Generalized Cohen's Kappa: A Novel Inter-rater Reliability Metric for Non-Mutually Exclusive Categories

Andrea Figueroa[0000−0002−4524−4661], Sourojit Ghosh[0000−0001−5143−6187], and Cecilia Aragon[0000−0002−9502−0965]

University of Washington, Seattle

Abstract.

Qualitative coding of large datasets has been a valuable tool for qualitative researchers. In terms of inter-rater reliability, existing metrics have not evolved to fit current approaches, presenting a variety of restrictions. In this paper, we propose Generalized Cohen's kappa, a novel IRR metric that can be applied in a variety of qualitative coding situations, such as variable number of coders, texts, and non-mutually exclusive categories. We show that under the preconditions for Cohen's kappa, GCK performs very similarly, thus demonstrating their interchangeability. We then extend GCK to the aforementioned situations and demonstrate it to be stable under different permutations.

Keywords: Inter-rater reliability · Cohen's kappa · Generalized Cohen's kappa · Qualitative Coding.

1 Introduction

A key component of qualitative research is *qualitative coding* of data, classifying textual, or other often human-generated items into nominal categories for future analysis [4]. To account for individual subjectivites, multiple researchers often encode the same dataset. It is thus useful to have a standardized metric of evaluating the overall agreement between coders.

This metric is called *inter-rater reliability* (IRR), which considers agreements and disagreements between coders to produce an overall score, usually between 0 and 1. Though several IRR metrics exist, most commonly used is Cohen's kappa [6], which measures agreement between two coders using mutually exclusive categories. For more than two coders, Fleiss' kappa [10] achieves the same goal, while Krippendorff's alpha [18] can be used for situations where coders encode unequal amounts of data.

However, one limitation of all these metrics is their inability to accommodate taxonomies of non-mutually exclusive categories. This is a major shortfall, since non-mutually exclusive categories can arise fairly commonly in taxonomies [2,12,13]. Furthermore, no single IRR metric currently exists that accommodates a variety of qualitative coding situations, such as a variable number of coders,

coders encoding unequal amounts of data, and using non-mutually exclusive categories. Since their inceptions, commonly used IRR metrics have not evolved to accommodate such modern qualitative coding situations [1,9,14,23], and there exists a need for a novel IRR metric that is robust enough to do so, yet performing similarly enough to current metrics under their preconditions.

We introduce the *Generalized Cohen's Kappa* (GCK), which aims to achieve this. Through Monte Carlo (MC) simulations, we first establish that this metric performs similarly to Cohen's kappa under the preconditions of the latter. We then show how it can be applied to taxonomies of non-mutually exclusive categories, and show how it is robust enough to handle variable numbers of coders encoding unequal amounts of data. In addition to reporting an overall kappa score per combination and overall, GCK also reports on agreement per category and different combinations of coders. We establish GCK as a viable and reliable IRR metric.

2 Related Work

2.1 Inter-rater Reliability

Inter-rater reliability (IRR) is a statistical measurement, defined as "the extent to which the variance in the ratings is attributable to differences among the objects rated." [22] IRR metrics provide an artifact of measuring consensus (or lack thereof) between coders [19], also accomodating for chance agreements [6,14]. An IRR score of 1 is considered as perfect agreement and a score of 0 is considered agreement by chance, meaning that all raters randomly coded the data. Table 1 shows an interpretation of IRR scores, which also applies to all the IRR metrics presented in this paper.

Table 1: Interpretation of the IRR scores.

IRR score	**Interpretation**
< 0.00	Poor agreement
0.00 to 0.20	Slight agreement
0.21 to 0.40	Fair agreement
0.41 to 0.60	Moderate agreement
0.61 to 0.80	Substantial agreement
0.81 to 1.00	Almost perfect

One of the first IRR metrics is Scott's pi [21], which "measures the reliability of classifying a large number of responses into nominal scale categories." This metric considers both observed and chance agreement between two coders with equal probabilities of using any of the mutually exclusive categories.

Soon after, Cohen's kappa [6] was introduced and quickly rose to become the most used IRR metric. It is defined as "the proportion of joint judgments in

which there is agreement after chance agreement is excluded" [6]. While they are defined similarly, Cohen's kappa considers the distribution of responses between coders differently than Scott's pi.

After Cohen's kappa was introduced, there was a need for a metric that could accommodate more than two coders. Fleiss' kappa [10] resolved this need, with the constraint that all coders must have coded the same number of data points. Krippendorff's alpha [18] is robust enough to accommodate situations where coders encode different amounts of data.

A common criticism of these IRR metrics is their lack of usability for taxonomies of non-mutually exclusive categories [9,14]. To overcome this limitation, Kirilenko and Stepchenkova [17] proposed a fuzzy kappa, extending Cohen's kappa to non-mutually exclusive categories with only two users. The authors clarify that further investigation and testing is needed for this metric to be used.

Even with the existence of this metric, there is still an absence of a single IRR metric that accommodates coders working with non-mutually exclusive categories, while encoding unequal amounts of data. For instance, consider a situation where a large dataset of tweets is being encoded for emotions (which can be non-mutually exclusive i.e. people can be expressing multiple emotions in the same tweet) by a team of coders encoding unequal numbers of tweets.

