A Less Overconservative Method for Reliability Estimation for Cohen's Kappa: Extended Version

Matt He[1], Ryan Baker[2], Stephen Hutt[3], and Jiayi Zhang[2]

[1] Northfield Mount Hermon, Gill, MA 01354, USA
[2] University of Pennsylvania, Philadelphia, PA 19104, USA
[2] University of Denver, Denver, CO 80204, USA

Abstract. Cohen's Kappa has been used in interrater reliability calculation for decades, often for small samples. Recently, researchers within the quantitative ethnography community have argued that Cohen's Kappa cannot validly be used without much larger samples [1, 2]. Within this paper, we argue that this conclusion is drawn based upon assumptions that are overly conservative. For example, not taking into account the amount of error in a Kappa estimate and using a statistical significance criterion rather than a statistical power analysis criterion for an analysis that is conceptually analogous to power analysis. We present a Monte Carlo analysis that assesses inter-rater reliability based on distance between the population Kappa and threshold Kappa (i.e., the degree of error), for a range of population Kappa values, threshold Kappa values, and sample sizes. Our findings indicate that Kappa can reasonably be used at the sample sizes often used in practice, either by raising threshold Kappa or by adopting the level of stringency used in statistical power analysis.

Keywords: Cohen's Kappa, Sample Size Calculation, Monte Carlo Analysis.

1 Introduction

One of the key first steps in the process of conducting epistemic network analysis is to transform data into a set of elements whose interrelationships can be visualized in one or more epistemic network graphs [3]. Though these elements can come from a variety of different sources, the most common source comes from the coding of data, typically a systematic process where data is analyzed qualitatively for meaning and assigned categorical codes [8, 10]. Today, a significant proportion of the data sets analyzed through epistemic network analysis use machine-coding of textual data, involving natural language processing [6]. However, even when analyzing data coded by natural language processing, it is still important within quantitative ethnography to align those codes to theory and/or human understanding [5, 7]. Creating this alignment, therefore, often still involves first obtaining a training data set of human-coded examples to bootstrap the process of machine learning [6] (although recent work has also investigated entirely

bottom-up methods of deriving codes from natural language data -- see [8] -- where alignment to theory would therefore occur later in the process).

Whether the eventual codes analyzed are produced by human coders or natural language processing, there is typically an initial phase of human coding. In this phase, the quantitative ethnography community -- like the qualitative research community before it (e.g., [9, 10]) -- often relies upon a process of validating that human coders can come to agreement on the mapping between codes and specific examples. In the key stage of this process, referred to as The Common Method for IRR Measurement, human coders separately code the same set of examples and then check how well they agree [1, 2].

The traditional and still most frequent method across fields for validating whether coders agree about specific examples better than chance is to use Cohen's Kappa [11]. Cohen's Kappa computes a base rate for the degree of agreement that could be expected by chance simply due to each coder's proportion of coding each category in the data. It then compares the actual degree of agreement to this base rate, to obtain a base-rate adjusted estimate of agreement.

There is an extensive literature of critique of Cohen's Kappa, which we will review in the following section. Perhaps most notable to the Quantitative Ethnography community, Eagan and colleagues [1, 2] argue that Kappa has an unacceptably high Type I error rate when used in the fashion it is commonly used within the Quantitative Ethnography community and related communities. They recommend instead using a different method, Shaffer's Rho, for inter-rater analysis.

In this paper, we offer a critique of the analysis of Kappa and other metrics offered by Eagan and colleagues [1, 2], comparing their approach's stringency and assumptions to power analysis. We find that their approach is analogous to statistical power analysis at two key steps, but is substantially more conservative (than statistical power analysis) at each of these steps. We then propose an alternative to the analysis conducted in this earlier work, that attempts to align more closely to the assumptions and degree of conservatism of statistical power analysis. This analysis, like the analysis in [1, 2] takes the form of a Monte Carlo analysis, but with different assumptions. Our analysis finds, as Eagan and colleagues do, that current practice is underpowered in some cases (i.e., specific choices of threshold and sample size seen in the literature are underpowered), but substantially less often than their analysis would suggest. We suggest a way for selecting appropriate sample sizes and a route forward for the continued principled use of Kappa.

