The Unreliability of Measures of Intercoder Reliability, and What to do About it

Justin Grimmer* Gary King† Chiara Superti ‡

October 13, 2015

Abstract

In both automated and traditional text analysis, human coders are regularly tasked with categorizing documents. Researchers then evaluate the success of this crucial step in the research process via one of many measures of intercoder reliability, such as Cronbachs alpha. They then improve coding practices until this measure reaches some arbitrary threshold, at which point remaining disagreements are resolved in arbitrary ways and ignored in subsequent analyses. We show that this common practice can generate severely biased estimates and misleading conclusions. The problem is the focus on measures of intercoder reliability which, except at the extreme, are unrelated to the quantities of interest, such as the proportion of documents in each category. We thus develop an approach that enables scholars to directly incorporate coding uncertainty into statistical estimation. The method offers an interval estimate which we prove contains the true proportion of documents in each category, under reasonable assumptions. We then extend this method to situations with multiple coders, when one coder is trusted more than another, and when the resulting document codes are used as inputs to another statistical model. We offer easy-to-use software that implements all our suggestions.

*Department of Political Science, Stanford University

†Institute for Quantitative Social Science, Harvard University, 1737 Cambridge Street, Cambridge MA 02138; GaryKing.org, king@harvard.edu, (617) 500-7570.

‡Institute for Quantitative Social Science, Harvard University

1 Introduction

We attempt to shore up an essential and well known, but relatively undiscussed, component of the science of automated and traditional text analysis. Researchers in these areas spend considerable time and resources managing teams that code documents into a chosen set of categories. Categorization is a fundamental component of human understanding, but the ambiguous and ever creative nature of human language makes the process difficult and perfection impossible (**?**, **?**, **?**, Appendix, **?**, p.10). The consequences for the ultimate conclusions drawn depend crucially on the extent and nature of these difficulties in this formative stage of analysis.

The usual practice involves defining a set of mutually exclusive and exhaustive categories based on substantive interest and a theoretical perspective; assigning two or more research assistants to code a small set of documents into these categories; discovering that levels of intercoder reliability are too low; adjusting, redefining, and clarifying the categories; and then retraining the research assistants and coding a new set of documents from scratch. This process is repeated until the level of intercoder reliability is deemed satisfactory. Although the level of intercoder reliability reached often goes unreported in many fields Lombard, Snyder-Duch and Bracken (2002); Riffe and Freitag (1997), the threshold "satisfactory" level in most fields is approximately 70–80% (**?**, pp. 4–5). At that point, the disagreements are ignored. The remaining documents are coded by a single coder, or by multiple coders who negotiate their differences informally or their codes are averaged. This practice is common across fields, including political science (Stewart and Zhukov, 2009; De Vreese et al., 2006; Druckman and Parkin, 2005; Jamal et al., 2014; Druckman, Kifer and Parkin, 2010, 2009), medicine (Hripcsak and Heitjan, 2002), education (Béchard and Grégoire, 2005), journalism (Leccese, 2009), sociology (Williams et al., 2009), business and marketing (Nazli Nik Ahmad and Sulaiman, 2004; Bortree and Seltzer, 2009), psychology (Zullow et al., 1988), and communication (Van Gorp, 2005).

For traditional hand coding projects, all documents are coded in this way. For many types of automated text analysis, only documents in the training set are coded via these procedures. For other types of machine learning analyses, the training set has high accuracy (due to known information, such as the author of a speech), but the classifier used to sort the test set into categories produces analogous errors that are typically as large as hand coding exercises. For any of these, the consequences for ignoring the remaining errors be devastating for quantities of interest of special interest to social scientists. In particular, social scientists usually have little interest in the numeric code for any one social media post or congressional speech (which they could of course merely read). Instead, the quantity of interest in the social sciences usually involves the percent of documents in each category, such as treaties involving strategic misperception or social media posts supporting each presidential candidate.

Our proposal to ameliorate this problem does not involve changing the best practices of coders in improving their levels of intercoder reliability. Researchers should continue the practice of improving their categories and training their coders, as has been the case (e.g., **?**). We also have no objection to the long history of work on measures of intercoder reliability, such as Cronbach's α (Cronbach, 1951), Scott's

π (Scott, 1955), Cohen's κ (Byrt, Bishop and Carlin, 1993), Krippendorff's α (Hayes and Krippendorff, 2007), and many others. This literature continues to provide more nuanced, improving ways of summarizing a confusion matrix in a scalar metric for intercoder reliability. This work should also continue.

The problem then is neither the researcher's coding practices nor the methodologist's measures of intercoder reliability. The problem is the near universal practice of using these measures to determine a threshold level, above which intercoder reliability may be safely ignored. In fact, as we show, no such level exists. Other than near perfect agreement, which does not occur with meaningful categories and documents, no level of intercoder reliability should be ignored. What is missing then, is a set of tools that enable researchers to incorporate the remaining disagreement in their analytical methods. We provide those here. Although generic methods of measurement error have been proposed (**?**, Appendix; **?**, Molinari 2008), the methods introduced here are tuned for exactly this problem and so are potentially more powerful or do not require additional assumptions that are unlikely to hold.

2 The Problem with Using Reliability to Assess Validity

In this section, we derive a mathematically relationship between reliability to validity. For now, we assume the absence of sampling variability, and use "probability" and "proportion" interchangably. We begin with notation, which is also summarized as Appendix .

Consider C coders ($c = 1, \ldots, C$) tasked with coding D_i ($d = 1, \ldots, D_i$) documents into K ($k = 1, \ldots, K$) categories. Each document d has a true category $\pi_d \in \{1, \ldots, K\}$, and each category k has a true proportion of documents, $\bar{\pi}_k = \text{mean}_d[I(\pi_d = k)]$, which we collect as a vector $\bar{\boldsymbol{\pi}} = \{\bar{\pi}_i : k = 1, \ldots, K\}$ (and where $\text{mean}_j(a_j) = \sum_{j=1}^{J} a_j/J$ and J is the number of objects contributing to the summation). Analogously, each coder c's decision on document d is $y_d^c \in \{1, \ldots, K\}$ and the proportion of documents coder c puts in category k is $\bar{y}_k^c = \text{mean}_d[I(y_d^c = k)]$. We also collect these results in a vector, $\bar{\boldsymbol{y}}^c = (\bar{y}_1^c, \bar{y}_2^c, \ldots, \bar{y}_K^c)$. For a raw or naive estimate, we average these results over coders, producing $\bar{y}_k = \text{mean}_c(\bar{y}_k^c)$ with vector $\bar{\boldsymbol{y}} = (\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_K)$.

We now explicitly relate estimates to the truth by explicitly writing the "data coding generation process". We begin in scalar notation with the observed proportion in category k as coded by c as a function of all the true proportions:

$$\bar{y}_k^c \quad = \quad \sum_{j=1}^{K} \epsilon_{jk}^c \bar{\pi}_j$$

where ϵ_{jk}^c is misclassification probability — the probability that coder c classifies a document into category k if it is in fact in category j. This means that $\sum_{j=1}^{J} \epsilon_{jk}^c = 1$ because the coder must make some decision with each document. We define ϵ_{kk}^c as coder c's *validity* for category k: the proportion of documents the coder correctly classifies in category k from category k, so that if $\epsilon_{kk}^c = 1$, then $\bar{y}_k^c = \bar{\pi}_k$.

We then move to matrix notation by collecting ϵ^c_{jk} misclassification probabilities into a $K \times K$ misclassification matrix $\boldsymbol{E}^c$, with coder c's validities ϵ^c_{kk} on the diagonal. The off diagonal entries measure the probabilities of errors that coder c makes. We then write the observed proportions for coder c simply as (see Kuha and Skinner, 1997; Mikhaylov, Laver and Benoit, 2012),

$$\bar{\boldsymbol{y}}^c \quad = \quad \boldsymbol{E}^c \bar{\boldsymbol{\pi}} \tag{2.1}$$

If we observe $\boldsymbol{E}^c$ then, a simple correction reveals the true $\boldsymbol{\pi}$,

$$\left(\boldsymbol{E}^c\right)^{-1} \bar{\boldsymbol{y}}^c \quad = \quad \bar{\boldsymbol{\pi}} \tag{2.2}$$

Of course not only is $\boldsymbol{E}^c$ rarely if ever observed, but researchers usually do not include features designed to estimate or influence it. Thus, instead of focusing on *validity*, researchers are often encouraged to focus on the agreement between coders, which we refer to as *reliability*. To do this, researchers set up coding tasks so that at least two coders categorize a subset of documents. The intuition expressed in the literature is that improving reliability may also improve validity, although the feeling is not universal (**?**, p. 130). A key goal of this paper is to formalize these intuitions and connect these two concepts mathematically.