Currently, no *single* IRR metric can measure the many different levels of agreement that can occur in such a situation. Brooks et al. [3] attempted to bridge this gap for their purpose. Extending their work, GCK uses a MC simulation, similar to prior work [5,16], and a novel approach of observed agreement calculation.

2.2 Cohen's kappa

Our proposed metric builds on Cohen's kappa [6], which calculates IRR between two coders using C mutually-exclusive categories over N items. It first calculates *observed agreement* (ρ_o) and *chance agreement* (ρ_e). *Observed agreement* is the proportion of items in which both coders agree, calculated by counting the number of items where both coders used the same category divided by the total number of items i.e.

$$\rho_o = \frac{\#items\ in\ agreement}{N} \tag{1}$$

Chance agreement is the proportion of items for which agreement is expected by chance. This is calculated by adding the probability that both coders use the same category, which is obtained by multiplying the probability of each coder using that category, for each category:

$$\rho_e \ = \frac{1}{N^2} \sum_{k=1}^{N} n_{k1} n_{k2} \tag{2}$$

where k is a category and n_{ki} is the number of times coder i used category k over the N items. The kappa score (κ) is the proportion of agreement after chance

agreement is removed. This is calculated using:

$$\kappa = \frac{\rho_o \ - \ \rho_e}{1 \ - \ \rho_e} \tag{3}$$

where ρ_o, $\rho_e \ \epsilon \ [0, 1]$. For complete agreement, $\rho_o = 1$ and $\rho_e \ = 0$, producing $\kappa = 1$. In unlikely cases of agreement less than chance, we obtain $\kappa < 0$.

3 Methods

3.1 Definition of Agreement

For GCK, we must establish definitions of ***agreement***, based on whether data is being coded with mutually exclusive or non-mutually exclusive categories. First, we define a ***combination*** of coders when they encode the same items. For instance, consider a team of three coders A, B and C where A and B code 100 data points, and C codes 50 of those 100 data points. Therefore, (A, B) is a combination of 100 data points, and (A, B, C) is a combination of 50 data points. When coders encode a different number of items between them, there will be multiple combinations. We thus define agreement as follows:

- ***Mutually exclusive categories***: Since coders may apply only one category for a given item, only one agreement is considered per text. If more than half the coders agree on any category, then agreement is 1, otherwise 0. This is the same definition of agreement used in Cohen's kappa [6].
- ***Non-mutually exclusive categories***: Since coders can apply more than one category to the same items and that the presence of a category does not mean the absence of other categories, agreements are considered per category. If one category is applied by more than half the coders in the combination or by none of them, then we count agreement for that code. If all coders agree that one category is not present, they reflect an implicit consensus within coders i.e. agree [11].

3.2 Calculation of Generalized Cohen's kappa

To calculate ***observed agreement*** (ρ'_o), we identify agreements across a dataset of N items depending on the type of categories used, mutually exclusive or otherwise, as described above. We calculate ρ'_o identically as Cohen's kappa.

$$\rho'_o = \frac{\#items \ with \ agreement}{N} \tag{4}$$

To calculate ***chance agreement*** (ρ_e) we use MC simulations to generate encodings of all items by computing probabilities of each coder applying each category. The probability ($\rho\_category_{ic}$) that a coder i uses a given category is calculated by dividing the number of times (n_{ic}) the coder uses the category c by the number of categories (C_i) used by the coder in all items coded.

$$\rho\_category_{ic} = \frac{n_{ic}}{C_i} \tag{5}$$

Then we compute the probability (ρ_number_{in}) that coder i uses x categories per item as the number of items n_{xi} coded with x categories divided by the total number of items coded (N_i).

$$\rho\_number_{in} = \frac{n_{xi}}{N_i} \tag{6}$$

To simulate the encoding of each datum by each coder in a combination, we randomly select the number of categories and the categories the coder will apply based on their probabilities $\rho\_category$ and ρ_number. In every simulation, we calculate variations in agreement in comparison to a **stability threshold**, explained in Section 3.3. After satisfactory simulations, we calculate total chance agreement ρ_e by averaging the simulated observed agreements of each combination weighted by the fraction of the overall dataset coded by each combination.

Finally, Generalized Cohen's kappa (κ') is calculated using observed and chance agreement using equation 3 from Cohen's kappa:

$$\kappa' = \frac{\rho'_o \ - \ \rho_e}{1 \ - \ \rho_e} \tag{7}$$

3.3 Execution time and experimental setup

The execution time of the algorithm depends on different factors, such as number of iterations needed to achieve the stability threshold, and number of unique combinations. The simulated datasets vary in size between 2,168 to 25,000 lines. The average execution time for all datasets is 9.32 seconds with a standard deviation of $\sigma = 1.45$ seconds. All the tests were executed on an AMD EPYC 7302P CPU @3GHz with 64 GB RAM under Debian 10 distribution. The GCK algorithm was implemented in Python3.

For MC simulations, we determined the *stability threshold* i.e., the value below which future simulations will not have significant variation, a common practice when using MC simulations [20]. After extensive testing, we set the stability threshold to 1×10^{-4}, ensuring stability in chance agreement to the third decimal place with an average standard deviation of 2×10^{-3}.

Generating simulated datasets. We simulated several datasets, each with different characteristics, for extensive testing of the GCK. With different permutations of the parameters (number of coders, number of categories, and number of items encoded) with initial agreements from 0.0 to 1.0 in increments of 0.05, we obtained 21 simulated datasets for each set of parameters.