2 Kappa: An Often-Criticized, Still-Used Metric

Cohen's Kappa [11] has been used for several purposes in the social sciences (and in research more broadly). Two of the most frequent uses in recent years are (1) to evaluate the degree of inter-rater reliability in qualitative data coding and (2) to evaluate models in machine learning. Kappa's popularity can be shown in the high levels of citation it has received – Cohen's 1960 paper [11] has been cited 14,500 times on Google Scholar in the last five years alone, with many more citations for other articles on the metric and substantial use of Kappa without citation.

Despite this enduring popularity, there have been concerns about Kappa from fairly shortly after its initial use, and work to extend Kappa through approaches such as Fleiss' Kappa [12], which allows for a larger number of raters, and Weighted Kappa [13], which works on data that is ordinal or where some labels are considered more related than others. In particular, Kappa has undesirable performance for very high base rates and very low base rates [14, 15], a finding replicated empirically by [16], who find that similar problems also impact several other widely used metrics, such as accuracy and F-score. Another key issue with Cohen's Kappa is that there is no agreed single way to compute a standard error for this metric [17]. This is because the probability density function is asymmetric in most cases, and shifts based upon the base rate, making it difficult to conduct statistical analysis. This discovery led to concern about initial table-based and mathematical methods for sample size selection for Kappa (cf. [18]). Researchers analyzing the properties of Cohen's Kappa have, therefore, often resorted to the use of simulations and Monte Carlo analyses (e.g., [1, 2, 16] and this paper).

More recently, alternatives such as the Matthews Correlation Coefficient/phi [19], PABAK [20], and Shaffer's rho [5] have emerged, with Shaffer's rho gaining particular prominence in the quantitative ethnography community. Metric-based approaches such as the Matthews Correlation Coefficient and PABAK provide an alternate way to mathematically compute agreement, whereas Shaffer's rho is a computational method of conducting a statistical significance test of agreement comparing coder Kappa to simulated Kappa values.

However, despite the availability of alternatives to Kappa, and general critiques of its use, Kappa continues to be used for several reasons. It remains expected by many reviewers and in many publication venues, due to its simplicity, standard interpretation, and predictable behavior for base rates that are neither extremely high nor extremely low. For this reason, it is also incorporated in the nCoder tools used within the quantitative ethnography community [6]. However, when researchers decide to use Kappa despite its limitations, question remains as to what sample size is appropriate [1, 2], which we detail in the next section.

3 An Examination of the Methods in Eagan et al.

Eagan and colleagues [1, 2] present Monte Carlo analyses where they tested whether a sample's Kappa is higher than the full population's Kappa. They repeatedly created data sets with 10,000 simulated data points (each having two codes) and sample from that data set to obtain sample subsets (referred to in those papers as test sets). In conducting each simulation, they specified a base rate for the simulated coders, and a sample set size. They then repeatedly generated 10,000 simulated data points and sampled from those codes, choosing sample set sizes of 20, 40, 80, 200, 400, 800, 2000, 4000, and 8000. They conducted 12,000 iterations of their analysis. True Kappa values (for the full population) were allowed to vary (i.e., a pre-chosen Kappa was not targeted), and were reported to vary from 0.3 to 1.0 [1]. A target threshold Kappa was then selected -- 0.65 in [2] (Eagan et al., 2017) and 0.7 in [1] (Eagan et al. 2020). Among each set of the 12,000 iterations, Eagan and colleagues counted the number of cases where

population Kappa was below threshold, and sample Kappa was above threshold. These were treated as false positives, and the proportion of these cases (count divided by 12,000) was treated as the Type I error rate. They then argued that any case where this Type I error rate was over 0.05 would represent evidence that the assessment of Kappa was flawed.

According to this metric, Eagan and colleagues argue that Kappa produces erroneous results more than 5% of the time for sample sizes under 400 [2] or 2000 [1], with the difference between these two values based on differences in base rate (an issue more systematically studied in [2] than in [1]). They therefore argue that the common practice of testing inter-rater reliability using Kappa on samples typically much smaller than these values is flawed and should be abandoned.