If two coders, 1 and 2, both code D documents, we define a confusion matrix element as the probability that coder 1 classifies a document as j and coder 2 classifies a document as k as $m^{12}_{jk} = \text{mean}_d[I(y^1_d = j, y^2_d = k)]$. The category-specific reliabilities are on the diagonal of this matrix, m^{12}_{kk}. Finally, as a summary, we define the overall reliability, the proportion of times coders 1 and 2 apply the same labels across all categories, as the sum of the diagonal elements of the confusion matrix:

$$a^{12} \quad = \quad \sum_{k=1}^{K} m^{12}_{kk}.$$

Some of the alternative scalar measures of reliability often begin with a^{12} and adjust for some concept of base levels of agreement that could be achieved by chance. The differences among these measures can be important for some purposes, but they do not materially affect our results.

Regardless of the measure used, researchers are often advised to focus on improving reliability, which is readily measured. Then, once reliability is improved as much as feasible given resources and time constraints, coder discrepancy is usually ignored for subsequent analyses. We show the problem with this approach by demonstrating that even high (but not perfect) levels of reliability do not satisfactorily constrain validity.

We begin formally connecting validity and reliability by assuming that each coder has a constant validity across categories.

Assumption 1. ***Constant Validity Assumption*** *Coder c's validity is constant across categories. This implies that it can be simplified as $\epsilon^c = \epsilon^c_{kk}$ for all k.*

Given Assumption 1 we can show

Proposition 1. *1) Suppose that $K > 2$ and coder 1 and coder 2 have overall reliability a^{12}. Then coder 1's validity is $\epsilon^1 \in [0,1]$ and coder 2's validity is $\epsilon^2 \in [0,1]$.*

2) Suppose $K \geq 2$ and Constant validity holds. Then Coder 1 and Coder 2 have maximum average validity $\dot{\epsilon}^{12} = \frac{1+a^{12}}{2}$ and if the coders have the maximum average validity $\epsilon^1 = a^{12} + 1 - \epsilon^2$.

Proof. Consider first the individual validity intervals for the coder. Fix an agreement rate between the coders as a^{12}. Suppose the coders agree on the same mistake a^{12} of the time and disagree and make distinct errors $1 - a^{12}$ of the time. Then if $K > 2$, $\epsilon^1 = 0$, $\epsilon^2 = 0$. If coder 1 is always correct and coder 2 is correct a^{12}, then $\epsilon^1 = 1$. If coder 2 is always correct and coder 1 is correct a_{12} then $\epsilon^2 = 1$.

Now consider the maximum joint validity $\dot{e}$. Note that for all pairs of documents the coders either agree or they disagree. For the coders to have maximum joint validity given the agreement rate then when they agree the coders must be correct and when they disagree one coder must be correct. Thus $a^{12} = P(\text{Coder 1 and Coder 2 Correct})$ and that $P(\text{Coder 1 Correct or Coder 2 Correct}) = 1$. Then, by inclusion/exclusion theorem:

$$\begin{aligned} a^{12} &= P(\text{Coder 1 Correct}) + P(\text{Coder 2 Correct}) - P(\text{Coder 1 Correct or Coder 2 Correct}) \\ &= \epsilon^1 + \epsilon^2 - 1 \end{aligned}$$

Algebraic manipulation then shows that the maximum joint validity $\dot{e} = \frac{1+a^{12}}{2}$ and that at this maximum $\epsilon^1 = a^{12} + 1 - \epsilon^2$.

□

Proposition 1 shows that without additional behavioral assumptions, even high coder agreement does not imply that we know that our coders are performing accurately. Even if coders agree perfectly, it may be the case that they always agree in error. To eliminate this possibility, we need to make an assumption that our coders are not *malicious*: able to perfectly misrepresent the answers in our study. Of course, we are often able to eliminate this assumption with careful selection, training, and monitoring of our coders. But obtaining a lower bound on the joint accuracy would still rely heavily on how pessimistic we are about our coders' joint validity. In Proposition 3 below show that the coders' joint validity can range from zero to the maximum described in Proposition 1. Therefore, an assumption alone would define the lower bound on the coders' performance.

Rather than focus on the pessimistic case, we first make the optimistic assumption that our coders are performing as well as possible, given their level of agreement. Formally, this is equivalent to assuming $\dot{e} = \frac{\epsilon^1+\epsilon^2}{2}$ is at a maximum, given the joint agreement. Proposition 1 does show that agreement provides an upper bound on the joint accuracy of our coders. We show that in the next section that assuming our coders have the maximum validity given their codes provides much more informative intervals, though requires more stringent assumptions about coder behavior. We view this assumption as much less optimistic than the current approach—assuming no coder error remains for high levels of agreement—but not so unrealistically pessimistic as to require an assumption that our coders are intentionally coordinating

to undermine our project. We will also see that our framework will provide a natural sensitive analysis to relax this optimistic assumption to the more pessimistic case where our coders are performing more poorly.

Proposition 1 leads to our second assumption:

Assumption 2. ***Wisdom of the Coders****. We will suppose that our coders' average validity is at a maximum. Equivalently, as we show below, this means if a plurality of coders of document d agree, they agree on the truth; if there is a tie, then at least one side identifies the truth.*

3 A Method for Incorporating Coder Error

In this section we will develop an algorithm that allows us to propagate the uncertainty from our coders, using Assumptions 1 and 2. Rather than only a point estimate, our procedure returns an interval, where the true value will lie if the assumptions are satisfied. We first provide intuition about how to use evaluation matrices to obtain bounds on the true proportions in each category.

3.1 Intuition about Bounds

To gain intuition about our approach, consider the simplest coding situation. Suppose that a coder is attempting to classify documents into two categories.

We suppose that our single coder is able to perfectly classify documents that truly belong in category 2, but is only correct ϵ^1 of the time when a document truly belongs to category 1. We can represent the coder's decisions with the following evaluation matrix

$$\boldsymbol{E}^1 = \begin{pmatrix} \epsilon^1 & 0 \\ (1-\epsilon^1) & 1 \end{pmatrix}$$

We can then represent the observed proportions, $\boldsymbol{E}^1\bar{\boldsymbol{\pi}} = \bar{\boldsymbol{y}}^1 = (\epsilon^1\bar{\pi}_1^1, (1-\epsilon^1)\bar{\pi}_1 + \bar{\pi}_2)$, where $\boldsymbol{\pi} = (\pi_1, \pi_2)$ is the true proportion in each category.

If we suppose that ϵ^1 is known, then we can solve for the true values of $\bar{\boldsymbol{\pi}}$. They are

$$\bar{\pi}_1 = \frac{\bar{y}_1^1}{\epsilon^1} \tag{3.1}$$

$$\bar{\pi}_2 = \bar{y}_2^1 - \frac{1-\epsilon^1}{\epsilon^1}\bar{y}_1^1 \tag{3.2}$$

Both equations are intuitive. The first notes that the only way our coder would classify something in category 1 is if it truly belongs to category 1, because she perfectly handles all documents that correctly belong to category 2. The proportion in Category 1 will be too small and dividing by ϵ^1 correctly adjusts the proportion. The second equation adjusts $\bar{y}_2^1$, removing $(1-\epsilon^1)$ of the incorrectly coded proportion of documents that should have been coded in Category 1 $\frac{\bar{y}_1^1}{\epsilon^1}$.

Suppose, now, that we don't know that the true values of ϵ^1, but rather some interval in which the values might lie—say $\epsilon^1 \in [a, 1]$. Then, we can plug in the values in Equation 3.1 and 3.2 to determine bounds on the proportion in each category. In particular, for this interval we find that $\pi_1 \in [\frac{\bar{y}_1^1}{1}, \frac{\bar{y}_1^1}{a}]$ and $\pi_2 \in [\bar{y}_2^1 - \frac{1-a}{a}\bar{y}_1^1, \bar{y}_2^1]$. Unlike other measurement error models, note that the observed proportions define the lower-bound for Category 1 and the upper bound for Category 2.

Throughout the paper we will consider a more general setting, where coders make errors in both categories, but we will assume each coder has equal validity across those categories (A1). This results in the following evaluation matrix

$$\boldsymbol{E}^1 = \begin{pmatrix} \epsilon^1 & (1-\epsilon^1) \\ (1-\epsilon^1) & \epsilon^1 \end{pmatrix}$$

Using this evaluation matrix and assuming we know ϵ^1 we can solve for $\bar{\pi}_1$ and $\bar{\pi}_2$ to obtain,

$$\bar{\pi}_1 = \frac{\epsilon^1}{2\epsilon^1 - 1}\bar{y}_1^1 - \frac{1-\epsilon^1}{2\epsilon^1 - 1}\bar{y}_2^1 \tag{3.3}$$

$$\bar{\pi}_2 = \frac{\epsilon^1}{2\epsilon^1 - 1}\bar{y}_2^1 - \frac{1-\epsilon^1}{2\epsilon^1 - 1}\bar{y}_1^1 \tag{3.4}$$

Like the case with errors in only one category, Equations 3.3 and 3.4 are intuitive. We want to reweight to reflect the errors made when assigning documents that should have been in a particular category to the other category, while also removing documents incorrectly placed in the category. And exactly like the simpler case, we might suppose that our validity lies in some interval $\epsilon^1 \in [a, 1]$. Even in this more general case, the observed proportions are at the extremes of our interval. If $\bar{y}_1^1 > \bar{y}_2^1$, then $\bar{\pi}_1 \in [\bar{y}_1^1, \frac{a}{2a-1}\bar{y}_1^1 - \frac{1-a}{2a-1}y_2^1]$.

