Cross-replication Reliability - An Empirical Approach to Interpreting Inter-rater Reliability

Ka Wong
Google Research
danicky@gmail.com

Praveen Paritosh
Google Research
pkp@google.com

Lora Aroyo
Google Research
l.m.aroyo@gmail.com

Abstract

When collecting annotations and labeled data from humans, a standard practice is to use inter-rater reliability (IRR) as a measure of data goodness (Hallgren, 2012). Metrics such as Krippendorff's alpha or Cohen's kappa are typically required to be above a threshold of 0.6 (Landis and Koch, 1977). These absolute thresholds are unreasonable for *crowdsourced data* from annotators with high cultural and training variances, especially on *subjective topics*. We present a new alternative to interpreting IRR that is more empirical and contextualized. It is based upon benchmarking IRR against baseline measures in a replication, one of which is a novel *cross-replication reliability* (xRR) measure based on Cohen's (1960) kappa. We call this approach the xRR framework. We opensource a replication dataset of 4 million human judgements of facial expressions and analyze it with the proposed framework. We argue this framework can be used to measure the quality of crowdsourced datasets.

1 Introduction

Much content analysis and linguistics research is based on data generated by human beings (henceforth, annotators or raters) asked to make some kind of judgment. These judgments involve systematic interpretation of textual, visual, or audible matter (e.g. newspaper articles, television programs, advertisements, public speeches, and other multimodal data). When relying on human observers, researchers must worry about the quality of the data — specifically, their reliability (Krippendorff, 2004). Are the annotations collected reproducible, or are they the result of human idiosyncrasies?

Respectable scholarly journals typically require reporting quantitative evidence for the inter-rater reliability (IRR) of the data (Hallgren, 2012). Cohen's kappa (Cohen, 1960) or Krippendorff's alpha (Hayes and Krippendorff, 2007) is expected to be

Kappa Statistic	Strength of Agreement
<0.00	Poor
0.00–0.20	Slight
0.21–0.40	Fair
0.41–0.60	Moderate
0.61–0.80	Substantial
0.81–1.00	Almost Perfect

Figure 1: Agreement measures for categorical data (Landis and Koch, 1977)

above a certain threshold to be worthy of publication, typically 0.6 (Landis and Koch, 1977). Similar IRR requirements for human annotations data have been followed across many fields. In this paper we refer to this absolute interpretation of IRR as the Landis-Koch approach (Fig. 1).

This approach has been foundational in guiding the development of widely used and shared datasets and resources. Meanwhile, the landscape of human annotations collection has witnessed a tectonic shift in recent years. Driven by the data-hungry success of machine learning (LeCun et al., 2015; Schaekermann et al., 2020), there has been an explosive growth in the use of *crowdsourcing* for building datasets and benchmarks (Snow et al., 2008; Kochhar et al., 2010). We identify *three paradigm shifts* in the scope of and methodologies for data collection that make the Landis-Koch approach not as useful in today's settings.

A rise in annotator diversity In the pre-crowdsourcing era lab settings, data were typically annotated by two graduate students following detailed guidelines and working with balanced corpora. Over the past two decades, however, the bulk of data are annotated by crowd workers with high cultural and training variances.

A rise in task diversity There has been an increasing amount of subjective tasks with genuine ambiguity: judging toxicity of online discussions (Aroyo et al., 2019), in which the IRR values range

7053

between 0.2 and 0.4; judging emotions expressed by faces (Cowen and Keltner, 2017), in which more than 80% of the IRR values are below 0.6; and A/B testing of user satisfaction or preference evaluations (Kohavi and Longbotham, 2017), where IRR values are typically between 0.3 and 0.5.

A rise in imbalanced datasets Datasets are no longer balanced intentionally. Many high-stakes human judgements concern rare events with substantial tail risks: event security, disease diagnostics, financial fraud, etc. In all of these cases, a single rare event can be the source of considerable cost. High class imbalance has led to many complaints of IRR interpretability (Byrt et al., 1993; Feinstein and Cicchetti, 1990; Cicchetti and Feinstein, 1990).

Each of these changes individually has a profound impact on data reliability. Together, they have caused a shift from *data-from-the-lab* to *data-from-the-wild,* for which the Landis-Koch approach to interpreting IRR is admittedly too rigid and too stringent. Meanwhile, we have seen a drop in the reliance on reliability. Machine learning, crowdsourcing, and data research papers and tracks have abandoned the use and reporting of IRR for human labeled data, despite calls for it (Paritosh, 2012). The most cited recent datasets and benchmarks used by the community such as SQuAD (Rajpurkar et al., 2016), ImageNet (Deng et al., 2009), Freebase (Bollacker et al., 2008), have never reported IRR values. This would have been unthinkable twenty years ago. More importantly, this is happening against the backdrop of a reproducibility crisis in artificial intelligence (Hutson, 2018).

With the decline of the usage of IRR, we have seen a rise of ad hoc, misguided quality metrics that took its place, including 1) agreement-%, 2) accuracy relative to consensus, 3) accuracy relative to "ground truth." This is dangerous, as IRR is still our best bet for ensuring data reliability. How can we ensure its continued importance in this new era of data collection?

This paper is an attempt to address this problem by proposing an empirical alternative to interpreting IRR. Instead of relying on an absolute scale, we benchmark an experiment's IRR against two baseline measures, to be found in a replication. Replication here is defined as re-annotating the same set of items with a slight change in the experimental setup, e.g., annotator population, annotation guidelines, etc. By fixing the underlying corpus, we can

ensure the baseline measures are *sensitive* to the experiment on hand. The first baseline measure is the annotator reliability in the replication. The second measure is the annotator reliability *between* the replications. In Section 3, we present a novel way of measuring this. We call it cross-kappa (κ_x). It is an extension of Cohen's (1960) kappa and is designed to measure annotator agreement between two replications in a chance-corrected manner.

We present in Appendix A the International Replication (IRep) dataset,[1] a large-scale crowdsourced dataset of four million judgements of human facial expressions in videos. The dataset consists of three replications in Mexico City, Budapest, and Kuala Lumpur.[2] Our analysis in Section 4 shows this empirical approach enables meaningful interpretation of IRR. In Section 5, we argue xRR is a sensible way of measuring the goodness of crowdsourced datasets, where high reliability is unattainable. While we only illustrate comparing annotator populations in this paper, the methodology behind the xRR framework is general and can apply to similarly replicated datasets, e.g., via change of annotation guidelines.