With these simulated datasets, our goal is to observe how chance agreement and kappa scores change when observed agreement changes in the GCK algorithm, and to understand the impact individual parameters have on overall kappa score. We will thus be able to understand the behavior of GCK in different settings and how it compares to Cohen's kappa.

3.4 Dataset and Parameters

Our data is a subset of 2000 reviews, from Yin et al.'s [24] fanfiction reviews dataset consisting over 176 million reviews, manually qualitatively coded by 7 different coders using two different categories: *emotions*, and *valence*. Emotion categories are non-mutually exclusive, and valence is mutually exclusive.

To compare GCK with Cohen's kappa, we tested them under the preconditions of the latter: mutually exclusive categories (*valence*), and two coders encoding 2000 reviews. We observe $\binom{7}{2} = 21$ combinations of coders exist, as shown in Table 2, and calculate GCK values for each combination.

Table 2: Pairwise combinations $P_{(i,j)}$ of coders i and j.

		Coder id						
		1	2	3	4	5	6	7
Coder id	1							
	2	$P_{(1,2)}$						
	3	$P_{(1,3)}$	$P_{(2,3)}$					
	4	$P_{(1,4)}$	$P_{(2,4)}$	$P_{(3,4)}$				
	5	$P_{(1,5)}$	$P_{(2,5)}$	$P_{(3,5)}$	$P_{(4,5)}$			
	6	$P_{(1,6)}$	$P_{(2,6)}$	$P_{(3,6)}$	$P_{(4,6)}$	$P_{(5,6)}$		
	7	$P_{(1,7)}$	$P_{(2,7)}$	$P_{(3,7)}$	$P_{(4,7)}$	$P_{(5,7)}$	$P_{(6,7)}$	

4 Results

In our experiments, we study how GCK performs in comparison to Cohen's kappa under the latter's preconditions. We establish that they perform similarly, and then extend GCK to situations using non-mutually exclusive categories.

4.1 Comparison of Chance Agreements

The primary way in which GCK differs from Cohen's kappa is in its calculation of observed agreement. GCK computes chance agreement in the same way that Cohen's kappa does. Therefore, before comparing the performances of GCK and Cohen's kappa, we first need to establish that our algorithm for calculating chance agreement via MC simulations does indeed compare well to the chance agreement of Cohen's kappa algorithm.

We computed chance agreements on the data referred to in Section 3.4, using the pairwise combinations of coders in Table 2. Because of the randomness in the MC simulations, we calculate chance agreement 10 times per pairwise combination and report means. Figure 1 shows the chance agreement scores obtained by the two metrics. We observe that the results are similar in every pairwise combination of coders. The average difference between chance agreements between both metrics is 2×10^{-3} with a standard deviation of $\sigma = 1 \times 10^{-3}$, thus confirming the validity of our chance agreement calculations.

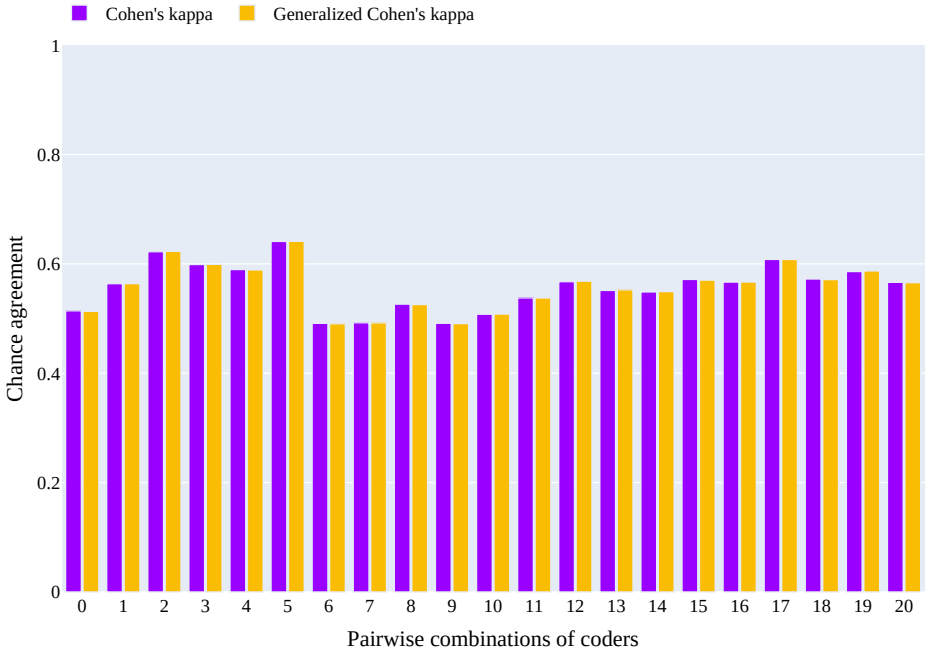

Fig. 1: Comparison of the chance agreement by pairwise combinations of coders. The x-axis represents the 21 different pairwise combinations while the y-axis represents the chance agreement. Both metrics are color-coded, Cohen's kappa in purple and Generalized Cohen's kappa in yellow.

4.2 Comparison to Cohen's kappa

To compare the overall performances of GCK and Cohen's kappa, we use the same data referred to in Section 3.4, and the pairwise combinations of coders in Table 2. We calculate observed agreement and GCK 10 times per pairwise combination and report means. Figure 2 shows kappa scores for Cohen's kappa and GCK. We observe similar results for both metrics with an average difference of 2×10^{-3} and a standard deviation of $\sigma = 1 \times 10^{-3}$.

