k-Rater Reliability: The Correct Unit of Reliability for Aggregated Human Annotations

Ka Wong
Google Research
danicky@gmail.com

Praveen Paritosh
Google Research
pkp@google.com

Abstract

Since the inception of crowdsourcing, aggregation has been a common strategy for dealing with unreliable data. Aggregate ratings are more reliable than individual ones. However, many natural language processing (NLP) applications that rely on aggregate ratings only report the reliability of individual ratings, which is the incorrect unit of analysis. In these instances, the data reliability is under-reported, and a proposed *k-rater reliability* (kRR) should be used as the correct data reliability for aggregated datasets. It is a multi-rater generalization of *inter-rater reliability* (IRR). We conducted two replications of the WordSim-353 benchmark, and present empirical, analytical, and bootstrap-based methods for computing kRR on WordSim-353. These methods produce very similar results. We hope this discussion will nudge researchers to report kRR in addition to IRR.

1 Introduction

Crowdsourcing has become a mainstay for data collection in NLP (Geva et al., 2019; Sabou et al., 2014). It can produce data in a scalable and cost effective manner. However, these benefits come at a cost: quality. The reliability of crowd workers is always of central concern. One common strategy to increase the data reliability is to collect multiple, independent judgements and to use the aggregated judgements instead. Indeed, early papers such as Snow et al. (2008) show that average ratings correlate more strongly with expert judgements. This makes sense, as average ratings are known to have a higher reliability than individual ones (Ebel, 1951).

A number of strategies have been proposed to address data quality issues, e.g. rater modeling, label correction, label pruning (Kumar and Lease, 2011), but aggregation remains very popular (Prabhakaran et al., 2021). Sheshadri and Lease (2013) present nine crowdsourced datasets across a wide range of NLP tasks to compare different aggregation methods. See Difallah and Checco (2021) for a recent review of aggregation techniques. In short, aggregation has become the default method for acquiring reliable data from the crowd.

Interestingly, after we adopted aggregation as a community, we forgot to update our reliability measures correspondingly. The field continues to report data reliability in terms of IRR, even when aggregate ratings are used. Focusing on IRR, we are unable to capture the increase in reliability due to aggregation. The actual data reliability is hence unknown. This has important consequences. Reliability is often used as a safeguard for *reproducibility*. Therefore conclusions about the reproducibility of a dataset drawn based the reliability of individual ratings may be different than that based on the reliability of aggregate ratings.

By reporting the correct reliability that is actually higher, this may even have a side effect of lessening the stigma on low-IRR datasets. As a result, this may create a path forward towards reliable data on subjective tasks, where a high IRR is difficult to obtain, such as emotions (Wong et al., 2021) and toxicity (Wulczyn et al., 2017). With a reproducibility crisis looming in the background (Baker, 2016; Hutson, 2018), more frequent and accurate reporting of reliability is our primary safeguard (Paritosh, 2012).

We denote the reliability of aggregate ratings as *k-rater reliability* (kRR), in order to differentiate it from inter-rater reliability. In this paper we present a few methods for computing kRR. First, we demonstrate a general, empirical approach that is based on replications. To that end, we conducted two replications of WordSim-353 (Finkelstein et al., 2001), a widely used word similarity dataset. We then discuss two other alternatives that do not require replications. One is a re-sampling-based bootstrap approach (Efron and Tibshirani, 1994). It is suitable for experiments with a high rating redun-

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Volume 2: Short Papers, pages 378 - 384

dancy. The other is an existing analytical approach based on intraclass correlation (ICC). It is suitable for continuous data where the aggregation is the mean. We conclude with recommendations for reporting reliability of crowdsourced annotations, and novel research questions to expand the usefulness of kRR.

2 Related Work

Various authors have stressed the importance of measuring reliability for the correct unit of analysis. Ebel (1951) asks "Is it better to estimate the reliability of individual ratings or the reliability of average ratings? If decisions are based upon average ratings, it of course follows that the reliability with which one should be concerned is the reliability of those averages." Shrout and Fleiss (1979) and Hallgren (2012) reiterate similar points.