Numerical Example to Obtain Bounds The reasoning thus far has been about a simple case with one coder. To derive our bounds we will use the information from double coded data to obtain information about each coders' validity and to therefore obtain bounds on the true proportion in each category. For a numerical example, we will suppose that the true proportion in each category is $\bar{\boldsymbol{\pi}} = (0.7, 0.3)$ and that the coders' proportions are,

$$\bar{\boldsymbol{y}}^1 = \begin{pmatrix} 0.8 & 0.2 \\ 0.2 & 0.8 \end{pmatrix} \bar{\boldsymbol{\pi}} = (0.62, 0.38)$$

$$\bar{\boldsymbol{y}}^2 = \begin{pmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{pmatrix} \bar{\boldsymbol{\pi}} = (0.66, 0.34)$$

With naive estimate $\bar{\boldsymbol{y}} = (0.64, 0.36)$. We will suppose that our coders have agreement rate 0.7.

Our approach to obtaining bounds will be to make assumptions that make use of information about our coders' agreement rate and their estimated proportions to inform the bounds that are used on the proportions. To gain intuition, we will walk through a set of assumptions that place increasingly specific assumptions on

the validity and therefore narrow the bounds. We will focus on obtaining bounds for the first category.

First, we will merely assume that our coders perform better than some baseline level and not make user of agreement rate. Specifically, we will initially suppose that our coders perform above some lower level of validity, $\epsilon^1_{11}, \epsilon^1_{22} \in [0.65, 1]$ and $\epsilon^2_{11}, \epsilon^2_{22} \in [0.65, 1]$ and make no other assumptions about the errors. By the logic of Equation 3.1 category 1 will be at a minimum if $\epsilon^1_{11} = \epsilon^2_{11} = 1$ and $\epsilon^1_{22} = \epsilon^2_{22} = 0.65$. This occurs because it implies that all the category 2 coding decisions are correct, while a portion of the category 1 decisions are erorrs. By the same logic the validity that would maximize category 1 is $\epsilon^1_{22} = \epsilon^2_{22} = 1$ and $\epsilon^1_{11} = \epsilon^2_{11} = 0.65$. Applying this logic yields a bound on $\bar{\pi}_1 \in [0.45, 0.98]$. We collect all the bounds in Table 1

Table 1: Bounds For Category 1 Under Different Assumptions

Assumption	Point Estimate	Minimum	Maximum
Truth	0.7	-	-
Naive	0.64	-	-
Minimum Validity	-	0.45	0.98
Constant Validity	-	0.64	0.97
Maximum Joint Validity	-	0.699	0.76
Maximum Joint Validity, Equality Constraint	0.7	-	-

In what follows we will make the more restrictive assumption that our coders' have constant validity—or that the probability a coder correctly classifies a document in each category is constant across categories. We defend the assumption below, but here we note that using Equation 3.3 for both coders implies that $\bar{\pi}_1 \in [0.64, 0.97]$.

At this point we can begin using our coders' agreement rate. In particular, we will suppose that our coders have maximum average validity (A1). This implies that $\epsilon^1 \in [0.7, 1]$ and that $\epsilon^2 = a^{12} + 1 - \epsilon^1$. Note, that this assumption eliminates the first two bounds, because those bounds require coders to perform either better than the agreement rate implies is possible or worse than the maximum joint validity. Once we constrain our search to only those values such that the coders have maximum joint validity, we obtain a much narrower interval $\bar{\pi}_1 \in [0.698, 0.76]$. Notice, also, that this interval does not contain the naive estimates.

In the two category case we are able to use one final piece of information to obtain the correct point estimate for the true proportions, under the assumptions we have made. (This is a special result that does not hold if $K > 2$ categories, though the intuition about the constraint remains useful). Notice that if we have the true evaluation matrix, $\boldsymbol{E}^1$, then $(\boldsymbol{E}^1)^{-1}\bar{\boldsymbol{y}}^1 = \bar{\boldsymbol{\pi}}$ and that $(\boldsymbol{E}^2)^{-1}\bar{\boldsymbol{y}}^2$. This implies that $(\boldsymbol{E}^1)^{-1}\bar{\boldsymbol{y}}^1 = (\boldsymbol{E}^2)^{-1}\bar{\boldsymbol{y}}^2$. Including this constraint—for the two category case—yields the exactly correct evaluation matrices and therefore provides the correct estimate of the proportion in Category 1, $\bar{\pi}_1 = 0.7$.

In what follows we use the intuition and assumptions from this section to derive a more genearl algorithm for estimating bounds on the proportion in categories.

3.2 Deriving the Intervals

To motivate our method for obtaining intervals, recall that Equation 2.2 shows that if we know $\boldsymbol{E}^c$ for coder c then we can use the coder's estimates to back out the true values $\bar{\boldsymbol{\pi}}$. Of course we do not know the true values of $\boldsymbol{E}^c$, but Proposition 1 provides an upper bound on the coders' validity—it defines a range for each coder that depends on the agreement between the pair of coders and the other coder's validity. Using this information, and other properties that a solution must have, we can define a set of pairs of matrices where the coders have maximum average validity. We then find the minimum and maximum values of $\bar{\pi}_k$ over this set of matrices.

To derive this interval, we first make the maximum average validity assumption (A2). (In the next section we generalize our procedure to make less optimistic assumptions about accuracy). The upper bound from Proposition 1 defines a range of potential pairs of values for ϵ^1 and ϵ^2: $\epsilon^2 \in [a^{12}, 1]$ and $\epsilon^1 = a^{12} + 1 - \epsilon^2$. For intuition note that if $\epsilon^2 = a^{12}$ then $\epsilon^1 = 1$ and if $\epsilon^2 = 1$ then $\epsilon^1 = a^{12}$.

Proposition 1 provides values for the diagonal elements of $\boldsymbol{E}^1, \boldsymbol{E}^2$, but provides no information about the off-diagonal elements, which we will optimize over to obtain intervals for each category. Before performing this optimization, however, we can use further constraints to limit the number of matrices we search over. First, as demonstrated in the previous section, if we know that if we have the true evaluation matrices, then $\left(\boldsymbol{E}^1\right)^{-1}\bar{\boldsymbol{y}}^1 = \left(\boldsymbol{E}^2\right)^{-1}\bar{\boldsymbol{y}}^2$. Second, we know that any solution must have all entries between zero and one. Therefore, we can restrict our attention to pairs of matrices $\boldsymbol{E}^c\bar{\boldsymbol{y}}^c \in \Delta^{K-1}$, where Δ^{K-1} is the $K-1$ dimensional simplex, for all coders c. Third, because of the structure of the evaluation matrices, we know that each column must sum to 1, so the off-diagonal elements for each column must sum to $1-\epsilon^c$ for $\boldsymbol{E}^c$. We can collect the pairs of matrices that satisfies these constraints into the set$\mathbb{E}$ with typical element $(\tilde{E}^1, \tilde{E}^2) \in \mathbb{E}$.

We can then optimize over the set of matrix pairs $\mathbb{E}$ to obtain an interval estimator for π_k. Note that because we have restricted $\mathbb{E}$ to contain matrices such that $(\tilde{E}^1, \tilde{E}^2) \in \mathbb{E}$ implies that $\left(\tilde{\boldsymbol{E}}^1\right)^{-1}\bar{\boldsymbol{y}}^1 = \left(\tilde{\boldsymbol{E}}^2\right)^{-1}\bar{\boldsymbol{y}}^2$, we know that

$$\min_{(\tilde{E}^1,\tilde{E}^2)\in\mathbb{E}} \frac{\left(\tilde{\boldsymbol{E}}^1\right)^{-1}\bar{\boldsymbol{y}}^1 + \left(\tilde{\boldsymbol{E}}^2\right)^{-1}\bar{\boldsymbol{y}}^2}{2} \quad = \quad \min_{(\tilde{E}^c)\in\mathbb{E}} \left(\tilde{\boldsymbol{E}}^c\right)^{-1}\bar{\boldsymbol{y}}^c$$

for coders $c = 1, 2$.

We will therefore define our interval estimator for $\bar{\pi}_k$ as ${\bar{\pi}_k}^{\text{int}}$

$${\bar{\pi}_k}^{\text{int}} \quad = \quad \left[\min_{\tilde{E}^c\in\mathbb{E}} \left(\tilde{\boldsymbol{E}}^c\right)^{-1}\bar{\boldsymbol{y}}^c|_k, \max_{\tilde{E}^i\in\mathbb{E}} \left(\tilde{\boldsymbol{E}}^c\right)^{-1}\bar{\boldsymbol{y}}^c|_k\right] \tag{3.5}$$

where $|_k$ denotes selecting the k^{th} element from a vector.

Proposition 2 shows that under our assumption of maximum average validity (Assumption 2) and constant category validity assumption (Assumption 1), ${\bar{\pi}_k}^{\text{int}}$ will contain the true values of $\bar{\pi}_k$. Further, the bounds are sharp, in the sense that making any additional use of the coding decisions—such as the error structure—would

require an additional set of assumptions about the proportions in the categories or further additional assumptions about each coders' validity.