2   Related Work

To position our research, we present a brief summary of the literature in two areas: metrics for measuring annotator agreement and their shortcomings (Section 2.1), comparing replications of an experiment (Section 2.2).

2.1   Annotator Agreement

Artstein and Poesio (2008) present a comprehensive survey of the literature on IRR metrics used in linguistics. Popping (1988) compare an astounding 43 measures for nominal data (mostly applicable to reliability of data generated by only two observers). Since then, Cohen's (1960) kappa and its variants (Carletta et al., 1997; Cohen, 1968) have become the de facto standard for measuring agreement in computational linguistics.

One of the strongest criticisms of kappa is its lack of interpretability when facing class imbalance. This problem is known as the *kappa paradox* (Feinstein and Cicchetti, 1990; Byrt et al., 1993;

---

[1]https://github.com/google-research-datasets/replication-dataset
[2]On this task, raters received average hourly wages of $12, $20, and $14 USD in Mexico City, Budapest, and Kuala Lumpur respectively. See Appendix A for annotation setup.

Warrens, 2010), or the 'base rates' problem (Uebersax, 1987). Bruckner and Yoder (2006) show class imbalance imposes practical limits on kappa and suggest one to interpret kappa *in relation* to the class imbalance of the underyling data. Others have proposed measures that are more robust against class imbalance (Gwet, 2008; Spitznagel and Helzer, 1985; Stewart and Rey, 1988). Pontius Jr and Millones (2011) even suggest abandoning the use of kappa altogether.

2.2 Agreement Between Replications

Replications are often being compared, but it is done at the level of per-item mean scores. Cowen and Keltner (2017) measure the correlation between the mean scores of two geographical rater pools. They use Spearman's (1904) correction for attenuation (discussed later in this paper) with split-half reliability. Snow et al. (2008) measure the Pearson correlations between the score of a single expert and the mean score of a group of non-experts, and vice versa. In this comparison the authors do not correct for correlation attenuation, hence the reported correlations may be strongly biased. Bias aside, correlation is not suitable for tasks with non-interval data or task with missing data. In this paper, we propose a general methodology for measuring rater agreement between replications with the same kind of generality, flexibility, and ease of use as IRR.

3 Cross-replication Reliability (xRR)

Data reliability can be assessed when a set of items are annotated multiple times. When this is done by a single rater, *intra-rater reliability* assesses a person's agreement with oneself. When this is done by two or more raters, *inter-rater reliability* (IRR) assesses the agreement between raters in an experiment. We propose to extend IRR to measure a similar notion of rater-rater agreement, but where the raters are taken from two different experiments. We call it *cross-replication reliability* (xRR). These replications can be a result of re-labeling the same items with a different rater pool, annotation template, or on a different platform, etc.

We begin with a general definition of Cohen's (1960) kappa. We extend it to *cross-kappa* (κ_x) to measure cross-replication reliability. We then use this foundation to define *normalized κ_x* to measure similarity between two replications.

3.1 Kappa and Its Generalizations

The class of IRR measures is quite diverse, covering many different experimental scenarios, e.g., different numbers of raters, rating scales, agreement definitions, assumptions about rater interchangeability, etc. Out of all such coefficients, Cohen's (1960) kappa has a distinct property that makes it most suitable for the task on hand. Unlike Scott's pi (Scott, 1955), Fleiss's kappa (Fleiss, 1971), Krippendorf's alpha (Krippendorff, 2004), and many others, Cohen's (1960) kappa allows for *two different marginal distributions*. This stems from Cohen's belief that two raters do not necessarily share the same marginal distribution, hence they should not be treated *interchangeably*. When we compare replications, e.g., two rater populations, we are deliberately changing some underlying conditions of the experiment, hence it is safer to assume the marginal distributions will not be the same. Within either replication, however, we rely on the rater interchangeability assumption. We think this more accurately reflects the current practice in crowd-sourcing, where each rater contributes a limited number of responses in an experiment, and hence raters are operationally interchangeable.

Cohen's (1960) kappa was invented to compare two raters classifying n items into a fixed number of categories. Since its publication, it has been generalized to accommodate multiple raters (Light, 1971; Berry and Mielke Jr, 1988), and to cover different types of annotation scales: ordinal (Cohen, 1968), interval (Berry and Mielke Jr, 1988; Janson and Olsson, 2001), multivariate (Berry and Mielke Jr, 1988), and any arbitrary distance function (Artstein and Poesio, 2008). In this paper we focus on Janson and Olsson's (2001) generalization, which the authors denote with the lowercase Greek letter iota (ι). It extends kappa to accommodate interval data with multiple raters, and is expressed in terms of pairwise disagreement:

$$\iota = 1 - \frac{d_o}{d_e}. \tag{1}$$

d_o in this formula represents the observed portion of disagreement and is defined as:

$$d_o = \left[n \binom{b}{2} \right]^{-1} \sum_{r<s} \sum_{i}^{n} D(x_{ri}, x_{si}), \tag{2}$$

where n is the number of items, b the number of annotators, i the item index, r and s the annotator

indexes; $\sum_{r<s}$ is the sum over all r and s such that $1 <= r < s <= b$. $D()$ is a pairwise disagreement defined as:

$$D(x_{ri}, x_{si}) = (x_{ri} - x_{si})^2 \qquad (3)$$

for interval data, and

$$D(x_{ri}, x_{si}) = \begin{cases} 0, & x_{ri} = x_{si} \\ 1, & \text{otherwise} \end{cases} \qquad (4)$$

for categorical data. Note we are dropping Janson and Olsson's multivariate reference in $D()$ and focusing on the univariate case. d_e in the denominator represents the expected portion of disagreement and is defined as:

$$d_e = \left[n^2 \binom{b}{2} \right]^{-1} \sum_{r<s} \sum_i^n \sum_j^n D(x_{ri}, x_{sj}). \qquad (5)$$

Janson and Olsson's expression in Eq. 1 is based on Berry and Mielke Jr (1988). While the latter use absolute distance for interval data, the former use squared distance instead. We follow Janson and Olsson's approach because squared distance leads to desirable properties and familiar interpretation of coefficients (Fleiss and Cohen, 1973; Krippendorff, 1970). Squared distance is also used in alpha (Krippendorff, 2004). Berry and Mielke Jr (1988) show if $b = 2$ and the scale is categorical, ι in Eq. 1 reduces to Cohen's (1960) kappa. For other rating scales such as ratio, rank, readers should refer to Krippendorff (2004) for additional distance functions. The equations for d_o and d_e are unaffected by the choice of $D()$.