These studies primarily focus on the reliability of the *mean*, which is just one of many different aggregation methods. There is a reason. Not only is the mean a popular choice, it is also the only known choice where the reliability of the aggregate ratings can be computed *analytically* from the reliability of individual ratings. This is done in the ICC framework. ICC is typically used to measure the reliability of single ratings, but it actually has a variant that can be used for mean ratings as well. Shrout and Fleiss (1979) list several types of ICC coefficient, one of which is for mean ratings. They call it ICC(k), where k is the number of ratings per item. In this generalized notation, ICC(1) is just the reliability of individual ratings, or the IRR. Note that McGraw and Wong (1996) use a slightly different notation, ICC($1, k$), to explicitly denote that it is for a one-way random effects model, where the raters are treated as interchangeable. That is a common assumption in most crowdsourcing experiments done on commercial platforms such as Amazon Mechanical Turk.

ICC(k) is an established way of measuring the reliability of mean ratings, hence it is readily usable by researchers. However, it has some drawbacks. Being part of the ICC family, ICC(k) is only applicable to continuous data. In addition, ICC(k) measures the reliability of *mean* ratings, therefore it cannot accommodate other aggregation functions. In other words, for other popular data types, such as majority votes of binary data, there is no known coefficients for measuring the reliability of aggregate ratings. Other than ICC(k), the authors are not aware of any multi-rater generalization for other coefficients such as Cohen's (1960) *kappa* or Krippendorff's *alpha* (Krippendorff, 2011). We therefore take ICC(k) as an inspiration and abstract away from it to define a class of reliability that describes the reliability of aggregate ratings for any data types. We denote it kRR.

3 Contributions

- We emphasise the reliability of aggregate ratings is higher than that of individual ratings.
- We give a general definition of kRR, extending from the definition of IRR, and discuss three methods for computing it.
- We conduct two replications of the WordSim-353 benchmark to validate these methods.

4 *k*-Rater Reliability

We define kRR as the chance-adjusted agreement between replications of aggregate ratings. This definition is very similar to IRR. In fact, they only differ in terms of interpretation. kRR is identical to IRR other than that each individual rating in the IRR calculation is replaced by a k-rater aggregate rating. After all, the mathematics in IRR are agnostics to how those labels are produced.

Just like IRR, a minimum of two replications is required to calculate kRR. Given two vectors of aggregate ratings, one can calculate the reliability between them using any IRR coefficients that fit the purpose. kRR is designed to be analogous to IRR so that we can build upon the rich IRR literature and the various coefficient choices for different experimental conditions and assumptions. For example, in a binary task, if all the items are rated by two fixed but distinct groups of raters (raters from different locales), Cohen's (1960) *kappa* is a suitable choice. Whereas if the raters groups are homogeneous, and the rating scale is ordinal (e.g. Likert), then Krippendorff's *alpha* (Krippendorff, 2011) can be used. Just like IRR, kRR is a general concept and is agnostic to the choice of coefficient.

This definition of kRR can be directly operationalized by creating replications. We call this approach to calculating kRR the *empirical approach*. We demonstrate it in the next section on the WordSim-353 benchmark. The empirical approach is the most direct and most general, with the drawback that a minimum of two replications

are required. We later present two narrower alternatives in Section 5 that do not require replications. The empirical results will be used as a golden reference to validate them.

4.1 Replicating the WordSim Dataset

WordSim-353 (Finkelstein et al., 2001) is a widely used benchmark for measuring a system's ability to compute similarity between two words, and has been cited over 1500 times. The dataset contains 353 word pairs. Each word pair is rated by the same 13 workers for their similarity on a scale from 1 to 10, to indicate how similar their meanings are. The 13 ratings on each word pair are then aggregated into a mean score. It is important to note that only the mean of the ratings are utilized by all the research using this dataset as a benchmark. So the unit of analysis is the aggregate of the 13 ratings, not individual ratings.

Nearly twenty years have elapsed since the creation of the WordSim dataset. It is impossible to recreate the original experimental conditions due to rater population changes. Therefore, we created two replications in order to approximate the kRR of the original dataset. Two is the minimum replication factor required for the empirical approach, though a higher replication would result in a more accurate measure of kRR.