Proposition 2. *Suppose that coder* 1 *and coder* 2 *have agreement* a^{12} *and that Assumptions 1 and 2 hold. Then* $\bar{\pi_k} \in \bar{\pi_k}^{int}$ *for each* k*.*

Proof. Assumptions 1 and 2 imply that the true evaluation matrices $(\boldsymbol{E}^1, \boldsymbol{E}^2) \in \mathbb{E}$. This follows because at the true evaluation matrices the coders will have constant validity, by Assumption 1, the inverted evaluation matrices will yield the correct answer and therefore be equal to each other, and the true proportions lie in the simplex. From Equation 2.2 this implies that $(\boldsymbol{E}^c)^{-1}\bar{\boldsymbol{y}}^c = \bar{\boldsymbol{\pi}}$ for both coder $i = 1, 2$. Thus $\bar{\pi_k} \in \bar{\pi}_k^{\text{int}}$. □

To obtain the interval that coder agreement, coders' decisions, and maximum joint validity implies we optimize over the pairs of matrices that are possible under the assumptions to find the minimum and maximum implied values of $\bar{\pi}_k$. The matrix inverses in Equation 3.5 make straightforward optimization difficult to obtain the identification region. Therefore, we use an iterative algorithm to obtain the interval estimates for $\bar{\pi}_k^{\text{int}}$. We describe this algorithm to obtain the interval in Appendix B.

3.3 What If There is No Truth?

Our algorithm is based on the assumption that there is a true proportion of documents in each category. This assumption will hold in many situations: coding rules should be written so that there is an unambiguous true proportion in each category. Yet, some scholars might dispute this assumption is reasonable. Instead they might argue that the truth depends on the model that coders have in mind. Even in the absence of an assumption of a gold standard truth our method still provides a useful and intuitive method for incorporating unceratinty from our hand coders.

To see that our method is still useful even if there is no ground truth, suppose for now that there is no true proportion of documents in each category. Rather, we will suppose that each coder has her own interpretation of the coding scheme, leading to a personal *model* of how documents should be assigned to categories. Under this assumption we can recast our algorithm as a method for interpolating across the coders' models. First, suppose that we are exclusively interested in coder 1's model. Then the evaluation matrix for coder 1 is the identity matrix—by defintion coder 1 has applied her model to the data. The evaluation matrix for coder 2 will have the agreement, a^{12} on the diagonal—where they agree coder 2 is applying coder 1's model—and then the disagreements have to be reassigned to the remaining categories. This will be possible if we can find a matrix with a^{12} down the diagonal $\boldsymbol{E}_{a^{12}}$ such that $\boldsymbol{E}_{a^{12}}^{-1}\bar{\boldsymbol{y}}_2 = \bar{\boldsymbol{y}}_1$. Likewise, if we impose coder 2's model, then the evaluation matrix for coder 2 is the identity model and coder 1's model has the agreement on the diagonal a^{12}.

Now, suppose that we are interested in a mixture between each coder's model of how the documents map into the categories. The coders agree a^{12} of the time, so they share the same model in those instances. Then, for the remaining (1 - a^{12})

we will suppose that we impose coder 1's model c_1 share of the documents and coder 2's model $c_2 = (1 - c_1)$ share of the documents. This implies that coder 1's evaluation matrix has $a^{12} + (1 - a^{12})c_1$ on the diagonal and coder 2's evaluation matrix has $a^{12} + (1 - a^{12})(1 - c_1)$. Note, that this implies that the average joint accuracy is $\frac{a^{12}+(1-a^{12})c_1+a^{12}+(1-a^{12})(1-c_1)}{2} = \frac{a^{12}+1}{2}$ or the maximum accuracy in the previous section.

For each c_1 we can find evaluation matrices $E^1_{c_1}$ and $E^2_{c_1}$ such that $\left(E^1_{c_1}\right)^{-1} \bar{\boldsymbol{y}}_1 = \left(E^2_{c_1}\right)^{-1} \bar{\boldsymbol{y}}_2$. For each category k we can then search over c_1 to obtain maximum and minimum values for each category. This, however, yields the equivalent interval as in Equation 3.5: optimizing over different values of c_1 is equivalent to searching over the set $\mathbb{E}$. Therefore, the interval estimator can also be interpreted as a method for characterizing the potential proportions, given coder disagreement.

3.4 Extending the Algorithm to Include Non-Overlapping Coders and To Include Sampling Uncertainty

It is often the case where a pair of coders will double code a subset of data and the remaining documents are coded by a single coder. We can modify our interval estimator to accomodate this additional classification data. Suppose that the coders double code a subset of documents $D = D_1 \cap D_2$. Define $\bar{\boldsymbol{y}}^c_D$ as the proportion for coder c in each category when applied to set D and define $\bar{\boldsymbol{y}}^c$ as the proportion in each category for all D_c coding decisions—including both double and single coded documents. Our constraint that the evaluation matrices, when inverted, return the same proportion can now only be applied to the subset of documents that are coded by both coders. That is $(E^1)^{-1} \bar{\boldsymbol{y}}^1_D = (E^2)^{-1} \bar{\boldsymbol{y}}^2_D$, but this need not necessarilly hold in general for $\bar{\boldsymbol{y}}^c$ because of differences in the documents included in coder c's sample–even if coder 1 and 2's evaluation matrices are identical.

To include the single-coded documents, we have to introduce an additional assumption: that our coders make the same kind of errors in the double-coded and the single-coded documents. We will call this the "evaluation matrix stability assumption".

Assumption 3. ***Evaluation Matrix Stability Assumption*** *Call coder c's evaluation for the double-coded documents $\boldsymbol{E}^c_D$ and for the single-coded documents $\boldsymbol{E}^c_{D_c}$. We will suppose that $\boldsymbol{E}^c_D = \boldsymbol{E}^c_{D_c} = \boldsymbol{E}^c$*

Assumption 3 is quite likely to hold in practice, particularly if the double-coded documents are a random sample from the entire collection of documents. We will define the set $\mathbb{E}$ as in Section 3.2, but only require the constraint to hold over the double-coded documents. Using these assumptions and optimizing over $\mathbb{E}$, Corollary 1 provides an interval estimator that uses all the coded data.

Corollary 1. *Suppose that coder 1 and coder 2 have agreement a^{12} and Assumptions 1, 2, and 3 hold. Define $\bar{\pi_k}^{int}$ as*

$$\bar{\pi_k}^{int} = \left[\min_{\tilde{E}^1, \tilde{E}^2 \in \mathbb{E}} \frac{(\tilde{E}^1)^{-1}\bar{\boldsymbol{y}}^1 + (\tilde{E}^2)^{-1}\bar{\boldsymbol{y}}^2}{2}|_k, \max_{\tilde{E}^1, \tilde{E}^2 \in \mathbb{E}} \frac{(\tilde{E}^1)^{-1}\bar{\boldsymbol{y}}^1 + (\tilde{E}^2)^{-1}\bar{\boldsymbol{y}}^2}{2}|_k \right]$$

Then $\bar{\pi}_k \in \bar{\pi}_k^{int}$

Proof. Under Assumptions 1, 2, and 3 the true evaluation matrices $(E^1, E^2) \in \mathbb{E}$. And if E^1 and E^2 are the true evaluation matrices across both double and single-coded documents, then, $\bar{\pi}_k = \frac{(E^1)^{-1}\bar{\boldsymbol{y}}^1 + (E^2)^{-1}\bar{\boldsymbol{y}}^2}{2}|_k$ □

Thus far we have assumed that we have estimates of the population agreement between coders and proportions. Of course, in any sample our estimates of those parameters will also be noisy. To capture the uncertainty from that noise we can perform a simple bootstrapping procedure. Specifically, we perform a vanilla bootstrap at the double-coded document level and then apply our algorithm. We can then obtain conservative bounds on the proportions by taking the minimum of the minimum of each category across the boot strap iterations, and the maximum of the maximum of each category of the boot strap iterations.

3.5 Simulation Evidence for the Bounding Algorithm

To demonstrate the performance of our algorithm we apply it to simulated data, generated under a variety of assumptions about how coders are performing. We then examine the proportion of time our algorithm covers the ground truth of the simulation and how the width of the intervals we construct depends upon the coders' agreement rate.

First consider the top-portion of Table 2. For these simulations we generate our synthetic coding data assuming our coders are performing with maximum joint validity for a varying number of documents (left-hand column). We then report whether we used a bootstrap to estimate the key quantities of interest (second column from left), whether the coders were assumed to have equal validity (or $\epsilon^c = \frac{1+a^{12}}{2}$) for $c = 1, 2$) or with one coder outperforming the other (third column from the left), and finally the proportion of times the estimated interval contains the true proportions. In each of these simulations, we vary the agreement rate from 0.6 to 0.95.

The top-portion of Table 2 shows that under a wide range of settings our interval estimator performs well. When there is a small sample, for example, the uncertainty in estimating the key quantities of interest leads the interval estimator to contain the truth a small number of times, 60% of the time. This, however, is remedied with the bootstrap. As the sample size increases the interval estimator contains the truth almost always. While the Propositions imply that this number should be 1 the interval will sometimes fail to contain the true value because our algorithm is an approximate inference procedure. We can easily improve the approximation by searching over a more granular set of values, at the cost of the algorithm taking longer to complete.