3.2 Definition of κ_x

Here we present κ_x as a novel reliability coefficient for measuring the rater agreement between two replications. In Janson and Olsson's generalized kappa above, the disagreement is measured within pairs of annotations taken from the same experiment. In order to extend it to measure cross-replication agreement, we construct annotation pairs such that the two annotations are taken from different replications. We do not consider annotation pairs from the same replication. We define *cross-kappa*, $\kappa_x(X,Y)$, as a reliability coefficient between replications X and Y:

$$\kappa_x(X,Y) = 1 - \frac{d_o(X,Y)}{d_e(X,Y)}, \qquad (6)$$

where

$$d_o(X,Y) = \frac{1}{nRS} \sum_{i=1}^n \sum_{r=1}^R \sum_{s=1}^S D(x_{ri}, y_{si}), \qquad (7)$$

and

$$d_e(X,Y) = \frac{1}{n^2 RS} \sum_{i=1}^n \sum_{j=1}^n \sum_{r=1}^R \sum_{s=1}^S D(x_{ri}, y_{sj}), \qquad (8)$$

where x and y denote annotations from replications X and Y respectively, n is the number of items, R and S the numbers of annotations per item in replications X and Y respectively. In this definition, the observed disagreement is obtained by averaging disagreement observed in nRS pairs of annotations, where each pair contains two annotations on the same item taken from two different replications. Expected disagreement is obtained by averaging over all possible n^2RS cross-replication annotation pairs. When each replication has only 1 annotation per item, and the data is categorical, it is easy to show κ_x reduces to Cohen's (1960) kappa.

κ_x is a kappa-like measure, and will have properties similar to kappa's. κ_x is bounded between 0 and 1 in theory, though in practice it may be slightly negative for small sample sizes. $\kappa_x = 0$ means there is no discernible agreement between raters from two replications, beyond what would be expected by chance. $\kappa_x = 1$ means all raters between two replications are in perfect agreement with each other, which also implies perfect agreement within either replication.

3.3 κ_x with Missing Data

As presented, the two replications can have different numbers of annotations per item. However, within either replication, the number of annotations per item is assumed to be fixed. We recognize this may not always be the case. In practice, items within an experiment can receive varying numbers of annotations (i.e., missing data). We now show how to calculate κ_x with missing data.

When computing IRR with missing data, weights can be used to account for varying numbers of annotations within each item. Janson and Olsson (2004) propose a weighting scheme for iota in Eq. 1. Instead, we follow the tradition of Krippendorff (2004) in weighting each annotation equally in computing d_o and d_e. That amounts to the following scheme. In d_o, we first normalize within each item, then we take a weighted average over

7056

all items, with weights proportional to the combined numbers of annotations per item. In d_e, no weighting is required.

Since R and S can now vary from item to item, we index them using $R(*)$ and $S(*)$ to denote that they are functions of the underlying items. We rewrite d_o and d_e as:

$$d_o(X,Y) = \sum_{i=1}^{n} \frac{R(i) + S(i)}{\mathbf{R} + \mathbf{S}} \sum_{r=1}^{R(i)} \sum_{s=1}^{S(i)} \frac{D(x_{ri}, y_{si})}{R(i) \cdot S(i)} \tag{9}$$

and

$$d_e(X,Y) = \frac{1}{\mathbf{R} \cdot \mathbf{S}} \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{r=1}^{R(i)} \sum_{s=1}^{S(j)} D(x_{ri}, y_{sj}), \tag{10}$$

with

$$\mathbf{R} = \sum_{i}^{n} R(i), \quad \mathbf{S} = \sum_{j}^{n} S(j), \tag{11}$$

where $\mathbf{R}$ is the total number of annotations in replications X, $R(i)$ the number annotations on item i in replication X, $r = 1, 2, \ldots, R(i)$ (on item i in replication X); and similarly for $\mathbf{S}$, $S(j)$, and s with respect to replication Y. $\sum_{r=1}^{R(i)}$ and $\sum_{s=1}^{S(j)}$ in Eq. 9 and 10 are inner summations, where i and j are indexes from the outer summations. Without missing data, $R(i) = R$ for all i, and $S(j) = S$ for all j, then $\mathbf{R} = nR$, $\mathbf{S} = nS$, reducing Eq. 9 and 10 to Eq. 7 and 8.

3.4 Normalization of κ_x

xRR is modeled closely after IRR in order to serve as its baseline. As IRR measures the agreement between raters, so does xRR. In other words, κ_x is really a measure of *rater agreement*, not a measure of *experimental similarity* per se. This distinction is important. If we want to measure how well we replicate an experiment, we need to measure its disagreement with the replication *in relationship* to their own internal disagreements. The departure between inter-experiment and intra-experiment disagreements is important in measuring experimental similarity.

This calls for a normalization that considers κ_x in relation to IRR. First, we take inspirations from Spearman's correction for attenuation (Spearman, 1904):

$$\rho_{xy} = \frac{r_{xy}}{\sqrt{reliability_x}\sqrt{reliability_y}}, \tag{12}$$

where r_{xy} is the observed Pearson product-moment correlation between x and y (variables observed with measurement errors), ρ_{xy} is an estimate of their true, unobserved correlation (in the absence of measurement errors), and $reliability_x$ and $reliability_y$ are the reliabilities of x and y respectively. Eq. 12 is Spearman's attempt to correct for the negative bias in r_{xy} caused by the observation errors in x and y.[3]

Eq. 12 is relevant here because of the close connection between Cohen's (1960) kappa and the Pearson correlation, r_{xy}. In the dichotomous case, if the two marginal distributions are the same, Cohen's (1960) kappa is equivalent to the Pearson correlation (Cohen, 1960, 1968). In the multi-category case, Cohen (1968) generalizes this equivalence to weighted kappa, under the conditions of equal marginals and a specific quadratic weighting scheme.