Within this method, there are thus two steps in evaluating performance across a set of simulations: first, determining how often population Kappa is below threshold and sample Kappa is above threshold. Second, determining if this proportion of cases is above 5%. In considering the first step in Eagan et al.'s process, we can note that -- like any sampling procedure -- a Kappa estimate from a small sample represents the central point on a probability density function of population Kappa values. Any Kappa estimate generated from a sample is simply the most likely value, and the true value is likely to deviate from that true value. Should a researcher or reviewer consider a coding scheme invalidated if the population Kappa is 0.69 and the sample Kappa is 0.71? Currently, the approach in [1, 2] treats this case the same as a case where population Kappa is 0.30 and sample Kappa is 0.71. Is the goal of selecting a Kappa threshold for a sample to determine if the true population Kappa is over that exact threshold (even though that exact threshold may vary across research communities and even between [1, 2])? Or is the goal to be confident that the true population Kappa cannot be a much lower level? There are probably good reasons to view a population/sample difference of 0.41 as a much more serious problem than a population/sample difference of 0.02. At the same time, Eagan et al's approach treats a difference in Kappa between, say, 0.64 and 0.23, as wholly non-problematic, since both are below threshold.

In considering the second step, we can note a clear analogy to statistical significance testing, with the adoption of the criteria that Type I error be less than 0.05; however, there are important differences between the two. Statistical significance testing seeks to determine whether a result of a certain magnitude could have been obtained were only chance variation occurring. However, Kappa itself does not represent a statistical test -- it represents something different, a strength of association. Strength of association is typically considered a measure of effect size rather than statistical significance.

To review, then, the method in [1, 2] sets a very stringent criterion at each of its two steps of evaluating performance: (1) considering a Kappa sample invalid based on an arbitrarily small difference between population and sample Kappa, so long as that difference crosses a specific threshold, and (2) allowing this threshold-crossing to occur no more than 5% of the time.

It is worth noting that substantially less stringent criteria have been adopted by other methods to choose sample size. Take the case of statistical power analysis to select sample size for statistical testing [21]. This task is analogous in some ways to our selection of a sample size for Kappa (in fact, this exact procedure was used by [18], but

relied upon a mathematical calculation of Kappa standard error later discovered to be unreliable [17]). The goal of statistical power analysis is to see if (1) a test with a known true effect size will obtain statistical significance, with (2) frequency over a pre-chosen threshold (typically 80%).

We can compare each of the steps of Eagan et al.'s performance evaluation procedure to statistical power analysis. For step 1 of their process, the statistical test most analogous to the case evaluation procedure in [1, 2] is the Wilcoxon rank-sum test [22], which looks for whether one sample's values are typically higher than another sample's values. A statistically significant value can be obtained for Wilcoxon without requiring that fewer than 5% of comparisons be in favor of the lower-valued sample; as such, this first element of [1, 2] is much more stringent than traditional statistical significance testing.

For step 2 of Eagan et al.'s process, statistical power analysis typically looks for whether a significant result is seen at least 80% of the time, which can be inverted to a failure mode occurring less than 20% of the time. By contrast, Eagan et al. [1, 2] look for whether a failure mode (a significant result) is seen at least 5% of the time. As such, the second step of Eagan et al.'s procedure is four times as stringent as statistical power analysis.

In concluding this section, we note that each of the two steps of Eagan et al.'s approach is substantially more stringent than statistical power analysis. In the following section, we propose a method that more explicitly considers the degree of difference between population Kappa and sample Kappa. We also consider the implications of using a second-step stringency criterion in line with statistical power analysis rather than statistical significance testing.

4 This Paper's Methods

Within this paper, we attempt to analyze the risks of false positives when obtaining Kappa values from small subsets of data, using a method attempting to achieve a level of conservatism more analogous to statistical power analysis. Our overall process is similar in structure to Eagan et al.'s [1, 2]. First, we create a simulated data set; then, we sample from that simulated data set; finally, we test whether that simulated data set represents a false positive. In all simulations, we sample from a population size of one million, and for all sets of parameters, 100,000 iterations are conducted.