The bottom-portion of Table 2 relaxes the maximum validity assumption, instead supposing that our coders are making independent coding decisions. Note that the assumption of independence is not necessarily a large departure from our assumption of maximum validity. For example, suppose that coders agree on 81% of the documents. If the code independently then we would suppose each coder's validity is 90%, while under maximum joint validity each coder would have validity

Table 2: Simulation Evidence

No. Coded	Bootstrap	Simulation Type	Proportion Contained
		Maximum Validity	
100	No	Equal	0.60
100	Yes	Equal	0.93
500	No	Equal	0.93
500	Yes	Equal	1
1000	No	Equal	0.99
1000	Yes	Equal	1
1000	No	Differential	0.96
10000	No	Equal	0.98
10000	No	Differential	0.99
30000	No	Equal	1
30000	No	Differential	0.99
		Independent Coders	
100	No	Equal	0.90
1000	No	Equal	1
10000	No	Equal	1
10000	No	Differential, Non-Constant	0.65

90.5%.[1] Perhaps, then, it is not surprising that our algorithm tends to perform well when coders are making independent coding decisions.

The bottom two rows of Table 2 relax the assumption that coders have constant validity across categories. Here, we see that the interval estimator does not perform as well as the other settings. That said, the estimator does often contain the truth, and in the next section we introduce a way to perform a sensitivity analysis that will provide more conservative bounds and therefore improve the coverage of the algorithm.

One might be concerned that the coverage rates in Table 2 are merely the results of creating large intervals that are therefore not particularly informative. Figure 1 shows that this is not the case. In it, we plot the width of all intervals from the top-half of Table 2. The thick-black lines is a local linear regression between the interval width and the agreement rate among coders (where the variability emerges because of the random characteristics of the particular simulation). This demonstrates that for high levels of agreement the intervals are often quite narrow, and yet, still tend to cover the truth regularly.

3.6 Income Inequality News Coverage

Our first application of the algorithm is to an analysis of income inequality news coverage over time. The coded statements come from McCall (2013), which seeks to refute arguments that Americans car little about income inequality. A key component of her analysis is assessment of coverage of income inequality coverage in the

[1] Of course, the difference can be larger, depending on the agreement rate.

Figure 1: Interval Width Declines as Coders Agree More

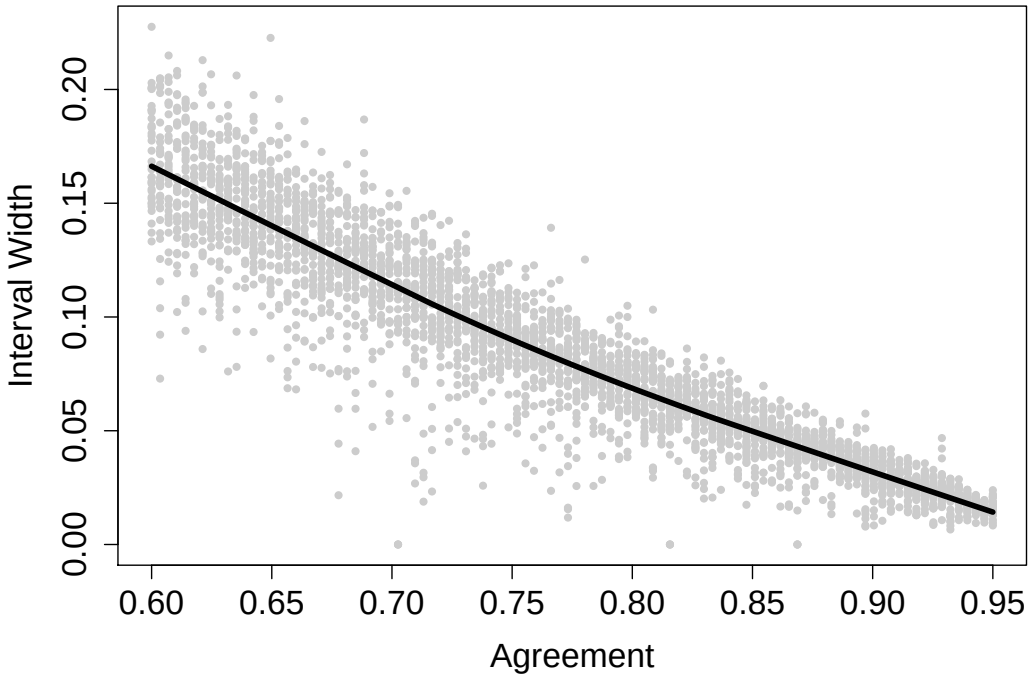


Table 3: Bounds on the Proportion of News Stories about Inequality

Method	Irrelevant	Inequality	Economy/Changes
Coder 1	0.46	0.12	0.41
Coder 2	0.47	0.11	0.42
Bounds, No Bootstrap, Min	0.46	0.07	0.41
Bounds, No Bootstrap, Max	0.49	0.12	0.44
Bounds, Bootstrap, Min	0.37	0.00	0.31
Bounds, Boostrap, Max	0.60	0.19	0.54

news. McCall (2013) employs a team of undergraduate coders to hand code news stories and we examine three categories the coders used: (1) whether the story was irrelevant for the analysis, (2) whether the story was about inequality, and (3) an aggregated category capturing whether the story covered stories broadly about changes in the economy. All together the validation set contained 121 total double-coded documents and the two coders have an agreement rate of 0.88.

Table 3 shows the average of the coders classifications across the three categories and provides the bounds both assuming the population variables are correct and for a bootstrapped values. The first two rows shows that both coders place a very similar proportion of documents in each category, reflecting the high agreement rate between the coders. Further, it shows that a relatively small proportion of the articles—only about 11.5%—are about inequality. The next two rows provide the bounds on the proportion in each category, treating the estimates of agreement and from our coders as population parameters. The bounds are fairly narrow, and show that between 7% and 12% of the articles are about inequality.

The bottom two rows, however, show the result of bootstrapping our algorithm 100 times and then forming the bounds. Including the estimation uncertainty in our bounding procedure widens the bounds considerably. In spite of the substantial intercoder agreement, the small number of double coded documents implies that there is still substantial uncertainty about the distribution of documents across the categories.

4 Relaxing the Maximum Joint Accuracy Assumption

Our analysis has relied upon the assumption that our coders are performing as accurately as possible given the level of agreement. Of course this could be a strong assumption. Coders are likely to make the same error or they may disagree and both be incorrect. In this section we describe a sensitivity analysis that relaxes the assumption of maximum average validity, providing a method for examining lower levels of joint validity.

Proposition 1 shows that the maximum average validity for the coders is $\frac{1+a^{12}}{2}$. The minimum average validity is always 0 (assuming $K > 2$ and that we have two coders). To relax the assumption of maximum joint validity, we can fix a level of joint validity $\gamma \in [0, \frac{1+a^{12}}{2}]$ and then examine the potential validity values at that

level of joint validity. Proposition 3 shows that at a wide range of values of γ, the agreement rate still provides information about the range of values for coder 1 and 2's validity.

Proposition 3. *Suppose $\dot{\epsilon} = \gamma$, coder 1 and coder 2 have agreement rate a^{12}, $K > 2$, and Assumption 1. Then if $\gamma \in [\frac{1-a^{12}}{2}, \frac{1+a^{12}}{2}]$, $\epsilon^1 \in [\gamma - \frac{1-a^{12}}{2}, \gamma + \frac{1-a^{12}}{2}]$ and $\epsilon^2 = 2\gamma - \epsilon^1$. If $\gamma \in [0, \frac{1-a^{12}}{2}]$ then $\epsilon^1 \in [0, 2\gamma]$ and $\epsilon^2 = 2\gamma - \epsilon^1$.*

Proof. Without loss of generality, consider ϵ^1 and fix $\gamma \in [\frac{1-a^{12}}{2}, \frac{1+a^{12}}{2}]$. Consider first the upper bound. The upper bound on ϵ^1 occurs if coder 1 is always correct when the coders disagree, which occurs $1 - a^{12}$. If this is true then coder 2 is only correct when she agrees with coder 1. This implies

$$\begin{aligned} \epsilon^1 &= \epsilon^2 + (1 - a^{12}) \\ 2\gamma &= \epsilon^1 + \epsilon^2 \end{aligned}$$

where the second equation follows from the definition of γ. Solving for ϵ^1 yields $\epsilon^1 = \gamma + \frac{1-a^{12}}{2}$. Note that at the maximum $\epsilon^2 = \gamma - \frac{1-a^{12}}{2}$.

While this same argument also provides the lower bound we will now derive it explicitly. Note that ϵ^1 is small as possible if coder 1 is wrong in all instances she disagrees with coder 2. This implies that $\epsilon^2 - (1 - a^{12})$ of the time coder 1 agrees with coder 2 and must be correct. Thus

$$\begin{aligned} \epsilon^1 &= \epsilon^2 - (1 - a^{12}) \\ 2\gamma &= \epsilon^1 + \epsilon^2 \end{aligned}$$

Solving for ϵ^1 yields $\epsilon^1 = \gamma - \frac{1+a^{12}}{2}$.