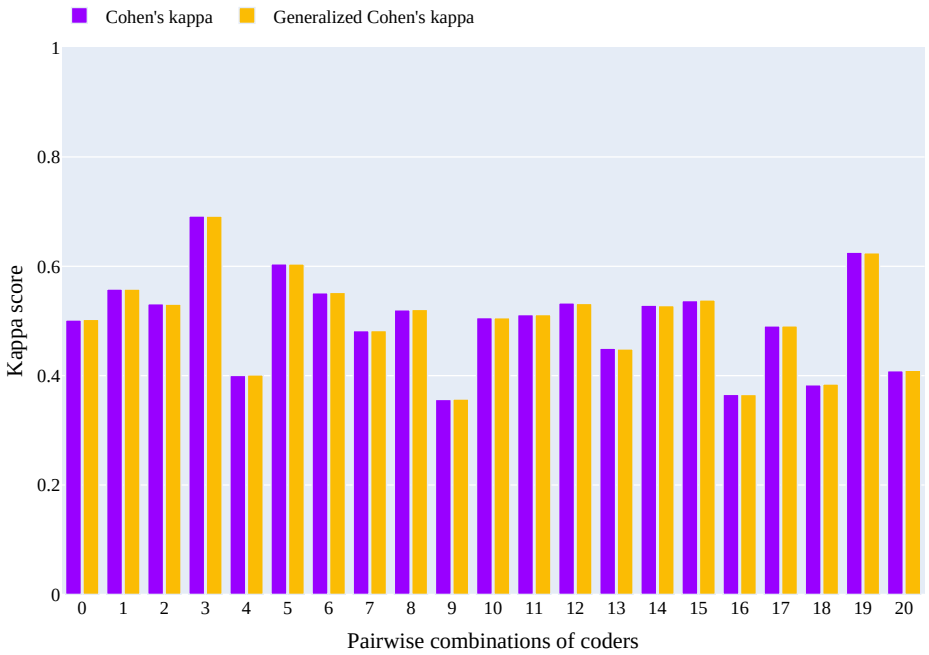

Fig. 2: Comparison of kappa scores by pairwise combinations of coders. The x-axis represents the 21 different pairwise combinations while the y-axis represents the kappa score. Both metrics are color-coded, Cohen's kappa in purple and Generalized Cohen's kappa in yellow.

4.3 Applying GCK to extended situations through simulation experiments

In this section, we extend GCK to situations involving non-mutually exclusive categories, and test it by varying the number of coders and in cases of unequal amounts of coded data. We first examine the relationship between observed agreements and the resultant values of Cohen's kappa, as depicted in Figure 3.

Since we intend to understand the behavior of GCK with non-mutually exclusive categories, the probability of applying one, two, up to n categories to one item by one coder was obtained from the manually qualitatively coded data from the fanfiction reviews using the non-mutually exclusive taxonomy of *emotions*. We simulated a variety of datasets for different parameters, as mentioned in Section 3.3, and generate random encodings per coder and item. For each item, we calculate agreements and categorize them into levels of agreement, ranging from 0 to n, where n is the number of categories. Table 3 provides an example of levels of agreement for $n = 3$.

As before, to account for the randomness in the simulation, all results presented here will be means of 10 executions with different seeds under the same parameters. Different experiments vary different parameters of the simulated

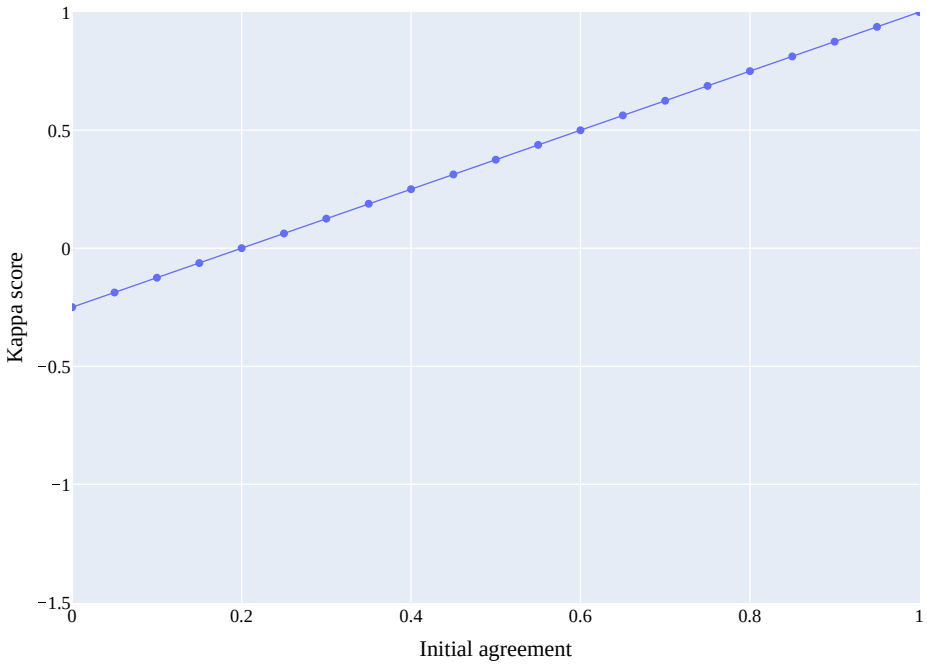

Fig. 3: Cohen's kappa scores varying with observed initial agreements. The two variables are seen to have a linear relationship.

datasets to understand how they affect the final Generalized Cohen's kappa score. We applied the GCK algorithm on each dataset with varying initial agreements. Figure 4 displays the behavior of the GCK with 3 coders, 5 non-mutually exclusive categories, and 1000 items.