In any given simulation run, we use three parameters: a sample size, a threshold Kappa (false positives have Kappa over threshold), and a value of Cohen's Kappa for the entire simulated population, which is selected in relation to the threshold Kappa. For example, we might select a threshold Kappa of 0.65 (as in [2]) and a population Kappa 0.2 less than the threshold Kappa, making the population Kappa parameter 0.45. Note that these parameters are different from the parameters used by Eagan and colleagues; they select a threshold but allow the population Kappa to vary.

We generate the population by first creating 200 data points with random binary codes (i.e., each of the two coders' decisions are selected randomly). We then iteratively add to the data set by repeatedly comparing the population's Kappa to the target Kappa.

If the population's current Kappa is higher than the population's target Kappa, we create a data point with disagreement between the raters (selecting the direction of disagreement randomly). If the current Kappa is lower than the target Kappa, then we create a data point with agreement (randomly selecting whether the agreement is code-present or code-absent). In the very large data set size used, this procedure reliably creates populations with the target Kappa.

After creating the population, we then repeatedly sample random data points from this population, with a preselected sample size (a parameter, as noted above). We conduct 100,000 iterations for each simulation. Within each of these 100,000 iterations, we test whether the sample Kappa is above or below the threshold Kappa. Choosing both the population Kappa and threshold Kappa (in relation to each other) enables us to avoid treating small levels of variation as a false positive. We can then calculate what proportion of time we have a sample Kappa above threshold, despite having a population Kappa substantially below threshold.

Several sets of simulations were run, using the following possible parameters:

- Threshold Kappa: 0.6, 0.65, 0.7, 0.75, 0.8 -- a larger set of values is used than in Eagan et al. (2017, 2020) to test for the range of values of Kappa preferred in a range of research areas quantitative ethnography is applied to.
- True population Kappa: Threshold - 0.05, Threshold - 0.1, Threshold - 0.2, Threshold - 0.3
- Sample size: 20, 40, 60, 80, 100, 200, 400, 500, 800, 1000, 2000, 4000, 8000.

For example, in a given run of 100,000 iterations, we might have selected a sample size of 60, a threshold Kappa of 0.7 (as in [1]), and a population Kappa 0.05 below that -- i.e., 0.65. This cell is in italic boldface in Table 1.

The full software used in this paper is available, open-source and free, at bit.ly/3wRhrt4

5 Findings

Having created these simulations, we can now check for the proportion of time a specific test produces a Kappa above threshold, despite having a lower true population Kappa. We consider first a sample size of 60 (Table 1) – a small dataset, but of a size often seen in inter-rater reliability checks within the quantitative ethnography and learning analytics communities. Table 1 reports the proportion of samples that have a Kappa value above the threshold, when the true value (population Kappa) is some amount less. We first note that for this sample size, approximately 30% of samples had a Kappa value greater than .05 larger than the true population Kappa across all thresholds. This suggests that there is a fairly high probability (around 30%) that a sample Kappa value barely over threshold may represent a population Kappa value barely below threshold, regardless of what that threshold is. As we increase the distance between threshold Kappa and the population Kappa (from .1 to .3), we consistently see a reduction in the number of samples that meet the threshold, with less than 1% of our samples achieving a threshold Kappa value .3 or more above the true population Kappa value. Though we

observe some variation across the different threshold values, results are somewhat consistent between thresholds of .6 and .75; that is, samples are similarly likely to have error/noise regardless of the threshold value at these levels. We do see a reduction in error for the threshold of .8, suggesting a higher threshold Kappa value may be more robust to error. Looking at the table, one notes that if a researcher were to select a level of conservatism comparable to power analysis, even a small sample of 60 data points is sufficient to be confident that a threshold is unlikely to represent a population Kappa more than 0.1 below that threshold. If a researcher instead chooses a level of conservatism comparable to statistical significance testing, a small sample of 60 data points is still sufficient to be confident that a threshold is unlikely to represent a population Kappa more than 0.2 below that threshold.

Table 1. The proportion of cases where Population Kappa was more than specific distances (row) below Threshold Kappa (cols), for a sample size of 60. (Italic boldface cell is referred to in Methods section).