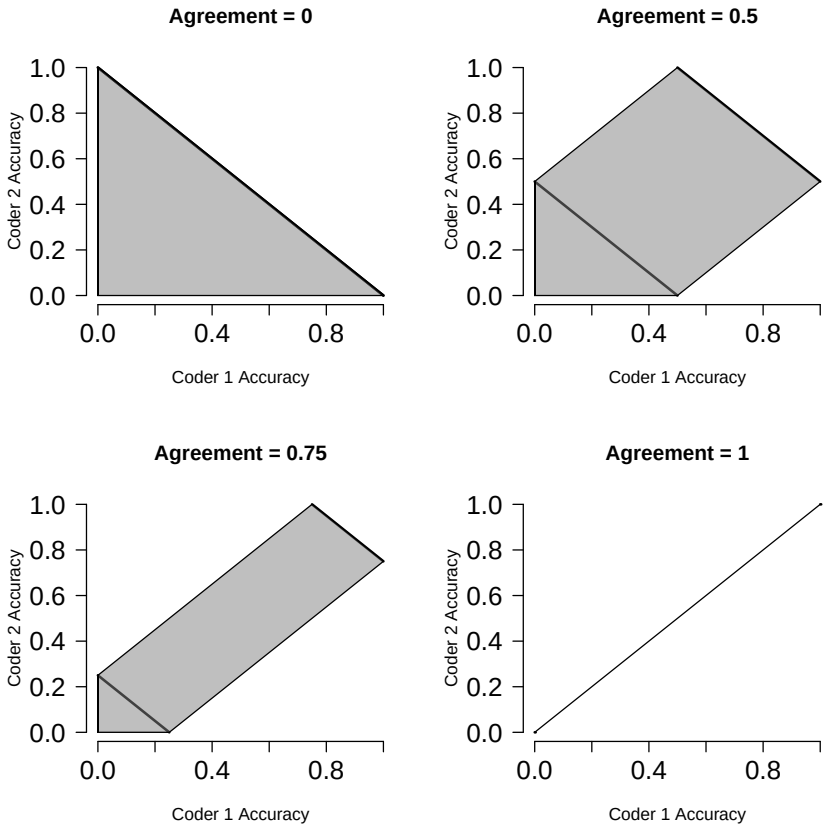
If $\gamma \in [0, \frac{1-a^{12}}{2})$, then the agreement rate provides no additional constraint on the average validity, because the value of γ is sufficiently small that the coders need not have any overlapping agreement. □

Proposition 3 shows that for a wide range of joint validity, the agreement rate provides information about how much the coders can disagree and therefore specifying a level of joint validity implies a range for the individual coders' validity. Proposition 3 provides a useful sensitivity analysis because it generalizes the upper bound in Proposition 1. To see this, note that the at the maximum joint validity $\gamma = \frac{1+a^{12}}{2}$. The maximum value for ϵ^1 is therefore $\epsilon^1 = \gamma + \frac{1-a^{12}}{2} = \frac{1+a^{12}}{2} + \frac{1-a^{12}}{2} = \frac{2}{2} = 1$. And the lower bound on $\epsilon^1 = \frac{1+a^{12}}{2} - \frac{1-a^{12}}{2} = \frac{2a^{12}}{2} = a^{12}$.

Figure 4 provides a visualization of how the range of joint validity that is possible under different levels of agreement. The horizontal axis in each plot is coder 1's accuracy, ϵ^1 and the vertical axis in each plot is coder 2's accuracy, ϵ^2. The grey area in each plot defines the set of possible values for our coder's accuracy for all the potential values of γ. The thick black lines in each plot define iso-joint validity lines where $\epsilon^1 + \epsilon^2 = 1 + a^{12}$ (line closest to the top-right corner) and $\epsilon^1 + \epsilon^2 = 1 - a^{12}$ (line closest to the origin). We can represent other iso-joint validity lines with the equation $\epsilon^2 = 2\gamma - \epsilon^1$.

Consider the plot in the top-left corner, which shows the range values when the coders disagree. In this case, all values in the lower-half of the unit square are

Figure 2: Geometric Interpretation of the Maximum Validity Assumption Suggests a Sensitivity Analysis



This figure provides a geometric interpretation of the relationship between agreement and accuracy. Note that as the agreement rate is lowered a wider range of accuracies are considered for each coder. This implies that examining how the interval increases as the agreement rate decreases provides a means of assessing the sensitivity of our assumption that our coders are performing as accurately as possible.

possible. For all potential values of γ under this agreement rate $\gamma = \frac{1+a^{12}}{2} = 0.5$, $\gamma[0, 0.5]$ the range of validity values for both coder 1 and coder 2 spans from 0 to 1. The top-right plot shows that as agreement increases, we have more information about the possible joint-validity values. And as the bottom right-plot shows, that as $a^{12} = 1$, then for any value of γ then only $e_1 = e_2 = \gamma$ is possible.

We use Proposition 3 and Figure 4 to relax the assumption that our coders are performing at maximum average validity. Our previous algorithm proceeds selecting the maximum value of $\gamma = \frac{1+a^{12}}{2}$. Setting lower values of γ then relaxes our optimistic assumption, allowing our coders to agree and be incorrect and for instances where they disagree and one coder is not correct. For any level of γ we can obtain a set of potential evaluation matrices with the appropriate joint accuracy, (E^1_γ, E^2_γ). Imposing the constraints, that $\left(E^1_\gamma\right)^{-1} \bar{\boldsymbol{y}}^1 = \left(E^2_\gamma\right)^{-1} \bar{\boldsymbol{y}}^2$ and that $\left(E^c_\gamma\right)^{-1} \bar{\boldsymbol{y}}^c \in \Delta^{K-1}$ for all c. Collect the pairs of matrices that satisfy the constraints for a given level of γ in the set $\mathbb{E}_\gamma$. For a given γ, the interval estimator is then:

$$\bar{\pi_k}^\gamma = \left[\min_{\tilde{E}^c \in \mathbb{E}_\gamma} \left(\tilde{E}^c\right)^{-1} \bar{\boldsymbol{y}}^c, \max_{\tilde{E}^c \in \mathbb{E}_\gamma} \left(\tilde{E}^c\right)^{-1} \bar{\boldsymbol{y}}^c\right] \tag{4.1}$$

where the difference between Equation 3.5 and Equation 4.1 the set of matrices considered at different values of γ. If we suppose that the coders average validity $\dot{\epsilon} \in [\gamma_{\text{low}}, \gamma_{\text{high}}]$, then we consider a range of potential γ values. We call this our $\tilde{\gamma}$-level average validity assumption,

Assumption 4. ***$\tilde{\gamma}$-level average validity*** *Coders have average validity $\dot{\epsilon} \in [\gamma_{low}, \gamma_{high}]$ where $\gamma_{high} \leq \frac{1+a^{12}}{2}$ and $\gamma_{low} \geq 0$.*

We can encode this broader set of joint validities by optimizing over this interval of gammas, $\pi^{\tilde{\gamma}}_k$,

$$\pi^{\tilde{\gamma}}_k = \left[\min_{\gamma \in \tilde{\gamma}} \pi^\gamma_k, \max_{\gamma \in \tilde{\gamma}} \pi^\gamma_k\right]$$

Proposition **??** shows that $\pi^{\tilde{\gamma}}_k$ will contain the true proportions.

Proposition 4. *Suppose Assumption 1 and 4 hold. Then $\bar{\pi}_k \in \pi^{\tilde{\gamma}}_k$.*

Proof. Assumptions 1 and 4 imply that the true evaluation matrices $(\boldsymbol{E}^1, \boldsymbol{E}^2) \in \mathbb{E}_{\tilde{\gamma}}$ and therefore $\bar{\pi}_k \in \pi^{\tilde{\gamma}}_k$. □