Varying number of coders. We varied the number of coders from 2 to 5, maintaining the number of texts and categories constant at 1000 texts and 5 mutually-exclusive categories. We simulated 21 datasets for each set of parameters with initial agreement from 0.0 to 1.0. We expected that the number of coders would change the minimum value of the kappa score as different numbers of coders have different likelihoods of obtaining agreement with the same number of categories. These results are represented in Figure 5.

Varying number of categories. Similar to the previous subsection, we varied the number of non-mutually exclusive categories while maintaining the number of texts as 1000 and the number of coders as 3. The numbers of categories tested are 3, 5, and 7. These results are represented in Figure 6.

Varying number of items coded. Finally, we varied the number of encoded items with the number of coders set to 3 and the number of non-mutually ex-

Table 3: Description of the level of agreements for 3 categories.

Level of Agreement	Description
0	Every coder disagrees on every category in the text.
1	Coders agree only on one out of three categories
2	Coders agree on two out of three categories
3	Coders agree in all categories

clusive categories to 5. We expect that this would have an impact on the kappa scores as the probabilities of achieving agreement or disagreement remain the same regardless of the number of texts. Figure 7 shows the kappa scores of the simulated datasets with different initial agreements and number of texts.

5 Discussion

5.1 Establishing GCK as a viable alternative to Cohen's kappa

As discussed in Section 4, the primary objective during testing was to first establish that, under the preconditions of applying Cohen's kappa, GCK would perform similarly. We show this first by testing the accuracy of the chance agreements generated by MC simulations and comparing those values to the chance agreements calculated by the Cohen's kappa algorithm, and then compare the overall metrics.

In Section 4.1, we demonstrate that the chance agreements calculated by the GCK algorithm are similar to those reported by the Cohen's kappa algorithm. This validates that MC simulations are a good approximation of chance agreement, and allows us to proceed with the overall comparison of the two metrics. We then observe, in Figure 2, the values are very similar to each other. In combination with the results in Section 4.1, these results validate our method of calculating observed agreement and demonstrate that under the preconditions of applying Cohen's kappa, the GCK performs similarly.

We also observe that both Cohen's kappa and GCK have a linear relationship with observed agreement, as shown in Figures 3 and 4 respectively. This further strengthens our claim of interchangeability between those metrics under the conditions of Cohen's kappa.

5.2 A metric for a wide range of situations

Having demonstrated that the MC simulations provide good estimates of chance agreement in Section 4.1 and that GCK performs similarly to Cohen's kappa in Section 4.2, we now aim to extend GCK to situations involving non-mutually exclusive categories. We confirm the stability of GCK by testing with variable numbers of coders encoding unequal amounts of data, varying each of these individually while keeping the others consistent. We do so by demonstrating that the linear relationship between observed agreement and GCK does not

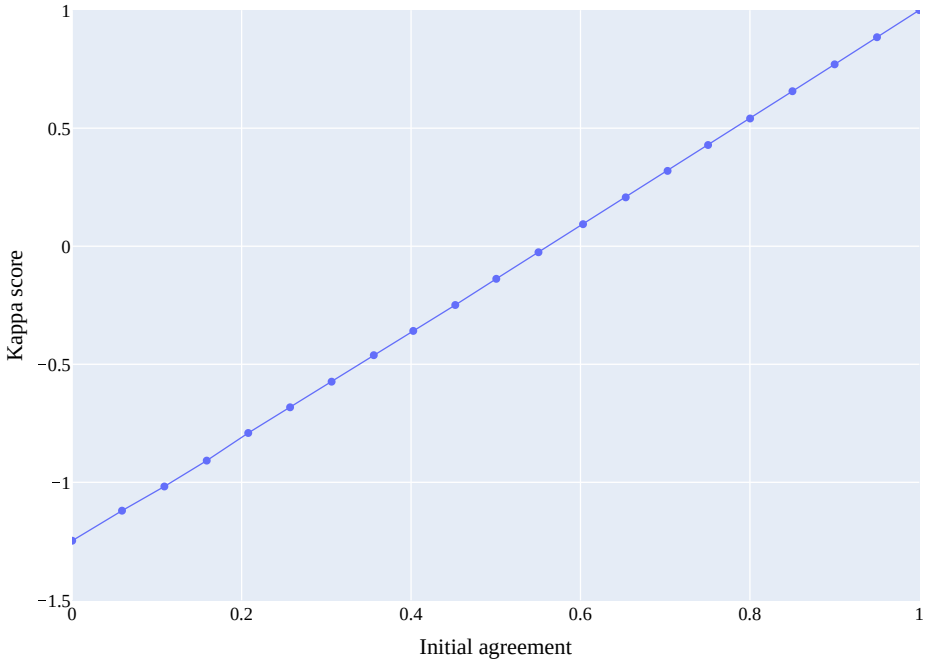

Fig. 4: Generalized Cohen's kappa scores of simulated datasets with varying initial agreement.

change when considering non-mutually exclusive categories, and while changing other parameters.