We used the original annotation guidelines on Amazon Mechanical Turk. Raters were paid on average USD 9.5 per hour. In each replication, we collected 13 judgements on each of the same 353 word pairs. There was a detail that we did not follow. In the original experiment, the authors employed 13 unique raters, and each one rated all 353 word pairs. In our replications, we followed more modern conventions and limited the contributions of each individual rater for better generalizability. This detail aside, these are our best attempts to replicate the original experiment. The data is publicly available at `https://github.com/google-research-datasets/wordsim-replications`.

4.2 Empirical kRR Results

We take k columns of ratings at random from each of the two replications, compute the k-rater mean scores for each replication, and measure the reliability between them using Krippendorf's *alpha*, the most widely used and general reliability index. We do this for $k = 1, 2, \ldots, 13$. The resulting kRR values are shown in Fig.1. At $k = 1$, the IRR is 0.574,

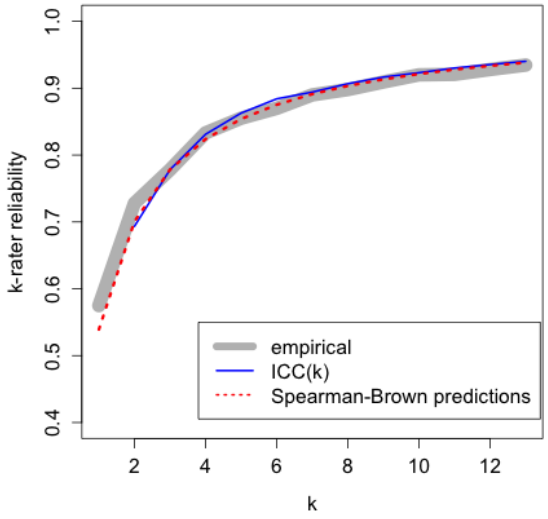

Figure 1: k-rater reliability for replications of WordSim benchmark, calculated using 3 different methods: 1) Empirical, based on replications, 2) ICC(k), analytical, and 3) SB predictions. Note ICC(1) is not available as we only have a single column of ratings available at $k = 1$. All SB predictions are based on only 2 ratings per item.

slightly lower than the 0.6 originally reported in Finkelstein et al. (2001). At $k = 13$, the k-rater reliability is 0.940, quite a bit higher than the IRR. In addition, Fig.1 shows the marginal returns on increasing the number of ratings on the replicated datasets.

5 Other Approaches to Computing kRR

The empirical approach is general, as it can accommodate any choice of rating scale, aggregation function, and reliability coefficient. However, it has a major drawback. As we see in Section 4.1, it can be difficult to do a perfect replication post-fact. This backward incompatibility will present a challenge to computing kRR for existing datasets. Below we present two alternatives that can work on existing datasets under some conditions without requiring any additional data collection. One is a re-sampling based bootstrap approach (Efron and Tibshirani, 1994), the other is ICC(k).

5.1 Bootstrap

Bootstrap (Efron and Tibshirani, 1994) is a re-sampling technique commonly used for quantifying uncertainty in statistical parameter estimation. One can bootstrap an NLP annotations dataset by re-sampling ratings within each annotation item with replacement at the same sample size. If one treats each bootstrap sample as a replication, then one can apply the technique discussed in Section 4

to obtain a *bootstrapped* kRR. Bootstrap is an approximate technique and works better with larger sample sizes, typically 20 observations and above for a single distribution. The 13-rating redundancy in the WordSim replications is arguably small for a typical bootstrap exercise, but it makes up for it with a large number of items.

Before we apply bootstrap to the original WordSim dataset, we first verify its soundness by comparing it against the empirical results in Section 4.2. When applied to one of the two recent replications, the bootstrapped kRR is 0.943. This is comparable to the 0.940 reported in Section 4.2. We then apply bootstrap to the original WordSim dataset and find a bootstrapped kRR of 0.953 (Table 1). The exact method introduced below produces a very similar value at 0.950.

