23))
 out6 <- t(c(G13_23.ml, LL.G13_23, UL.G13_23))
 out7 <- t(c(G3.ml, LL.G3, UL.G3))
 out <- rbind(out1, out2, out3, out4, out5, out6, out7)
 colnames(out) <- c("Estimate", "LL", "UL")
 rownames(out) <- c("G{1,2}", "G{1,3}", "G{2,3}", "G{1,2}-G{1,3}", "G{1,2}-G{2,3}",
 "G{2,3}-G{1,3}","G(3)")
 return(out)
}
```

Example

```
f = c(100, 6, 4, 40, 20, 1, 9, 120)
ci.3rater(.05, f)
                Estimate          LL         UL
G{1,2}         0.56666667  0.46601839  0.6524027
G{1,3}         0.50000000  0.39564646  0.5911956
G{2,3}         0.86666667  0.79701213  0.9135142
G{1,2}-G{1,3}  0.06666667  0.00580397  0.1266464
G{1,2}-G{2,3} -0.30000000 -0.40683919 -0.1891873
G{2,3}-G{1,3} -0.36666667 -0.46222023 -0.2662566
G(3)           0.64444444  0.57382971  0.7068720
```

```
ci.4rater <- function(alpha, n, f1, f2) {
 # Computes adjusted Wald confidence interval for G{1,2} - G{3,4}
 # with 2-category ratings.
 # Arguments:
 #   alpha:  alpha level for 1-alpha confidence
 #   n:      sample size
 #   f1:     f1211 + f2111 + f1222 + f2122
 #   f2:     f1121 + f2221 + f1112 + f2212
 #           first subscript represents rating of rater 1
 #           second subscript represents rating of rater 2
 #           third subscript represent rating of rater 3
 #           fourth subscript represent rating of rater 4
 # Returns:
 #   ML estimate, SE, and adjusted Wald confidence interval for
 #   difference in G-indices(G{1,2} - G{3,4})
 z <- qnorm(1 - alpha/2)
 p1.ml <- f1/n;                   p1 <- (f1 + 1)/(n + 2)
 p2.ml <- f2/n;                   p2 <- (f2 + 1)/(n + 2)
 diff.ml <- 2*(p1.ml - p2.ml);    diff <- p1 - p2
 se.ml <- sqrt((p1.ml + p2.ml - (p1.ml - p2.ml)^2)/n)
 se.diff <- sqrt((p1 + p2 - (p1 - p2)^2)/(n + 2))
 LL <- 2*(diff - z*se.diff)
 UL <- 2*(diff + z*se.diff)
 out <- t(c(diff.ml, se.ml, LL, UL))
 colnames(out) <- c("Estimate", "SE", "LL", "UL")
 return(out)
}
```

Example

```
ci.4rater(.05, 300, 78, 52)
      Estimate         SE         LL        UL
[1,] 0.1733333 0.03767502 0.02432764 0.3200432
```

```
ci.meta <- function(alpha, f, n, r) {
 # Computes confidence interval for average G-index estimated
 # from m studies.
 # Arguments:
 #   alpha:  alpha level for 1-alpha confidence
 #   f:      m x 1 vector of agreement frequencies
 #   n:      m x 1 vector of sample sizes
 #   r:      number of rating categories
 # Returns:
 #   estimated averages, standard errors, confidence intervals
 m <- length(f)
 z <- qnorm(1 - alpha/2)
 nt <- sum(n)
 p.ml <- f/n;                      p <- (f + 2/m)/(n + 4/m)
 G.ml <- (r*p.ml - 1)/(r - 1);     G <- (r*p - 1)/(r - 1)
 ave.G.ml <- sum(G.ml)/m;          ave.G <- sum(G)/m
 var.G <- (r/(r - 1))^2*p*(1 - p)/(n + 4/m)
 se.ave <- sqrt(sum(var.G)/m^2)
 LL <- ave.G - z*se.ave
 UL <- ave.G + z*se.ave
 out <- cbind(ave.G.ml, LL, UL)
 cat(paste("Total sample size =", nt), fill = TRUE)
 cat(paste("Confidence level =", (1 - alpha)), fill = TRUE)
 colnames(out) <- c("Estimate", "LL", "UL")
 return (out)
}
```

Example

```
f = c(41, 58)
n = c(50, 70)
ci.meta(.05, f, n, 2)
Total sample size = 120
Confidence level = 0.95
      Estimate       LL        UL
[1,] 0.6485714 0.487966 0.7663075
```