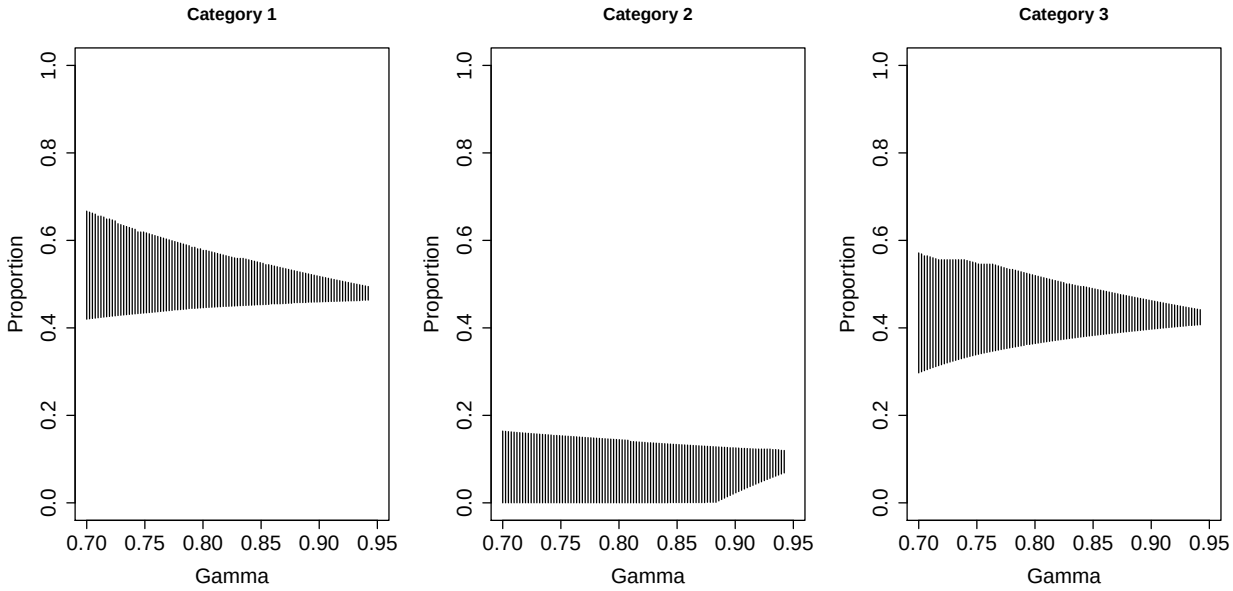
4.1 Examining the Proportion of Stories About Income Inequality

In Section 3.6 we reanalyze double-coded from McCall (2013) and find that sampling variability can result in wide intervals from our method, even when there is a relatively high rate of intercoder agreement. In this section we again analyze McCall (2013) and we show that even if we assume there is no sampling variability, but relax

the assumption of maximum validity, there can still remain considerable uncertainty about the true proportion of documents in each category.

Figure 3 shows how the intervals we obtain over the same set of documents from McCall (2013) vary as we consider a wider range of potential levels of joint validity. At the far right of each plot is the interval assuming maximum joint validity, $\frac{1+0.88}{2} = 0.94$ and those intervals replicate the intervals obtain from assuming no sampling variability in our analysis in Section 3.6.

Figure 3: Relaxing the Assumption of Maximum Joint Validity Yields Wider Intervals



As we move to the left in Figure 3, we consider intervals under lower levels of average joint validity. It is immediately clear that interval width increases substantially as we included instances where our coders are more error prone. And at the far-left of the plot, where we assume the coders are correct 70% of the time average, we obtain the widest intervals for each category. The interval for the Inequality category is [0.42, 0.67]; the interval for the inequality category is [0.0, 0.16]; and the for economy changing category we obtain [0.3, 0.57]. Again, even in a setting where there is a relatively high amount of intercoder reliability, we have certainty about the estimates only if we are willing to assume our coders are performing close to the maximum average joint validity.

5 Uncertainty When One Coder is Trusted More and Incorporating Uncertainty From Hand Codes When Evaluating Machine Learning Methods

Hand coding is increasingly used as an input for automatic classification algorithms and as a means to evaluate the performance of those methods. The prevailing standard for assessing the performance is to compare the performance of the classification algorithm to a gold standard. The most common gold standard are a set of hand coded documents from the coders. In any instance with a true gold standard—a set of classifications that are made without error—then we have all the information we would need to estimate $\boldsymbol{E}$ and to assess our model's performance.

Gold standards are also the prevailing method to assess how well teams of hand coders are performing. Researchers will hand code a subset of documents themselves and compare their coding decision to the decisions' of hired coders. The assumption is that a researcher on the project is able to produce infallible classifications of the coding scheme.

A true gold standard, however, is rarely available. Even when professors and advanced graduate students are coding documents, they are likely to make errors. A common reply is that the professor's coding decision becomes the *de facto* correct answer. Or, groups of coders may come together and determine a true label after deliberation. Both procedures, however, render the classification both unreplicable and obscures the true meaning of the categories. For the codes to be replicated in the future and to be derived from the stated rules, they should not exist merely in one person's head or as the result of deliberation as a group. They must be clearly stated as a set of rules that other scholars could use.

Further, recognizing that all coders makes errors implies that there is no true gold standard. The literature describes this error as alloying the gold standard (Wacholder, Armstrong and Hartge, 1993)—with the notion that there are errors alloyed with the true correct answers. The presence of the error implies that the computed evaluation matrix $\boldsymbol{E}$ no longer is correct and the adjustments that result will provide inaccurate measures of the true underlying proportions (Wacholder, Armstrong and Hartge, 1993).

An alloyed gold standard, however, can still be useful for assessing a model's performance. Our algorithm in the previous section and Proposition 2 assumes that we have no specific information about the coders' accuracies, so either could perform better. An alloyed gold standard, however, implies that one coder is more accurate than the other, which narrows the accuracy interval for our coders.

Specifically, suppose that our pair of coders, 1 and 2 have agreement a^{12}. Then Proposition 1 shows that $\epsilon^1 + \epsilon^2 = 1 + a^{12}$. If the coders have equal accuracy then $\epsilon^2 = \epsilon^1 = \frac{1+a^{12}}{2}$. We can write then write the difference in the accuracy of the coders as

$$\epsilon^1 \quad = \quad \frac{1+a^{12}}{2} + \lambda \frac{1-a^{12}}{2}$$

where $\lambda \in [-1, 1]$. If $\lambda = 0$, then the coders have equal accuracy, if $\lambda = -1$ then coder 2 is the gold standard ($\epsilon^2 = 1$), and if $\lambda = 1$ then coder 1 is providing

a true gold standard ($\epsilon^1 = 1$). Therefore λ captures how much more accurate one coder is relative to the other coder. In the absence of additional information about how well our coders perform we might restrict the values of λ that we consider for the intervals. If we make strongers assumptions about λ—narrowing the interval of values where we think it resides—will yield smaller intervals for our estimates. At the cost, of course, of a stronger set of assumptions to obtain that narrow identification region.

Our method for analysis also provides a natural way to incorporate the potential uncertainty from hand coding when analyzing machine learning methods. While imperfect intercoder agreement is a universally recognized fact of every hand coding exercise, evaluation of classification methods often fail to incorporate the uncertainty this lack of agreement implies when evaluating methods. This might be problematic, because it can lead to over confidence that one particular method will perform better in the future, or potentially lead to overconfidence in the future performance of a method.

We can extend our approach in two ways to provide information about how well a particular method performs after incorporating coder uncertainty: creating best and worst case scenario bounds on agreement with a gold standard for a new coder and by providing a simulation procedure to better understand how a method might fare if new coders were asked to complete the same task.

5.1 Best and Worst Case Bounds on Method Agreement After Obtaining a New Coder

Using the assumptions we have made thus far and the agreement between an automatic classification method and a proposed gold standard provides all the information necessary to derive best and worst case scenario bounds for a method's agreement if a new coder is solicited, provided we are willing to assume the coder has a particulate agreement rate with the previous gold standard.

In particular, suppose that we have a proposed gold standard $\boldsymbol{y}^{\text{gold}}$, with corresponding proportions in each category $\bar{\boldsymbol{y}}^{\text{gold}}$ and a machine classification of those same documents $\boldsymbol{y}^{\text{machine}}$. We will suppose that the gold standard and the machine have confusion matrix $\boldsymbol{C}$ where entry $m_{ij}^{\text{machine,gold}}$ counts the number of times the machine the object in category i and the gold standard codes the object in category j. We will define a column normalized version of the matrix $\tilde{\boldsymbol{C}}$ where we require each column sum to 1.

Given this confusion matrix, we can define best and worst case bounds on the agreement rate between the machine method and the new "gold standard" we would obtain from a new coder, provided we fix an agreement rate between our human coders. Suppose that a^{12} represents the proportion of the time that we suspect the human coders will agree. To obtain the maximum new agreement rate, note that this is obtained if (1) the new human coder agrees with the machine in all instances where the original coder agreed and (2) the new human coder agrees with the machine when the original human coder disagreed—subject to the constraint on the agreement rate between the human coders. For each category, k, the new max agreement between the machine and human is,

$$a_{\text{max},k}^{\text{gold,machine}} = \min(a^{12}, \tilde{m}_{kk}) + \sum_{j \neq k}^{K} \max\left(\min(\tilde{m}_{jk}, (1 - a^{12}) - \sum_{l \neq k}^{k-1} \tilde{m}_{lk}), 0\right)$$

The upper bound new agreement is then obtained by weighting the per-category agreement by the proportion in each category from the original gold standard: $a_{\text{max}}^{\text{gold,machine}} = \sum_{k=1}^{K} a_{\text{max},k}^{\text{gold,machine}} \bar{y}_k^{\text{machine}}$.

The minimum agreement rate between the new gold standard and the machine coder is obtained if the new coder disagrees as much with machine as possible, subject to the constraint that the new coder agree as much as required with the original gold standard—a^{12}. Specifically, for category k the minimum agreement rate between the new gold standard and the machine is,

$$a_{\text{min},k}^{\text{gold,machine}} = \max\left(a^{12} - \sum_{j \neq k}^{K} \tilde{m}_{jk}, 0\right)$$

Intuitively, the method says that the minimum will be obtained if as much as the possible agreement between the human coders occurs in cases where the original coder and the machine disagreed, subject to the constraint that the human coders agree at least (a^{12}) of the time.