5.2 Intraclass Correlation

Intraclass correlation is a popular reliability coefficient for continuous data in behavioral and medical sciences. ICC gives researchers granular control over assumptions about the raters. For example, each annotation item can be rated by the same set of raters, or different sets of raters (interchangeability). In the former, the raters can be treated as either fixed or randomly drawn from a population. Shrout and Fleiss (1979) and McGraw and Wong (1996) give very extensive treatment on different ICC types for different rater assumptions.

In this paper, we focus on the most basic definition, one that treats raters as interchangeable. The ICC for k-rater averages is denoted as ICC(k) using McGraw and Wong's notation. The reliability of individual ratings is thus given by ICC(1). ICC(k) can be computed by summing squares of differences on the data matrix. Please see Appendix A for derivation and an illustration. Otherwise, software implementations of ICC are also widely available, e.g. in R and Python.

We first verify ICC(k)'s accuracy by comparing it against the empirical results in Section 4.2. To do that, we calculate ICC(k) for one of the two recent WordSim replications for $k = 1, 2, \ldots, 13$ and overlay the results (solid blue) over the empirical curve in Fig.1. We can see ICC(k) matches the empirical results quite well.

After verifying the technique, we compute ICC(k) on the original WordSim dataset. We report in Table 1 both ICC(1) and ICC(13) to show the increase in reliability. They are respectively 0.590 and 0.950.[1]

Unit of analysis	Method	reliability
single-rating	ICC(1)	0.590
13-rating mean	ICC(13)	0.950
13-rating mean	bootstrap	0.953

Table 1: Reliability of the original WordSim benchmark. First two rows are analytical estimates ICC(1) and ICC(13). Both computed using all 13 available ratings. Third row is a resampling-based bootstrapped estimate based on 100 bootstrap samples.

5.3 Spearman-Brown Formula

Given an experiment with a k-rating redundancy, ICC(k) quantifies the reliability of the k-rater average. If this reliability is too low, the researcher may want to increase the value of k. In this case, it would be helpful to know how additional ratings would impact reliability. This is analogous to calculating the required sample size for a given margin of error in a poll. For this purpose, the Spearman-Brown prophecy formula (Spearman, 1910; Brown, 1910) can be a useful tool. It predicts ICC(k) for any value of k based on ICC(1) in the current experiment:

$$\text{ICC}(k) = \frac{k \cdot \text{ICC}(1)}{1 + (k-1) \cdot \text{ICC}(1)}. \qquad (1)$$

Warrens (2017) and de Vet et al. (2017) recently proved that SB and ICC(k) are indeed equivalent in expectation,[2] even though they look nothing alike and were derived in very different contexts. These findings confirm past observations that SB predicts empirical results accurately (Remmers et al., 1927). A limitation of SB is clearly that it only works with ICC. However, Fleiss and Cohen (1973) show ICC is actually equivalent to weighted-kappa with quadratic weights, so it likely has wider applicability.

To verify the formula, we apply SB to one of the two recent WordSim replications and overlay the results (dotted red) over the empirical curve obtained earlier. When computing SB, we only provide it with 2 ratings, in order to assess its predictive accuracy. That is, we first compute ICC(1) with 2 randomly drawn ratings from each word

[1] The former is computed using two-way random without interaction ICC(1), the latter two-way random without interaction ICC(13). The equivalent one-way models yield identical point estimates.

[2] The only exception is two-way mixed model with interaction (Warrens, 2017).

pair, then we plug this ICC(1) value into Eq.1 for $k = 1, 2, \ldots, 13$. The SB curve is overlaid over the empirical curve in Fig.1. We see that SB tracks the empirical results very well even at high k. This is remarkable as the empirical approach requires 26 ratings for $k = 13$, whereas SB merely requires 2 for any value of k.

6 Conclusions and Discussion

We pointed out where aggregated ratings are used, as is the case in many crowdsourced datasets, reliability of aggregate ratings is the correct accounting of data reliability. We introduced k-rater reliability (kRR) as a multi-rater extension of IRR. We emphasise the reliability of aggregate ratings is higher than that of individual ratings. We present analytical and bootstrap-based methods for computing the kRR on the original WordSim dataset. Both methods produce similar estimates for 13-rater reliability ranging from 0.940 to 0.953. We conduct two replications of the entire WordSim-353 benchmark to validate these methods. We make our replication data publicly available on GitHub.