```
ci.contrast <- function(alpha, f, n, h, r) {
 # Computes confidence interval for a linear contrast of G-indices
 # estimated from m studies.
 # Arguments:
 #   alpha:  alpha level for 1-alpha confidence
 #   f:      m x 1 vector of agreement frequencies
 #   n:      m x 1 vector of sample sizes
 #   h:      m x 1 vector of contrast coefficients
 #   r:      number of rating categories
 # Returns:
 #   estimated averages, standard errors, confidence intervals
 m <- length(f) - length(which(h==0))
 z <- qnorm(1 - alpha/2)
 nt <- sum(n)
 p.ml <- f/n
 p <- (f + 2/m)/(n + 4/m)
 G.ml <- (r*p.ml - 1)/(r - 1)
 G <- (r*p - 1)/(r - 1)
 con.G.ml <- t(h)%*%G.ml
 con.G <- t(h)%*%G
 var.G <- (r/(r - 1))^2*p*(1 - p)/(n + 4/m)
 se.con <- sqrt(t(h)%*%(diag(var.G))%*%h)
 LL <- con.G - z*se.con
 UL <- con.G + z*se.con
 out <- cbind(con.G.ml, LL, UL)
 colnames(out) <- c("Estimate", "LL", "UL")
 return (out)
}
```

Example

```
f = c(41, 58, 85)
n = c(50, 70, 90)
h = c(-.5, -.5, 1)
ci.contrast(.05, f, n, h, 2)
      Estimate         LL        UL
[1,] 0.2403175 0.07122621 0.4123622
```

```
size.ci.diff <- function(alpha, G1, G2, r, w) {
 # Computes the sample size per group required to estimate a difference
 # in G-indices (two-rater) in 2-sample design with desired precision.
 # Arguments:
 #   alpha:  alpha level for 1-alpha confidence
 #   G1:     planning value of G-index in group 1
 #   G2:     planning value of G-index in group 2
 #   r:      number of rating categories
 #   w:      desired confidence interval width
 # Returns:
 #   required per group sample size
 z <- qnorm(1 - alpha/2)
 v <- (G1 + 1/(r - 1))*(1 - G1) + (G2 + 1/(r - 1))*(1 - G2)
 n0 <- ceiling(4*v*(z/w)^2)
 p01 <- ((r - 1)/r)*(G1 + 1/(r - 1))
 p02 <- ((r - 1)/r)*(G2 + 1/(r - 1))
 p1 <- (p01*n0 + 1)/(n0 + 2)
 p2 <- (p02*n0 + 1)/(n0 + 2)
 se <- sqrt(p1*(1 - p1)/(n0 + 2) + p2*(1 - p2)/(n0 + 2))
 LL <- (r/(r - 1))*(p1 - p2 - z*se)
 UL <- (r/(r - 1))*(p1 - p2 + z*se)
 w0 <- UL - LL
 n <- ceiling(n0*(w0/w)^2)
 return(n)
}
```

Example

```
size.ci.diff(.05, .8, .7, 4, .3)
[1] 93
```

```
size.ci.qrater <- function(alpha, G, r, q, w) {
 # Computes the sample size required to estimate a G-index in a
 # 1-sample design (two or more raters) with desired precision.
 # Arguments:
 #   alpha:  alpha level for 1-alpha confidence
 #   G:      planning value of G-agreement
 #   r:      number of rating categories
 #   w:      desired confidence interval width
 #   q:      number of raters
 # Returns:
 #   required sample size
 z <- qnorm(1 - alpha/2)
 n0 <- ceiling(4*(G + 1/(r^(q - 1) - 1))*(1 - G)*(z/w)^2)
 a <- r^(q - 1)
 p0 <- ((a - 1)/a)*(G + 1/(a - 1))
 p <- (n0*p0 + 2)/(n0 + 4)
 se <- sqrt(p*(1 - p)/(n0 + 4))
 LL <- a*(p - z*se)/(a - 1) - 1/(a - 1)
 UL <- a*(p + z*se)/(a - 1) - 1/(a - 1)
 w0 <- UL - LL
 n <- ceiling(n0*(w0/w)^2)
 return(n)
}
```

Example 1 (three rating categories and two raters)

```
size.ci.qrater(.05, .8, 3, 2, .25)
[1] 69
```

Example 2 (two rating categories and three raters)

```
size.ci.qrater(.05, .8, 2, 3, .25)
[1] 59
```