In Figure 4, we observe a perfectly linear relationship between observed agreements and GCK values when testing on data coded with non-mutually exclusive categories, with an observed agreement of 1.0 generating a GCK score of 1.0, similar to the situation in Figure 3.

Figure 5 shows GCK scores varying linearly as expected, when varying numbers of coders. The similarity between trends for 4-5 coders are comparable as those for 2-3 coders. This can be explained by the majority within those numbers of coders: 2 for 2-3 coders and 3 for 4-5 coders. This affects how agreement is counted in each case e.g. if we have a majority vote within 4 coders, adding one extra will not change the result.

From Figure 6, we observe that the number of categories does not drastically change GCK scores. The range of values is the same in all cases and there is a slight change in the curve of each number of categories. This can be explained. With 3 categories and 3 coders, it is difficult to have an agreement level of 0, because each coder can apply multiple categories to the same text and agreement by absence is more likely to occur, thus decreasing the kappa score. With 5 categories, there is a balance between the number of categories and coders where it is less likely to have a higher chance agreement, resulting in a higher score.

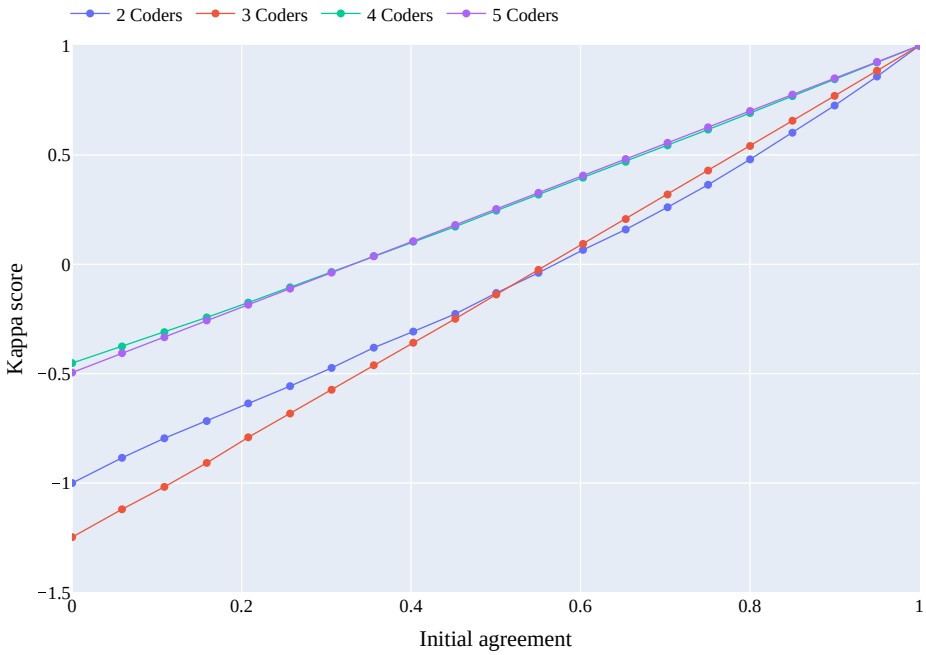

Fig. 5: Generalized Cohen's kappa scores of simulated datasets with varying initial agreement and number of coders. The number of coders is color-coded and represented with labels.

Finally, with 7 categories, the choices are much more, and it is more possible to have an agreement by absence, thus decreasing the kappa score. This shows how, based on likelihood, the approximation of chance agreement of MC simulations remove agreement that would occur by chance only. Thus, GCK behaves well with varying numbers of non-mutually exclusive categories.

Finally, the occluding nature of all the lines in Figure 7 indicate that the number of texts is not a determining factor, and that GCK still maintains a linear relationship to observed agreements. Therefore, based on these extensive tests, we determine that Generalized Cohen's kappa can be used under the preconditions of Cohen's kappa, as well as in situations consisting of variable numbers of coders using non-mutually exclusive categories on unequal numbers of items.

5.3 Explaining dissimilarities in edge values

The main observed difference between GCK and Cohen's kappa is for edge values. From Figures 3 and 4, we observe that though both metrics show similarly linear relationships with observed agreement, they differ in their ranges of values. While GCK shows a minimum value of -1.238, Cohen's kappa's minimum is -0.249. This dissimilarity can be explained by the difference in the procedure of agreement