Based on this strong connection, we propose replacing r_{xy} in Eq. 12 with κ_x and define normalized κ_x as:

$$\text{normalized } \kappa_x = \frac{\kappa_x(X,Y)}{\sqrt{\text{IRR}_X}\sqrt{\text{IRR}_Y}}. \tag{13}$$

Defined this way, one would expect normalized κ_x to behave like ρ_{xy}. That is indeed the case. When we apply both measures to the IRep dataset, we obtain a Pearson correlation of 0.99 between them (see Section 4.5). This leads to two insights. First, we can interpret normalized κ_x like a disattenuated correlation, ρ_{xy} (see (Muchinsky, 1996) for a rigorous interpretation). Second, normalized κ_x approximates the true correlation between two experiments' item-level mean scores.

Despite their affinity, ρ_{xy} is not a substitute for normalized κ_x for measuring experimental similarity. Normalized κ_x is more general as it can accommodate non-interval scales and missing data.

3.5 Connection between xRR and IRR

By connecting normalized κ_x to ρ_{xy}, we can also learn a lot about κ_x itself. To the extent that normalized κ_x approximates ρ_{xy}, we can rewrite Eq. 13 as:

$$\kappa_x(X,Y) \approx \rho_{xy}\sqrt{\text{IRR}_X}\sqrt{\text{IRR}_Y}. \tag{14}$$

This formulation shows κ_x behaves like a product of ρ_{xy} and the geometric mean of the two IRRs.

---

[3]Spearman relied on the assumption that the errors are uncorrelated with each other and with x and y.

This has important consequences, as we can deduce the following. 1) Holding constant the mean scores, and hence ρ_{xy}, the lower the IRRs, the lower the κ_x. Intra-experiment disagreement inflates inter-experiment disagreement. 2) In theory $\rho_{xy} \leq 1.0$,[4] hence κ_x is capped by the greater of the two IRRs. I.e., Intra-experiment agreement presents a ceiling to inter-experiment agreement. 3) If x and y are identically distributed, e.g., in a perfect replication, $\rho_{xy} = 1$ and $\kappa_x(X, Y) = \text{IRR}_X = \text{IRR}_Y$. Thus, when a low reliability experiment is replicated perfectly, κ_x will be as low, whereas normalized κ_x will be 1. This explains why normalized κ_x is more suitable for measuring experimental similarity.

In this section, we propose κ_x as a measure of rater agreement between two replications, and normalized κ_x is as an experimental similarity metric. In the next section, we apply them in conjunction with IRR to illustrate how we can gain deeper insights into experiment reliabilities by triangulating these measures.

4 Applying xRR to the IRep Dataset

As a standalone measure, IRR captures the reliability of an experiment by encapsulating many of its facets: class imbalance, item difficulty, guideline clarity, rater qualification, task ambiguity, etc. As such, it is difficult to compare the IRR of different experiments, or to interpret their individual values, because IRR is tangled with all the aforementioned design parameters. For example, we cannot attribute a low IRR to rater qualification without first isolating other design parameters. This is the problem we try to solve with xRR by contextualizing IRR with meaningful baselines via a replication. We will demonstrate this by applying this technique to the IRep Dataset (Appendix A). We focus on a subset of 5 emotions for illustration purposes, with the rest of the reliability values provided in Appendix B. In our analysis, IRR is measured with Cohen's (1960) kappa and xRR with κ_x. We will refer to them interchangeably.

4.1 IRR Variability Across Emotions

First we illustrate in Fig. 2 that different emotions within the same city can have very different IRR. For instance, the labels *awe* and *love* in Mexico City have an IRR of 0.1208 and 0.597 respectively (Table 1). *Awe* and *love* are completely different

---

[4]Spearman's correction can occasionally produce a correlation above 1.0 (Muchinsky, 1996).

emotions with different levels of class imbalance and ambiguity, and without controlling for these differences, the gap in their reliabilities is not unexpected. That is exactly the problem about comparing IRRs – such comparisons are not meaningful. We need something directly comparable to *awe* in order to interpret its low IRR. If we do not compare emotions, and just consider *awe* using the Landis-Koch scale, that would not be helpful either. We would not be able to tell if its low IRR is a result of poor guidelines, general ambiguity in emotion detection, or ambiguity specific to *awe*. It's more meaningful to compare replications of *awe* itself.
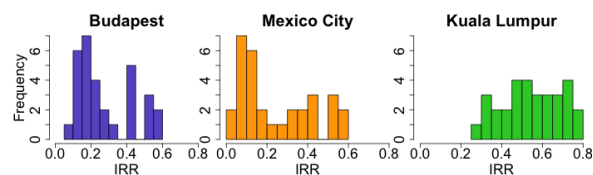

Figure 2: Histograms of 31 emotion labels' IRR in 3 cities. The x-axis denotes buckets of IRR values. The y-axis denotes the number of emotion labels in each of those buckets. There is a lot of variation between emotion labels within each city.

IRR	Mexico City	Kuala Lumpur	Budapest
contemplation	0.0514	0.745	0.1201
love	0.5973	0.6692	0.5421
awe	0.1208	0.3395	0.117
sadness	0.5147	0.6578	0.5175
contentment	0.4494	0.6363	0.2282

Table 1: IRR values of 5 emotion labels in 3 cities.

4.2 IRR Variability Across Replications

While the aforementioned variation in IRR between emotions is expected, IRR of the *same emotion* can vary greatly between replications as well. Fig. 3 shows two contrasting examples. On the one hand, the IRR of *love* is consistent across replications. On the other hand, the IRR of *contemplation* varies a lot. We know the IRR variation in *contemplation* is strictly attributed to rater pool differences because the samples, platforms and annotation templates are the same across experiments. Such variation in IRR will be missed entirely by sampling based approaches for error-bars (e.g. standard error, bootstrap), which assume a fixed rater population.