		Threshold Kappa (th)				
		0.6	**0.65**	**0.7**	**0.75**	**0.8**
Population Kappa	**th - 0.05**	0.316	0.316	0.274	0.300	0.242
	th - 0.1	0.175	0.175	***0.139***	0.151	0.107
	th - 0.2	0.035	0.035	0.024	0.025	0.014
	th - 0.3	0.005	0.035	0.003	0.002	0.001

* th: Threshold Kappa

Achieving a threshold of 0.7 and still having substantial probability of population Kappa of 0.6 (13.9%) may be too risky for many researchers. As in the substantially more conservative analysis in [1, 2], increasing the sample size decreases the risk of having an inflated estimate of Kappa. If we increase the sample size to 100 (Table 2) or 200 data points (Table 3), the proportion of cases where population Kappa is much lower than threshold Kappa decreases as well. While more than 5% of cases (Eagan et al's maximum for acceptability) can have population Kappa 0.05 lower than threshold Kappa, this proportion has dropped under 20% for a sample size of 200. In addition, fewer than 5% of cases have a population Kappa 0.1 lower than threshold for a sample of 200, regardless of the threshold Kappa's value (in the range studied). As such, by increasing to a still very feasible sample size of 200, we can be confident that a sample Kappa over 0.7 is unlikely to represent a population Kappa below 0.6. We again note that choosing a Kappa threshold of 0.8 leads to higher degrees of certainty than lower values of threshold Kappa.

Table 2. The proportion of cases where Population Kappa was more than specific distances (row) below Threshold Kappa (cols), for a sample size of 100

		Threshold Kappa (th)				
		0.6	**0.65**	**0.7**	**0.75**	**0.8**
Population Kappa	th - 0.05	0.269	0.268	0.235	0.244	0.197
	th - 0.1	0.117	0.110	0.090	0.088	0.061
	th - 0.2	0.012	0.009	0.006	0.005	0.003

	th - 0.3	<0.001	<0.001	<0.001	<0.001	<0.001

* th: Threshold Kappa

We further investigate the impact of different sample sizes in Tables 4-8, considering sample sizes ranging from 20 to 8000. We note that very small differences between threshold Kappa and population Kappa can be achieved for large samples. If we increase the sample size to 800 the likelihood of a .05 difference to the population drops below 5%, at 2000 samples, it drops below 1%.

Similar to the above results, the results are consistent between .6 and .75 thresholds but lower for a threshold of .8. Therefore, researchers in communities that choose thresholds of .8 may be able to confidently use smaller samples than researchers in other communities.

Table 3. The proportion of cases where Population Kappa was more than specific distances (row) below Threshold Kappa (cols), for a sample size of 200

		Threshold Kappa (th)				
		0.6	**0.65**	**0.7**	**0.75**	**0.8**
Population Kappa	**th - 0.05**	0.193	0.176	0.164	0.146	0.126
	th - 0.1	0.047	0.039	0.032	0.024	0.016
	th - 0.2	0.001	<0.001	<0.001	<0.001	<0.001
	th - 0.3	<0.001	<0.001	<0.001	<0.001	<0.001

* th: Threshold Kappa

Table 4. The proportion of cases where Population Kappa was more than a certain distance (cols) below a Threshold Kappa of 0.6, for varying sample sizes (rows)

		Population Kappa			
		0.55(th - 0.05)	**0.5 (th - 0.1)**	**0.4 (th - 0.2)**	**0.3 (th - 0.3)**
Sample Size	**20**	0.399	0.307	0.165	0.080
	40	0.350	0.231	0.077	0.020
	60	0.317	0.175	0.036	0.006
	80	0.292	0.145	0.021	0.002
	100	0.269	0.117	0.012	<0.001
	200	0.193	0.047	0.001	<0.001
	400	0.112	0.009	<0.001	<0.001
	500	0.086	0.004	<0.001	<0.001
	800	0.042	<0.001	<0.001	<0.001
	1000	0.028	<0.001	<0.001	<0.001
	2000	0.003	<0.001	<0.001	<0.001
	4000	<0.001	<0.001	<0.001	<0.001
	8000	<0.001	<0.001	<0.001	<0.001

Table 5. The proportion of cases where Population Kappa was more than a certain distance (cols) below a Threshold Kappa of 0.65, for varying sample sizes (rows)