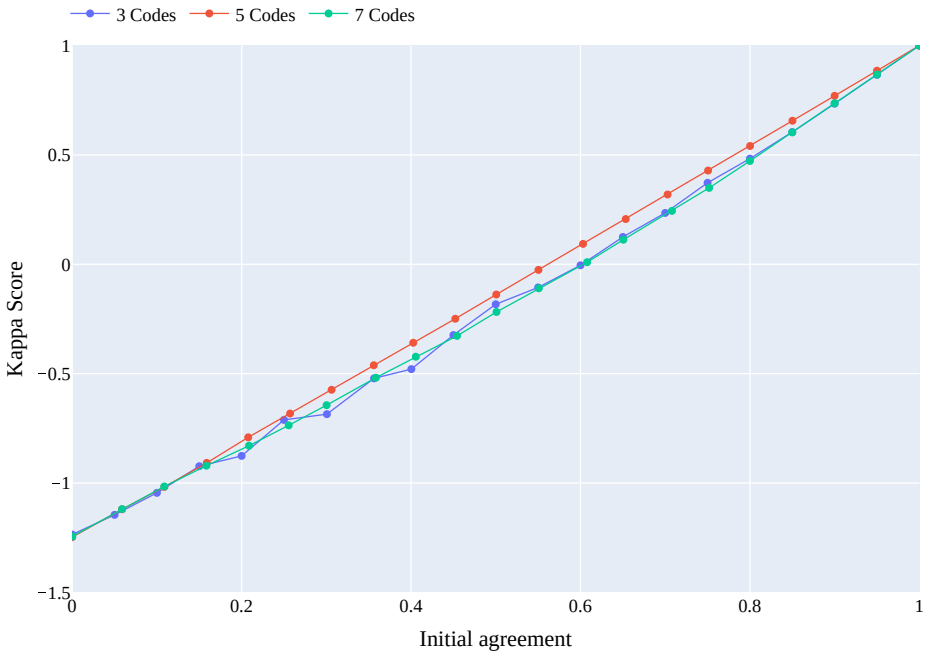

Fig. 6: Generalized Cohen's kappa scores of simulated datasets with varying initial agreement and number of categories. The number of categories is color-coded and represented with labels.

calculation. While Cohen's kappa considers coders agreeing or disagreeing on each datum, GCK's agreement is considered both per datum and per category. With 3 coders and 5 codes, agreement can go from 0-5 in each datum, making it much less likely to have 0 agreement in every datum than Cohen's kappa. Thus, chance agreement is likely to be higher for GCK since the MC simulations will have higher agreements, ultimately resulting in lower GCK scores than Cohen's kappa with observed agreements closer to 0.

5.4 A detailed understanding of agreement

One of the disadvantages of current IRR metrics is that one score is awarded to the whole qualitative coding process, which can be unrepresentative of agreement distributions among coders and categories. We overcome this drawback by reporting additional data of changes in agreement between different combinations of coders and categories. This can be helpful in agreement discussions to obtain deeper understandings of disagreement, removing the need of guessing which category or combination of coders did or did not agree.

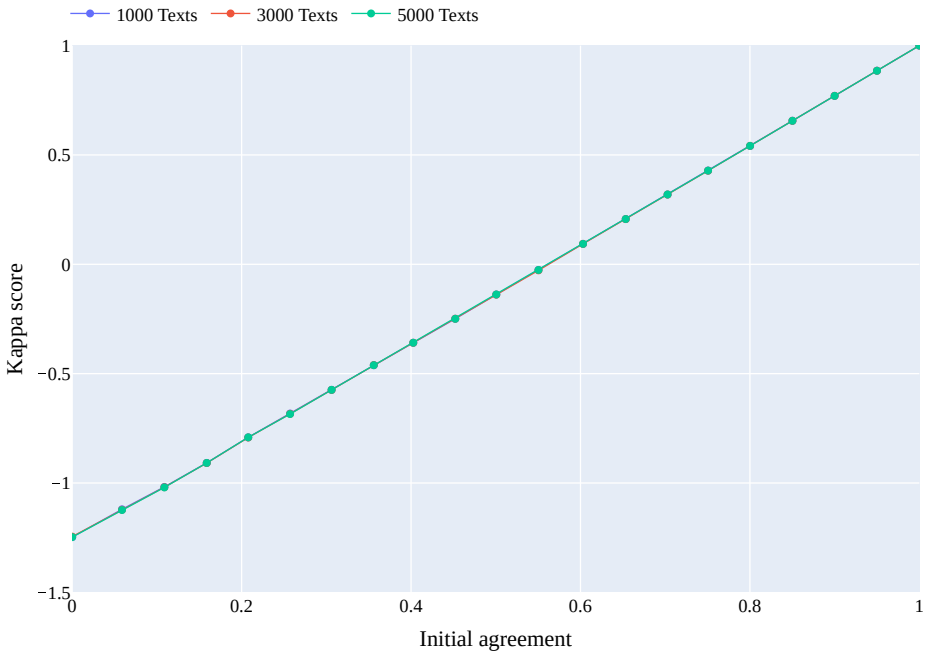

Fig. 7: Generalized Cohen's kappa scores of simulated datasets with varying initial agreement and number of texts. The number of texts is color-coded and represented with labels.

6 Conclusion and Future Work

Through this work, we present a novel algorithm to calculate inter-rater agreement: the Generalized Cohen's kappa. We demonstrate that it performs similarly to the widely-accepted Cohen's kappa under the latter's preconditions, and then demonstrate its viability in situations involving non-mutually exclusive categories. We find GCK to be stable under a variety of qualitative coding situations, such as using non-mutually exclusive categories being used by large teams of coders on large datasets.

We acknowledge that some known issues of Cohen's kappa are also carried forward into GCK. Some examples are reporting different agreements in data with the same initial agreement but different distribution of the data [15], and having high rates of Type I errors, meaning it reports higher scores than actual agreements [7,8]. We also note that applying a numeric metric to the qualitative coding process is not ideal, since it obscures so many aspects of the process. We aim to provide more intermediate information to overcome this issue, but there is still some loss of information in computing GCK, as with any metric.