4.3 Cross-replication Rater Agreement

As shown, replication can facilitate comparisons of IRR by producing meaningful baselines. However, IRR is an internal property of a dataset, it does not allow us to compare two datasets directly. To
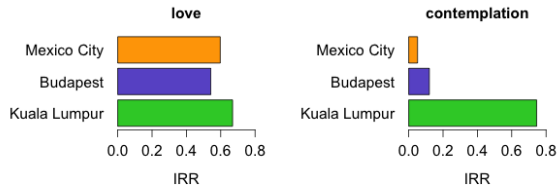
Figure 3: IRR values for label *love* (left) and *contemplation* (right) across the 3 cities. There are different degrees of IRR variability in the two emotion labels.

that end, we can apply κ_x to quantify the rater agreement between two datasets, as IRR quantifies the rater agreement within a dataset. Interestingly, not only is κ_x useful for comparing two datasets, but it also serves as another baseline for interpreting their IRRs.

IRR is a step toward ensuring reproducibility, so naturally we wonder how much of the observed IRR is tied to the specific experiment and how much of it generalizes? This is of particular concern when raters are sampled in a clustered manner, e.g., crowd workers from the same geographical region, grad students sharing the same office. We rarely make sure raters are diverse and representative of the larger population. High IRR can be the result of a homogeneous rater group, limiting the generality of the results. In the context of the IRep dataset, that two cities having similar IRRs does not imply their raters agree with each other at a comparable level, or at all. We will demonstrate this with two contrasting examples.
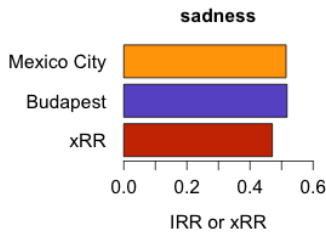
Mexico City and Budapest both have a moderate IRR for *sadness*, 0.5147 and 0.5175 respectively, and their κ_x is nearly the same at 0.4709 (Fig. 4). This gives us confidence that the high IRR of *sadness* generalizes beyond the specific rater pools. In contrast, on *contentment* Mexico City and Kuala Lumpur have comparable levels of IRR, 0.4494 and 0.6363 respectively, but their κ_x is an abysmal -0.0344 [5] (Fig. 5). In other words, the rater agreement on *contentment* is limited to within-pool observations only. This serves as an important reminder that IRR is a property of a specific experimental setup and may or may not generalize beyond that. κ_x allows us to ensure the internal agreement has external validity.

4.4 Replication Similarity

κ_x is a step towards comparing two replications, but it is not a good standalone measure of replication similarity. To do that, we must also account for both replications' internal agreements, e.g., via normalized κ_x in Eq. 13. Fig. 6 shows an example. Mexico City and Budapest have a low κ_x of 0.0817 on *awe*. On the surface, this low agreement may seem attributable to differences between the rater pools. However, there is a similarly low IRR in either city: 0.1208 in Mexico City, and 0.117 in Budapest. After accounting for IRR, normalized κ_x is much higher at 0.6872 (Table 2), indicating a decent replication similarity between the two cities.

Figure 4: IRR values of *sadness* in Mexico City and Budapest and their κ_x value. Both cities have as much internal agreement as cross-replication agreement.

Figure 6: IRR of *awe* in Mexico City and Budapest and their xRR. The low xRR is primarily a reflection of their low IRRs.
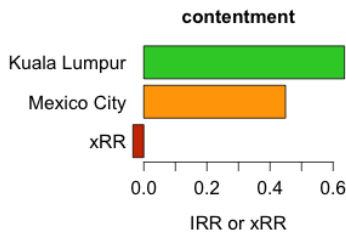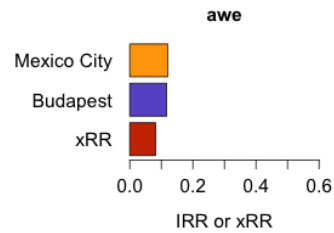

Figure 5: IRR of *contentment* in Kuala Lumpur and Mexico City and their κ_x. Both cities have high internal agreement, but no discernible cross-replication agreement.

	Mexico City x Budapest	Kuala Lumpur x Budapest	Mexico City x Kuala Lumpur
contemplation	0.070 (0.899)	0.040 (0.135)	0.036 (0.185)
love	0.519 (0.912)	0.480 (0.798)	0.604 (0.955)
awe	0.081 (0.687)	0.056 (0.283)	0.043 (0.213)
sadness	0.470 (0.912)	0.498 (0.854)	0.491 (0.845)
contentment	0.271 (0.848)	-0.014 (-0.03)	-0.034 (-0.064)

Table 2: κ_x and normalized κ_x (in parentheses) of 5 emotion labels in 3 replication pairs.

---

[5]Negative xRR value due to estimation error.

4.5 Connection to ρ_{xy}

We apply Spearman's correction for attenuation in Eq. 12 to all 31 emotion labels in 3 replication pairs. The resulting ρ_{xy} is plotted against the corresponding normalized κ_x in Fig. 7. Both measures are strongly correlated with a Pearson correlation of 0.99. This justifies interpreting normalized κ_x as a disattenuated correlation like ρ_{xy}.
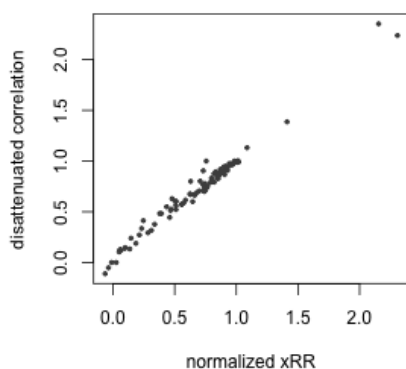

Figure 7: Scatter plot of ρ_{xy} (y-axis) and normalized κ_x (x-axis). Each dot is an emotion label in a pair of replications.

5 Measuring the Quality of a Crowdsourced Dataset

The IRep dataset is replicated and is conducive to xRR analysis. However, in practice most datasets are not replicated. Is xRR still useful? We present a specific use case of xRR in this section and argue that it is worth replicating a crowdsourced dataset in order to evaluate its quality.