While aggregation makes it possible to have reliable benchmarks on subjective topics, some readers may feel uneasy about increasing reliability via gathering additional ratings, as opposed to other traditional means such as improving rater guidelines. We suggest to mediate this concern by reporting both IRR and kRR. In fact, kRR is not meant to replace IRR, but rather complement it. IRR speaks to the reliability of the labeling process, whereas kRR quantifies the reliability of the aggregated data we consume. We urge researchers to report both where possible. In fact, Hallgren (2012) states, "In cases where single measures ICCs are low but average-measures ICCs are high, the researcher may report both ICCs to demonstrate this discrepancy."

This research also raises interesting questions for future research:

1. How do we derive multi-rater generalizations for coefficients other than ICC? A lot of NLP annotations are binary and multi-class. Such a generalization for majority voting would be particularly useful to the field.

2. Should we apply the Landis and Koch (1977) style of reliability cutoffs to kRR, or should kRR go by a different set of standards?

We urge researchers to report both IRR and kRR of aggregated human annotations, and for further inquiry around the above fundamental questions about reliability.

Acknowledgement

We thank Lora Aroyo and Chris Welty for sharing their WordSim replication datasets. We thank Michael Quinn and Jeremy Miles for their insightful discussions and comments. We also thank all the crowd workers for providing us with valuable annotations data.

References

Monya Baker. 2016. Reproducibility crisis. *Nature*, 533(26):353–66.

William Brown. 1910. Some experimental results in the correlation of mental abilities 1. *British Journal of Psychology, 1904-1920*, 3(3):296–322.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Henrica C.W. de Vet, Lidwine B. Mokkink, David G. Mosmuller, and Caroline B. Terwee. 2017. Spearman-brown prophecy formula and cronbach's alpha: different faces of reliability and opportunities for new applications. *Journal of Clinical Epidemiology*, 85:45–49.

Djellel Difallah and Alessandro Checco. 2021. Aggregation techniques in crowdsourcing: Multiple choice questions and beyond. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 4842–4844.

Robert L Ebel. 1951. Estimation of the reliability of ratings. *Psychometrika*, 16(4):407–424.

Bradley Efron and Robert J Tibshirani. 1994. *An introduction to the bootstrap*. CRC press.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2001. Placing search in context: The concept revisited. In *Proceedings of the 10th international conference on World Wide Web*, pages 406–414.

Joseph L Fleiss and Jacob Cohen. 1973. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and psychological measurement*, 33(3):613–619.

Mor Geva, Yoav Goldberg, and Jonathan Berant. 2019. Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets. *arXiv preprint arXiv:1908.07898*.

Kevin A Hallgren. 2012. Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorials in quantitative methods for psychology*, 8(1):23–34.

Matthew Hutson. 2018. Artificial intelligence faces reproducibility crisis.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

Abhimanu Kumar and Matthew Lease. 2011. Modeling annotator accuracies for supervised learning. In *Proceedings of the Workshop on Crowdsourcing for Search and Data Mining (CSDM) at the Fourth ACM International Conference on Web Search and Data Mining (WSDM)*, pages 19–22.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–74.

David Liljequist, Britt Elfving, and Kirsti Skavberg Roaldsen. 2019. Intraclass correlation–a discussion and demonstration of basic features. *PloS one*, 14(7):e0219854.

Kenneth O McGraw and Seok P Wong. 1996. Forming inferences about some intraclass correlation coefficients. *Psychological methods*, 1(1):30.

Praveen Paritosh. 2012. Human computation must be reproducible. In *WWW 2012, Lyon.*

Vinodkumar Prabhakaran, Aida Mostafazadeh Davani, and Mark Diaz. 2021. On releasing annotator-level labels and information in datasets. In *Proceedings of The Joint 15th Linguistic Annotation Workshop (LAW) and 3rd Designing Meaning Representations (DMR) Workshop*, pages 133–138, Punta Cana, Dominican Republic. Association for Computational Linguistics.