		Population Kappa			
		0.6 (th - 0.05)	**0.55 (th - 0.1)**	**0.45 (th - 0.2)**	**0.35 (th - 0.3)**
	20	0.402	0.300	0.154	0.068
	40	0.307	0.192	0.059	0.013
	60	0.314	0.175	0.036	0.036
	80	0.266	0.122	0.016	0.001
	100	0.268	0.110	0.009	<0.001
	200	0.176	0.039	<0.001	<0.001
Sample Size	**400**	0.099	0.006	<0.001	<0.001
	500	0.079	0.003	<0.001	<0.001
	800	0.035	<0.001	<0.001	<0.001
	1000	0.022	<0.001	<0.001	<0.001
	2000	0.003	<0.001	<0.001	<0.001
	4000	<0.001	<0.001	<0.001	<0.001
	8000	<0.001	<0.001	<0.001	<0.001

Table 6. The proportion of cases where Population Kappa was more than a certain distance (cols) below a Threshold Kappa of 0.7, for varying sample sizes (rows)

		Population Kappa			
		0.65(th - 0.05)	**0.6 (th - 0.1)**	**0.5 (th - 0.2)**	**0.4 (th - 0.3)**
	20	0.314	0.224	0.103	0.041
	40	0.298	0.182	0.051	0.011
	60	0.275	0.139	0.024	0.003
	80	0.254	0.114	0.013	0.001
	100	0.235	0.090	0.006	<0.001
	200	0.164	0.032	<0.001	<0.001
Sample Size	**400**	0.084	0.004	<0.001	<0.001
	500	0.064	0.002	<0.001	<0.001
	800	0.027	<0.001	<0.001	<0.001
	1000	0.016	<0.001	<0.001	<0.001
	2000	0.001	<0.001	<0.001	<0.001
	4000	<0.001	<0.001	<0.001	<0.001
	8000	<0.001	<0.001	<0.001	<0.001

Table 7. The proportion of cases where Population Kappa was more than a certain distance (cols) below a Threshold Kappa of 0.75, for varying sample sizes (rows)

		Population Kappa			
		0.7 (th - 0.05)	**0.65 (th - 0.1)**	**0.55 (th - 0.2)**	**0.45 (th - 0.3)**
Sample Size	**20**	0.397	0.289	0.137	0.057
	40	0.284	0.162	0.041	0.008
	60	0.301	0.152	0.026	0.003
	80	0.236	0.096	0.009	<0.001
	100	0.244	0.088	0.005	<0.001
	200	0.146	0.024	<0.001	<0.001
	400	0.072	0.003	<0.001	<0.001
	500	0.054	0.001	<0.001	<0.001
	800	0.020	<0.001	<0.001	<0.001
	1000	0.012	<0.001	<0.001	<0.001
	2000	0.001	<0.001	<0.001	<0.001
	4000	<0.001	<0.001	<0.001	<0.001
	8000	<0.001	<0.001	<0.001	<0.001

Table 8. The proportion of cases where Population Kappa was more than a certain distance (cols) below a Threshold Kappa of 0.8, for varying sample sizes (rows)

		Population Kappa			
		0.75(th - 0.05)	**0.7 (th - 0.1)**	**0.6 (th - 0.2)**	**0.5 (th - 0.3)**
Sample Size	**20**	0.292	0.196	0.081	0.030
	40	0.265	0.144	0.035	0.006
	60	0.243	0.108	0.015	0.001
	80	0.218	0.082	0.007	<0.001
	100	0.197	0.061	0.003	<0.001
	200	0.126	0.016	<0.001	<0.001
	400	0.054	0.001	<0.001	<0.001
	500	0.040	<0.001	<0.001	<0.001
	800	0.014	<0.001	<0.001	<0.001
	1000	0.006	<0.001	<0.001	<0.001
	2000	<0.001	<0.001	<0.001	<0.001
	4000	<0.001	<0.001	<0.001	<0.001
	8000	<0.001	<0.001	<0.001	<0.001