5.1 Data Target

Given a set of items, it is possible that annotations of the highest attainable quality still fail to meet the Landis-Koch requirements. Task subjectivity and class imbalance together impose a practical limit on kappa (Bruckner and Yoder, 2006). In these situations, the experimenter can forgo a data collection effort for reliability reasons. Alternatively, the experimenter may believe that data of sufficiently high quality can still have scientific merits, regardless of reliability. If so, what guidance can we use to *ensure the highest quality data*, especially when collecting data via crowdsourcing? This paper is heavily motivated by this question.

xRR allows us to interpret IRR not on an absolute scale, but against a replication, a reference of sorts. By judging a crowdsourced dataset against a reference, we can decide if its meets a certain quality bar, albeit a relative one. In the IRep dataset,

all replications are of equal importance. However, in practice, we can often define a *trusted source* as our target. This trusted source can consist of linguists, medical experts, calibrated crowd workers, or the experimenters themselves. They should have enough expertise knowledge and an adequate understanding of the task. The critical criterion in choosing a target is its ability to remove common quality concerns such as rater qualification and guideline effectiveness.

5.2 Internal Agreements

By replicating a random subset of a crowdsourced dataset with trusted annotators,[6] one can compare the two IRRs and make sure they are at a similar level. If the crowd IRR is much higher, that may be an indication of collusion, or a set of overly simplistic guidelines that have deviated from the experiment fidelity (Sameki et al., 2015). If the crowd IRR is much lower, it may just be a reflection of annotator diversity, or it can mean under-defined guidelines, unequal annotator qualifications, etc. Further investigation is needed to ensure the discrepancy is reasonable and appropriate.

5.3 Mutual Agreement

Suppose the two IRRs are similar, that is not to say that both datasets are similar. Both groups of annotators can have high internal agreement amongst themselves, but the two groups can agree on different sets of items. If our goal is to collect crowdsourced data that closely mirror the target, then we have to measure their mutual agreement, in addition to comparing their internal agreements. Recall from Section 3.5 that if an experiment is replicated perfectly, κ_x should be identical to the two IRRs. Or more concisely, normalized κ_x should be equal to 1. Thus a high normalized κ_x can assure us that the crowdsourced annotators are functioning as an extension of the trusted annotators, based on which we form our expectations.

5.4 Relation to Gold Data

At a glance, this approach seems similar to the common practice of measuring the accuracy of crowdsourced data against the ground truth (Resnik et al., 2006; Hripcsak and Wilcox, 2002). However, they are actually fundamentally different approaches. κ_x is rooted in the reliability literature that does not rely on the existence of a *correct*

---

[6]2 or more ratings per item are needed to measure the IRR.

answer. The authors argue this is an unrealistic assumption for many crowdsourcing tasks, where the input involves some subjective judgement. Accuracy itself is also a flawed metric for annotations data due to its inability to handle data uncertainty. For instance, when the reliability of the gold data is less than perfect, accuracy can never reach 1.0. Furthermore, accuracy is not chance-corrected, so it tends to inflate with class imbalance.

5.5 Extending an Existing Dataset

The aforementioned technique can also measure the quality of a dataset extension. The main challenge in extending an existing dataset is to ensure the new data is consistent with the old. The state-of-the-art method in computer vision is *frequency matching* – ensuring the same proportion of yes/no votes in each image class. Recht et al. (2019) extended ImageNet[7] using this technique, concluding there is a 11% - 14% drop in accuracy across a broad range of models. While frequency matching controls the distribution of some statistics, the impact of the new platform is uncontrolled for. Engstrom et al. (2020) pointed out a bias in this sampling technique. Overall, it is difficult to assess how well we are extending a dataset. To that end, xRR can be of help. A high normalized κ_x and a comparable IRR in the new data can give us confidence in the uniformity and continuity in the data collection.

6 Discussion

There has been a tectonic shift in the scope of and methodologies for annotations data collection due to the rise of crowdsourcing and machine learning. In many of these tasks, a high reliability is often difficult to attain, even under favorable circumstances. The rigid Landis-Koch scale has resulted in a decrease in the usage and reporting of IRR in most widely used datasets and benchmarks. Instead of abandoning IRR, we should adapt it to new ways of measuring data quality. The xRR framework presents a first-principled way of doing so. It is a more empirical approach that utilizes a replication as a reference point. It is based on two metrics – κ_x for measuring cross-replication rater agreement – and normalized κ_x for measuring replication similarity.

We opensource a large-scale replication dataset of facial expression judgements analyzed with the proposed framework. We show this framework can

---

[7] http://www.image-net.org/

be used to guide our crowdsourcing data collection efforts. This is the beginning of a long line of inquiry. We outline future work and limitations below:

Confidence intervals for κ_x Confidence intervals for κ_x and normalized κ_x are required for hypothesis testing. Though one can use the block-bootstrap for an empirical estimate, large sample behavior of these metrics needs to be studied.

Sensitivity of κ_x with high class-imbalance The xRR framework sidesteps the effect of class-imbalance by comparing replications on the same item set. Further analysis needs to confirm the sensitivity of κ_x metrics in high class-imbalance.

Optimization of κ_x computation Our method requires constructing many pairs of observations: n^2RS. This may get prohibitively expensive, when the number of items is large. Using algebraic simplification and dynamic programming, this can be made much more efficient.

Alternative normalizations of κ_x We provided one particular normalization technique, but it may not suit all applications. For example, when comparing crowd annotations to expert annotations, one can consider, $\kappa_x/\mathrm{IRR}_{expert}$.

Alternative xRR coefficients Our proposed xRR coefficient, κ_x, is based on Cohen's (1960) kappa for its assumption about rater non-interchangeability. It may be useful to consider Krippendorff's alpha and other agreement statistics as alternatives for other assumptions and statistical characteristics.

We hope this paper and dataset will spark research on these questions and increase reporting of reliability measures for human annotated data.

Acknowledgments

References

Lora Aroyo, Lucas Dixon, Nithum Thain, Olivia Redfield, and Rachel Rosen. 2019. Crowdsourcing subjective tasks: The case study of understanding toxicity in online discussions. In *Companion Proceedings of The 2019 World Wide Web Conference*, WWW '19, page 1100–1105, New York, NY, USA. Association for Computing Machinery.