HH Remmers, NW Shock, and EL Kelly. 1927. An empirical study of the validity of the spearman-brown formula as applied to the purdue rating scale. *Journal of Educational Psychology*, 18(3):187.

Marta Sabou, Kalina Bontcheva, Leon Derczynski, and Arno Scharl. 2014. Corpus annotation through crowdsourcing: Towards best practice guidelines. In *LREC*, pages 859–866. Citeseer.

Aashish Sheshadri and Matthew Lease. 2013. Square: A benchmark for research on computing crowd consensus. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 1.

Patrick E Shrout and Joseph L Fleiss. 1979. Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin*, 86(2):420.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Ng. 2008. Cheap and fast – but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 254–263, Honolulu, Hawaii. Association for Computational Linguistics.

Charles Spearman. 1910. Correlation calculated from faulty data. *British Journal of Psychology, 1904-1920*, 3(3):271–295.

Matthijs J Warrens. 2017. Transforming intraclass correlation coefficients with the spearman–brown formula. *Journal of clinical epidemiology*, 85:14–16.

Ka Wong, Praveen Paritosh, and Lora Aroyo. 2021. Cross-replication reliability - an empirical approach to interpreting inter-rater reliability. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7053–7065, Online. Association for Computational Linguistics.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th international conference on world wide web*, pages 1391–1399.

	Rating			
Item	1	2	$\dots j$	$\dots k$
1	x_{11}	x_{12}	$\dots x_{1j}$	$\dots x_{1k}$
2	x_{21}	x_{22}	$\dots x_{2j}$	$\dots x_{2k}$
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
i	x_{i1}	x_{i2}	$\dots x_{ij}$	$\dots x_{ik}$
.	.	.	.	.
.	.	.	.	.
n	x_{n1}	x_{n2}	$\dots x_{nj}$	$\dots x_{nk}$

Figure 2: A convenient data matrix and notational system for data used in calculating intra-class correlation coefficients

A Appendix on ICC(k)

ICC is a family of coefficients. It has slightly different formulations to accommodate different experimental designs. One of them, ICC(k), quantifies the reliability of average ratings based on k raters, where the raters are treated as interchangeable. We illustrate its close form calculation here. It is mainly re-expressing results from previous works on ICC calculation, such as Liljequist et al. (2019) and McGraw and Wong (1996).

ICC(k) predicates on the one-way random effects model being the data generation process. The model takes the form

$$x_{ij} = \mu + \phi_i + \epsilon_{ij},$$

where x_{ij} is the rating on item i from rater j, μ is the grand mean, ϕ_i is the mean of item i, and ϵ_{ij} is a random perturbation term. Assume a data matrix with n rows (item) and k columns (raters) with no missing data, as one shown in Fig. 2. Let

$$\bar{x}_{..} = \frac{1}{nk} \sum_{j=1}^{k} \sum_{i=1}^{n} x_{ij}$$

be the sample grand mean, and

$$\bar{x}_{i.} = \frac{1}{k} \sum_{j=1}^{k} x_{ij}$$

be the i^{th} sample item mean. Let

$$SSW = \sum_{j=1}^{k} \sum_{i=1}^{n} (x_{ij} - \bar{x}_{i.})^2$$
$$SSB = k \sum_{i=1}^{n} (\bar{x}_{i.} - \bar{x}_{..})^2$$

be respectively the sum of squares due to differences *within* items and the sum of squares due to differences *between* items. Then the estimator for the variance of ϵ, σ_ϵ^2, and the estimator for the variance of ϕ, σ_ϕ^2, are respectively

$$\hat{\sigma}_\epsilon^2 = \frac{SSW}{n(k-1)}$$
$$\hat{\sigma}_\phi^2 = \frac{SSB}{k(n-1)} - \frac{\hat{\sigma}_\epsilon^2}{k}.$$

Then ICC(k) can be computed as

$$\frac{\hat{\sigma}_\phi^2}{\hat{\sigma}_\alpha^2 + \hat{\sigma}_\epsilon^2/k}.$$

If we apply the above formula to individual ratings, with $k = 1$, the resulting reliability is known as inter-rater reliability. For any $k > 1$, it is an instance of the *k-rater reliability* proposed in this paper.