We believe that using visualizations to supplement GCK scores to provide a more comprehensive understanding of the coding process, offering support to

coders to better understand their process and improve their inter-rater reliability and discussion agreement. We plan to explore this question in future works. We also intend to examine the potential for GCK to provide intermediate instances of agreement by different categories and combinations of coders, helping researchers assess the qualitative coding process at different stages. This can help assist researchers in refining open-coding taxonomies, or identifying challenges with a particular combination of coders or a particularly divisive category.

We believe that this novel metric can adapt to different situations and will allow for a better assessment of inter-rater reliability during and after the qualitative coding process. Generalized Cohen's kappa can assess agreement in different settings present in qualitative coding, such as fixed or varying numbers of coders and mutually and non-mutually exclusive categories, or combination of them, making this metric adaptable and helpful for agreement discussions.

References

1. Andrés, A.M., Marzo, P.F.: Delta: A new measure of agreement between two raters. British journal of mathematical and statistical psychology **57**(1), 1–19 (2004)
2. Bazeley, P.: Issues in mixing qualitative and quantitative approaches to research. Applying qualitative methods to marketing management research **141**, 156 (2004)
3. Brooks, M., Kuksenok, K., Torkildson, M.K., Perry, D., Robinson, J.J., Scott, T.J., Anicello, O., Zukowski, A., Harris, P., Aragon, C.R.: Statistical affect detection in collaborative chat. In: Proceedings of the 2013 conference on Computer supported cooperative work. pp. 317–328 (2013)
4. Charmaz, K.: Constructing grounded theory: A practical guide through qualitative analysis. sage (2006)
5. Cicchetti, D.V., Shoinralter, D., Tyrer, P.J.: The effect of number of rating scale categories on levels of interrater reliability: A monte carlo investigation. Applied Psychological Measurement **9**(1), 31–36 (1985)
6. Cohen, J.: A coefficient of agreement for nominal scales. Educational and psychological measurement **20**(1), 37–46 (1960)
7. Eagan, B., Brohinsky, J., Wang, J., Shaffer, D.W.: Testing the reliability of inter-rater reliability. In: Proceedings of the Tenth International Conference on Learning Analytics & Knowledge. pp. 454–461 (2020)
8. Eagan, B.R., Rogers, B., Serlin, R., Ruis, A.R., Arastoopour Irgens, G., Shaffer, D.W.: Can we rely on irr? testing the assumptions of inter-rater reliability. In: International Conference on Computer Supported Collaborative Learning. pp. 529–532 (2017)
9. Epstein, M.H., Harniss, M.K., Pearson, N., Ryser, G.: The behavioral and emotional rating scale: Test-retest and inter-rater reliability. Journal of Child and Family Studies **8**(3), 319–327 (1999)
10. Fleiss, J.L.: Measuring nominal scale agreement among many raters. Psychological bulletin **76**(5), 378 (1971)
11. Fleiss, J.L.: Measuring agreement between two judges on the presence or absence of a trait. Biometrics pp. 651–659 (1975)
12. Ghosh, S., Figueroa, A.: Establishing tiktok as a platform for informal learning: Evidence from mixed-methods analysis of creators and viewers. Proceedings of the 56th Hawaii International Conference on System Sciences pp. 2431–2440 (2023)

13. Ghosh, S., Froelich, N., Aragon, C.: "i love you, my dear friend": Analyzing the role of emotions in the building of friendships in online fanfiction communities. In: Proceedings of the 15th International Conference on Social Computing and Social Media in the context of the 25th International Conference on Human-Computer Interaction (HCI International). Springer (2023)
14. Gisev, N., Bell, J.S., Chen, T.F.: Interrater agreement and interrater reliability: key concepts, approaches, and applications. Research in Social and Administrative Pharmacy **9**(3), 330–338 (2013)
15. Gwet, K.: Kappa statistic is not satisfactory for assessing the extent of agreement between raters. Statistical Methods for Inter-Rater Reliability assessment **1**(6), 1–6 (2002)
16. Gwet, K.L.: Computing inter-rater reliability and its variance in the presence of high agreement. British Journal of Mathematical and Statistical Psychology **61**(1), 29–48 (2008)
17. Kirilenko, A.P., Stepchenkova, S.: Inter-coder agreement in one-to-many classification: fuzzy kappa. PloS one **11**(3), e0149787 (2016)
18. Krippendorff, K.: Computing krippendorff's alpha-reliability (2011)
19. McDonald, N., Schoenebeck, S., Forte, A.: Reliability and inter-rater reliability in qualitative research: Norms and guidelines for cscw and hci practice. Proceedings of the ACM on Human-Computer Interaction **3**(CSCW), 1–23 (2019)
20. Owen, A.: Monte Carlo Theory, Methods and Examples. Stanford (2013)
21. Scott, W.A.: Reliability of content analysis: The case of nominal scale coding. Public opinion quarterly pp. 321–325 (1955)
22. Tinsley, H.E., Weiss, D.J.: Interrater reliability and agreement. In: Handbook of applied multivariate statistics and mathematical modeling, pp. 95–124. Elsevier (2000)
23. Uebersax, J.S.: Diversity of decision-making models and the measurement of interrater agreement. Psychological bulletin **101**(1), 140 (1987)
24. Yin, K., Aragon, C., Evans, S., Davis, K.: Where no one has gone before: A metadataset of the world's largest fanfiction repository. In: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems. pp. 6106–6110 (2017)