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Kenneth J. Berry and Paul W. Mielke Jr. 1988. A generalization of cohen's kappa agreement measure to interval measurement and multiple raters. *Educational and Psychological Measurement*, 48(4):921–933.

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: A collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, SIGMOD '08, page 1247–1250, New York, NY, USA. Association for Computing Machinery.

Cornelia Taylor Bruckner and Paul Yoder. 2006. Interpreting kappa in observational research: Baserate matters. *American journal on mental retardation*, 111(6):433–441.

Ted Byrt, Bishop Janet, and John Carlin. 1993. Bias, prevalence and kappa. *J Clin Epidemiol*, 46(5):423–9.

Jean Carletta, Amy Isard, Stephen Isard, Jacqueline C Kowtko, Gwyneth Doherty-Sneddon, and Anne H. Anderson. 1997. The reliability of a dialogue structure coding scheme.

Domenic V Cicchetti and Alvan R Feinstein. 1990. High agreement but low kappa: Ii. resolving the paradoxes. *Journal of clinical epidemiology*, 43(6):551–558.

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

Jacob Cohen. 1968. Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213.

Alan Cowen and Dacher Keltner. 2017. Self-report captures 27 distinct categories of emotion bridged by continuous gradients. *Proceedings of the National Academy of Sciences of the United States of America*, 114.

J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255.

Logan Engstrom, Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Jacob Steinhardt, and Aleksander Madry. 2020. Identifying statistical bias in dataset replication. In *International Conference on Machine Learning*, pages 2922–2932. PMLR.

Alvan R Feinstein and Domenic V Cicchetti. 1990. High agreement but low kappa: I. the problems of two paradoxes. *Journal of clinical epidemiology*, 43(6):543–549.

J.L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

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.

Kilem Li Gwet. 2008. Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, 61(1):29–48.

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.

Andrew F Hayes and Klaus Krippendorff. 2007. Answering the call for a standard reliability measure for coding data. *Communication methods and measures*, 1(1):77–89.

George Hripcsak and Adam Wilcox. 2002. Reference standards, judges, and comparison subjects: roles for experts in evaluating system performance. *Journal of the American Medical Informatics Association*, 9(1):1–15.

Matthew Hutson. 2018. Artificial intelligence faces reproducibility crisis.

Harald Janson and Ulf Olsson. 2001. A measure of agreement for interval or nominal multivariate observations. *Educational and Psychological Measurement*, 61(2):277–289.

Harald Janson and Ulf Olsson. 2004. A measure of agreement for interval or nominal multivariate observations by different sets of judges. *Educational and Psychological Measurement*, 64(1):62–70.

Shailesh Kochhar, Stefano Mazzocchi, and Praveen Paritosh. 2010. The anatomy of a large-scale human computation engine. In *Proceedings of the ACM SIGKDD workshop on Human Computation*, pages 10–17.

Ron Kohavi and Roger Longbotham. 2017. Online controlled experiments and a/b testing. *Encyclopedia of machine learning and data mining*, 7(8):922–929.

Klaus Krippendorff. 1970. Bivariate agreement coefficients for reliability of data. *Sociological methodology*, 2:139–150.

Klaus Krippendorff. 2004. *Content analysis: An introduction to its methodology*. Sage publications.

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

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *nature*, 521(7553):436–444.

Richard J. Light. 1971. Measures of response agreement for qualitative data: Some generalizations and alternatives. *Psychological Bulletin*, 76(5):365–377.

Paul M Muchinsky. 1996. The correction for attenuation. *Educational and psychological measurement*, 56(1):63–75.

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

Robert Gilmore Pontius Jr and Marco Millones. 2011. Death to kappa: birth of quantity disagreement and allocation disagreement for accuracy assessment. *International Journal of Remote Sensing*, 32(15):4407–4429.

Roel Popping. 1988. On agreement indices for nominal data. In *Sociometric research*, pages 90–105. Springer.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. 2019. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning*, pages 5389–5400. PMLR.

Philip Resnik, Michael Niv, Michael Nossal, Gregory Schnitzer, Jean Stoner, Andrew Kapit, and Richard Toren. 2006. Using intrinsic and extrinsic metrics to evaluate accuracy and facilitation in computer-assisted coding. In *Perspectives in Health Information Management Computer Assisted Coding Conference Proceedings*, pages 2006–2006.

Mehrnoosh Sameki, Aditya Barua, and Praveen Paritosh. 2015. Rigorously collecting commonsense judgments for complex question-answer content. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 3.

Mike Schaekermann, Christopher Homan, Lora Aroyo, Praveen Paritosh, Kurt Bollacker, and Chris Welty. 2020. The ai bookie bet: Will machine learning outgrow human labeling? *AI Magazine*, 41(4):123–126.

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

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.

C. Spearman. 1904. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1).

Edward L. Spitznagel and John E. Helzer. 1985. A proposed solution to the base rate problem in the kappa statistic. *Archives of General Psychiatry*, 42(7):725–728.

Gavin W. Stewart and Joseph M. Rey. 1988. A partial solution to the base rate problem of the k statistic. *Archives of general psychiatry*, 45(5):504–505.

John S Uebersax. 1987. Diversity of decision-making models and the measurement of interrater agreement. *Psychological bulletin*, 101(1):140.

Matthijs J Warrens. 2010. A formal proof of a paradox associated with cohen's kappa. *Journal of Classification*, 27(3):322–332.

Appendices

A   The IRep Dataset: Facial Expressions Replication Dataset

The IRep Dataset is a human annotated dataset with a list of 30 emotion labels from a set of emotion classes defined in Cowen and Keltner (2017) plus one additional label 'unsure'. During the data collection process, raters used the set of 30 facial expression labels to annotate their perception of the facial expression present in a video. As the purpose of this dataset is to illustrate how replications of human labeling experiments can be used to determine the overall quality of the resulting annotations, we have omitted the reference to the actual video content. The annotated videos are thus referred to as 'items' with a set of indices (item_IDs), e.g. item_1, item_2, etc, stored in a CSV format. Raters are referred to as Rater_1 or Rater_2 across all rater pools. They are just placeholders and do not imply particular individuals (Table 3).