6 Discussion and Conclusions

Overall, our Monte Carlo analyses show that the situation for Kappa is not quite so grim as Eagan et al. [1, 2] argue. Whereas they argued that Kappa could only be confidently used for samples of 400 [2] or 2000 [1] data points or higher, we find that only small

differences are seen for much smaller samples. As discussed above, our approach differs in three fashions. First, Eagan and colleagues allow simulated population Kappa to vary randomly, where we select specific known population Kappas and then test for the false positive rate for a specific known threshold Kappa. This methodological choice allows us to focus on specific possible cases. Second, Eagan and colleagues considered any case where sample Kappa was over threshold and population Kappa was under threshold to be problematic -- we look at the actual degree of difference. Third, Eagan and colleagues select a stringency level of 0.05, in line with statistical significance testing -- we argue for a stringency level of 0.2, in line with statistical power analysis. Note that readers of this paper do not need to agree with this third recommendation in order to make use of the analyses presented here. Even if Eagan et al.'s original stringency recommendations (0.05) are kept, looking at the magnitude of difference in Kappa values still leads to different conclusions than is seen in their work.

What our results indicate is that if we are willing to accept that a sample Kappa may be slightly higher than a population Kappa, fairly small sample sizes are needed to use Kappa with confidence. If we are willing for a threshold Kappa of 0.6 to actually represent a population Kappa of 0.501 5% of the time, then a sample of 200 is sufficient. If we are willing for a threshold Kappa of 0.8 to actually represent a population Kappa of 0.751 20% of the time, then a sample of 100 is sufficient. And if we are willing for a threshold Kappa of 0.7 to actually represent a population Kappa of .601 10% of the time, then a sample of 100 is sufficient. Even a very small sample of 40 can be acceptable in some cases -- for instance, if we are willing for a threshold Kappa of 0.7 to actually represent a population Kappa of .601 20% of the time.

Ultimately the difference between our approach and Eagan et al.'s approach is the question -- do we intend that a sample Kappa over threshold must indicate that population Kappa is also above threshold? Or do we think that a sample Kappa over threshold must indicate that population Kappa is probably near the threshold? It is not clear why a researcher would require population Kappa to be over a specific threshold, rather than close to it, given the disagreement between communities as to what the threshold should be.

That said, if a researcher wants to adopt this standard, our findings also show that there is a simpler approach than massively increasing sample size (or discarding Kappa entirely), as recommended by Eagan et al. A researcher can simply choose a higher threshold Kappa than their actual intended threshold Kappa. For example, a researcher who wants very high confidence that their Kappa will be strictly over 0.65 can instead choose a threshold of 0.75, and a sample size of 200. Then, as Table 7 shows, there would only be a 2.4% chance of population Kappa being below 0.65.

In other words, our findings provide evidence that Kappa can be acceptable for many uses and assumptions, even with smaller sample sizes than our community typically uses. We concur with Eagan et al. that Kappa becomes a more reliable metric with larger sample sizes, as do many other metrics (cf. [1]). However, a sample of 400-2000 is not necessary in order to use Kappa, under different assumptions. If we discard one of three assumptions -- that population Kappa must be strictly over threshold, that under 5% of cases should have population Kappa values below threshold, or that we cannot

simply use a higher threshold with our sample Kappa -- then the use of Kappa with smaller sample sizes is again acceptable.

There is ultimately a trade-off in the sample size labeled and the confidence which we can place in our findings. With a larger sample, there is less likelihood of noise in the data and our estimate of Kappa can be more precise. But the cost of this is more time spent in coding data for inter-rater checking. Some accounts have suggested that so much data must be coded to use Kappa that Kappa is essentially infeasible to use (i.e., [1]) -- our findings suggest that this tradeoff is perhaps not quite so grim as it may have looked.

To aid researchers in deciding what sample size to use, we have made the code for our experiments open-source (link redacted for review) so that researchers may experiment with different parameters and use our codebase as a decision-making tool.