The dataset contains 3,939,418 annotations on 38,499 unique items. The size of the dataset is 15MB, and the dataset is released on GitHub: `https://github.com/google-research-datasets/replication-dataset`. To produce the replications for this labeling experiment, we used rater pools in three different cities - Budapest, Kuala Lumpur and Mexico City - on the same labeling platform. In Table 4 we show the distribution of items and ratings across the different rater pools.

B IRR, xRR, and normalized xRR values for the IRep dataset

In Table 5 we report the IRR, κ_x, and normalized κ_x obtained from the entire IRep dataset.

Item_ID	Annotator_pool	Rater	Label_1	...	Label_30	Unsure
STRING	STRING	STRING	BOOLEAN	BOOLEAN	BOOLEAN	BOOLEAN
Item_1, ... Item_N	Budapest, Kuala Lumpur, Mexico city	Rater 1, Rater 2	0, 1	0, 1	0, 1	0, 1

Table 3: Schema of the CSV file: Each line in the IRep csv file corresponds to one item_ID annotated by Rater_1 and Rater_2 with some of the emotion labels (Label_1 … Label_30) annotated on the corresponding. There is one additional column for "unsure" indicating when it was not possible to determine which expression was expressed.

Pool	# of items	# of ratings	# of unique items	# of ratings per item
Mexico City	45,751	1,418,281	22,955	
			22,796	2
			159	1
Kuala Lumpur	26,811	831,141	13,422	
			13,389	2
			33	1
Budapest	54,516	1,689,996	27,666	
			26,850	2
			816	1
TOTALS	127,078	3,939,418	38,499	

Table 4: Distribution of items and ratings across different rater pools, where every item is annotated by a maximum of 2 raters from each pool. Here we show what fraction of the unique items were annotated by one or two raters in each rating pool.

	MC	KL	Bud	MC x Bud	KL x Bud	MC x KL	MC x Bud normalized	KL x Bud normalized	MC x KL normallized
shame	0.0192	0.4382	0.1261	0.106	0.0793	0.2116	2.1543	0.3373	2.3069
amusement	0.1238	0.6287	0.4401	0.2537	0.3919	0.1756	1.0869	0.745	0.6294
disgust	0.1725	0.5921	0.2475	0.2106	0.2389	0.2331	1.0192	0.6241	0.7294
pain	0.5156	0.5282	0.4383	0.4819	0.3581	0.4485	1.0137	0.7443	0.8594
desire	0.556	0.7548	0.4261	0.4886	0.4573	0.5643	1.0038	0.8064	0.8711
triumph	0.3215	0.3492	0.2778	0.2921	0.2592	0.3303	0.9774	0.8322	0.9858
anger	0.425	0.7413	0.526	0.4577	0.5308	0.5152	0.968	0.85	0.9179
surprise	0.3967	0.714	0.4007	0.376	0.453	0.4946	0.9431	0.8469	0.9293
distress	0.0799	0.3944	0.1587	0.1036	0.143	0.1183	0.92	0.5716	0.6664
love	0.5973	0.6692	0.5421	0.5193	0.4807	0.6042	0.9126	0.7981	0.9557
sadness	0.5147	0.6578	0.5175	0.4709	0.4987	0.4919	0.9124	0.8547	0.8454
fear	0.4054	0.6475	0.4363	0.3796	0.4174	0.4756	0.9026	0.7853	0.9283
contemplation	0.0514	0.745	0.1201	0.0707	0.0406	0.0363	0.8998	0.1357	0.1855
confusion	0.1708	0.4791	0.2314	0.1769	0.2207	0.2146	0.8898	0.6628	0.7502
unsure	0.5072	0.7802	0.5922	0.4743	0.167	0.247	0.8654	0.2457	0.3926
concentration	0.2973	0.5953	0.326	0.2675	0.2108	0.2972	0.8592	0.4785	0.7065
boredom	0.3624	0.7092	0.5701	0.3883	0.5525	0.4152	0.8543	0.8689	0.819
interest	0.1247	0.554	0.1917	0.1319	0.154	0.1342	0.8531	0.4726	0.5106
contentment	0.4494	0.6363	0.2282	0.2717	-0.0141	-0.0344	0.8484	-0.037	-0.0643
disappointment	0.0904	0.459	0.1823	0.0983	0.1477	0.1673	0.7657	0.5106	0.8213
pride	0.0451	0.5392	0.2016	0.0721	0.1539	0.2201	0.7561	0.4668	1.4114
elation	0.3052	0.6756	0.196	0.1813	0.2544	0.3974	0.7413	0.6991	0.8752
sympathy	0.086	0.5095	0.1148	0.0727	0.105	0.155	0.7317	0.4342	0.7405
awe	0.1208	0.3395	0.117	0.0817	0.0564	0.0433	0.6872	0.283	0.2138
ecstasy	0.1429	0.5449	0.1523	0.0953	0.0276	0.0138	0.646	0.0958	0.0495
neutral	0.213	0.4467	0.2615	0.148	0.0337	0.0175	0.6271	0.0986	0.0567
contempt	0.1234	0.3863	0.1281	0.0739	0.1022	0.1592	0.5878	0.4594	0.7292
relief	0.0948	0.4915	0.0982	0.0538	0.0684	0.1104	0.5576	0.3113	0.5114
doubt	0.1273	0.3235	0.1937	0.0874	0.0365	0.0768	0.5566	0.1458	0.3785
embarrassment	0.0841	0.4628	0.1508	0.0567	0.0613	0.1785	0.5035	0.232	0.9048
realization	0.0548	0.2656	0.1005	0.0019	0.0105	-0.0014	0.0256	0.0643	-0.0116

Table 5: The fist column shows the 30 emotion labels + "unsure" in the IRep dataset. The next 3 columns are their IRR measured by Cohen's (1960) kappa in Mexico City (MC), Kuala Lumpur (KL), and Budapest (Bud). The next 3 columns are the κ_x in the 3 pairs of cities, and the last 3 columns are the corresponding normalized κ_x.