References

1. Eagan, B.R., 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). https://doi.org/10.1145/3375462.3375508.
2. Eagan, B.R., Rogers, B., Serlin, R., Ruis, A.R., Irgens, G.A., Shaffer, D.W.: Can we rely on IRR? Testing the assumptions of inter-rater reliability. Presented at the International Conference on Computer-supported Collaborative Learning, Philadelphia, PA (2017).
3. Shaffer, D.W., Collier, W., Ruis, A.R.: A tutorial on epistemic network analysis: Analyzing the structure of connections in cognitive, social, and interaction data. Journal of Learning Analytics. 3, 9–45 (2016).
4. Kaliisa, R., Misiejuk, K., Irgens, G.A., Misfeldt, M.: Scoping the Emerging Field of Quantitative Ethnography: Opportunities, Challenges and Future Directions. In: International Conference on Quantitative Ethnography. pp. 3–17 (2021). https://doi.org/10.1007/978-3-030-67788-6_1.
5. Shaffer, D.W.: Quantitative Ethnography. Cathcart Press, Madison, WI (2017).
6. Cai, Z., Siebert-Evenstone, A., Eagan, B., Shaffer, D.W., Hu, X., Graesser, A.C.: nCoder+: A Semantic Tool for Improving Recall of nCoder Coding. In: International Conference on Quantitative Ethnography. pp. 41–54 (2019). https://doi.org/10.1007/978-3-030-33232-7_4.
7. Shaffer, D.W., Ruis, A.R.: How we code. In: International Conference on Quantitative Ethnography. pp. 62–77 (2021).
8. Cai, Z., Siebert-Evenstone, A., Eagan, B., Shaffer, D.W.: Using Topic Modeling for Code Discovery in Large Scale Text Data. In: International Conference on Quantitative Ethnography. pp. 18–31 (2021). https://doi.org/10.1007/978-3-030-67788-6_2.
9. Davey, J.W., Gugiu, P.C., Coryn, C.L.: Quantitative methods for estimating the reliability of qualitative data. Journal of MultiDisciplinary Evaluation. 6, 140–162 (2010).
10. Burla, L., Knierim, B., Barth, J., Liewald, K., Duetz, M., Abel, T.: From Text to Codings: Intercoder Reliability Assessment in Qualitative Content Analysis. Nursing Research. 57, 113–117 (2008). https://doi.org/10.1097/01.NNR.0000313482.33917.7d.
11. Cohen, J.: A Coefficient of Agreement for Nominal Scales. Educational and Psychological Measurement. 20, 37–46 (1960). https://doi.org/10.1177/001316446002000104.

12. Fleiss, J.L.: Measuring nominal scale agreement among many raters. Psychological Bulletin. 76, 378–382 (1971). https://doi.org/10.1037/h0031619.
13. Cohen, J.: Weighted Kappa: Nominal Scale Agreement with Provision for Scaled Disagreement or Partial Credit. Psychological Bulletin. 70, 213–220 (1968).
14. Delgado, R., Tibau, X.-A.: Why Cohen's Kappa should be avoided as performance measure in classification. PLoS ONE. 14, e0222916 (2019). https://doi.org/10.1371/journal.pone.0222916.
15. Feinstein, A.R., Cicchetti, D.V.: High agreement but low Kappa: I. the problems of two paradoxes. Journal of Clinical Epidemiology. 43, 543–549 (1990). https://doi.org/10.1016/0895-4356(90)90158-L.
16. Jeni, L.A., Cohn, J.F., De La Torre, F.: Facing Imbalanced Data--Recommendations for the Use of Performance Metrics. In: Humaine Association Conference on Affective Computing and Intelligent Interaction. pp. 245–251 (2013). https://doi.org/10.1109/ACII.2013.47.
17. Rigby, A.S.: Statistical methods in epidemiology. v. Towards an understanding of the kappa coefficient. Disability and Rehabilitation. 22, 339–344 (2000). https://doi.org/10.1080/096382800296575.
18. Cantor, A.B.: Sample-Size Calculations for Cohen's Kappa. Psychological Methods. 1, 150–153 (1996).
19. Matthews, B.W.: Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochimica et Biophysica Acta (BBA) - Protein Structure. 405, 442–451 (1975). https://doi.org/10.1016/0005-2795(75)90109-9.
20. Byrt, T., Bishop, J., Carlin, J.B.: Bias, prevalence and kappa. Journal of Clinical Epidemiology. 46, 423–429 (1993). https://doi.org/10.1016/0895-4356(93)90018-V.
21. Cohen, J.: Statistical Power Analysis. Current Directions in Psychological Science. 1, 98–101 (1992).
22. Wilcoxon, F.: Individual comparisons by ranking methods. Biometrics Bulletin. 1, 80–83 (1945).