We can then weight across categories to obtain the new minimum agreement rate $a_{\text{min}}^{\text{gold,machine}} = \sum_{k=1}^{K} a_{\text{min},k}^{\text{gold,machine}} \bar{y}_j^{\text{machine}}$

This procedure can be useful, but often provides extremely wide bounds on the machine's performance—even if the coder generates a new gold standard under agreement rates in relatively high agreement with the original coder. Further, the bounds rely upon an exceedingly unlikely coding decisions that are new coders might make. In the next section, we define a simulation based procedure that generates potential new agreement rates, given rates of agreement among new coders.

5.2 A Simulation Procedure to Encode Coder Error in Gold Standards

In this section we propose narrower bounds on agreement. Specifically, we introduce a simulation based procedure that encodes the uncertainty that comes from our coders. To perform the simulation we simulate confusion matrices between our original proposed gold standard and hypothetical new gold standards. Then, using the new synthetic confusion matrix, we generate new gold standard codings and assess agrement. Repeating the procedure many times over simulates drawing new hand coders.

Specifically, we first set an agreement rate a^{12}. We generate a new simulated confusion matrix $\boldsymbol{C}^{\text{sim}}$, which is normalized so each column sums to 1. We set all diagonal entries $m_{jj}^{\text{sim}} = a^{12}$. For each category, we then draw a realization from a $K-1$ Dirichlet distribution, with parameters $\boldsymbol{z}_k$, $\boldsymbol{g}_k \sim \text{Dirichlet}(\boldsymbol{z}_k)$. The values of $\boldsymbol{z}_k$ can be set to reflect beliefs about confusion among future coders or set equal to

a vector of 1's to reflect a uniform distribution. The off diagonal elements are then $\boldsymbol{m}_{-k,k} = (1 - a^{12}) \times \boldsymbol{g}_k$.

For each document d we then draw the new gold standard's label, $y_d^{\text{new gold}}$ according to $y_d^{\text{new gold}} \sim \text{Multinomial}(1, \boldsymbol{c}_{y_d^{\text{gold}}})$, or from a multinomial distribution that corresponds to the label of the current gold standard. For each iteration of the simulation we can then obtain a new agreement rate with the machines prediction $a^{\text{new gold,machine}}$ by taking the agreement rate between our simulated gold standard and the machine's predictions. Performing this simulation many times provides a distribution of agreement rates that incorporate one type of uncertainty from hand coders.

5.3 Incorporating Coder Uncertainty With Deep Learning Models

Tai, Socher and Manning (2015) reports the highest agreement to date on the Stanford Sentiment Tree Bank, classifying the sentiment of the short phrases. We use the simulation based algorithm to examine how the methods compare as we introduce the uncertainty from intercoder disagreement of their reported. We focus on the three best performing methods, based on the reported agreement with the gold standard, along with RNTN, a previous top performing method (Socher et al., 2013).[2]

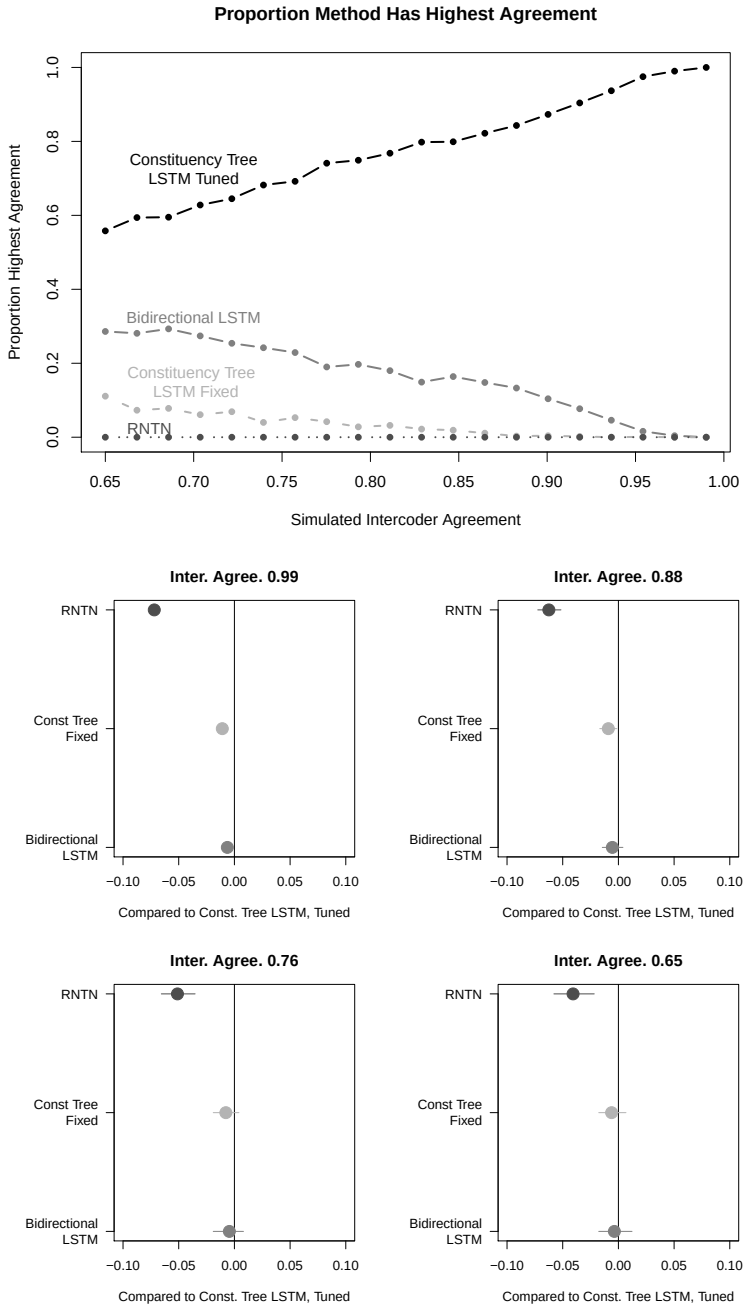
The left-hand plot in Figure 4 shows the proportion of simulations the top performing method, Constituency Tree LSTM Tuned, remains the top performing method as we allow for more disagreement between the gold standards. Indeed, at very high levels of simulated agreement between the coders—0.8 agreement between the original and new gold standards—Constituency Tree LSTM with tuned vectors is only the top performing method 76.8% of the time. At the lowest simulated level of agreement between the hand coders, Constituency Tree LSTM is the highest performing method in only 55.8% of the simulations. The right-hand plot shows the difference in agreement rate between the best performing method, Constituency Tree LSTM Tuned, and the other methods as agreement rate with the new coder decreases. This figure shows that even at reasonable levels of intercoder agreement the differences shrinks as the coders disagree more often—even though the agreement rates between the coders are relatively high for hand coding exercises.

6 Uncertainty With More than Two Coders for Each Document

We have so far focused on developing intervals when we have two coders, because this is the most common and often most affordable method used to assess the reliability and validity of document codes. In this section we show how adding additional coders to code a single document can provide even more informative bounds on coders' maximum possible accuracy. This more informative bound arises because our

[2] In our supplementary appendix we provide the results for all 9 methods.

Figure 4: Including Uncertainty from Intercoder Agreement in Evaluating Sentiment Classification Methods



assumption of the maximum possible joint validity and the addition of coders allows us to potentially identify instances where some some coders are *wrong*, allowing us to bound the maximum accuracy of a coder away from 1.

To examine the case with more than two coders, we first introduce some additional notation. For simplicity (and without loss of generality) suppose that our C coders have coded D documents into one-of-K categories. Define the K^C set of potential labels for each document in the set $\boldsymbol{\mathcal{T}}$, with one instance for document d as $\boldsymbol{y}_d \in \boldsymbol{\mathcal{T}}$. The C-element long vector $\boldsymbol{y}_d$ collects the coding decisions across coders and y_d^c describes the coding decision for coder c on document d. We will collect the coding decisions into an $C \times D$ matrix $\boldsymbol{Y}$.

To establish our bounds on coder accuracy with multiple coders, we will examine aggregations of our coder's decisions. An equivalent view of the coding process is that we are asking our coders to vote on the label for a particular document, where we tally the codes from our coders as votes for a document's label. Formally, we define the function $v : \boldsymbol{T} \rightarrow \mathbb{C}^K$ as a function that performs this aggregation. For any vector $\boldsymbol{y}_d$ we will say that coder c is in the plurality for document d if $y_d^c = \max_k v(\boldsymbol{y}_d)$ and $|\max_k v(\boldsymbol{y}_d)| = 1$: coder c is in the plurality when her label for document d agrees with the most popular label for document d adn there is a single most-popular label for document d. We will denote the proportion of times that coder c is in the plurality with $p^c = \text{mean}_d\left(I(y_d^c = \max_k v(\boldsymbol{y}_d), |\max_k v(\boldsymbol{y}_d)| = 1\right)$. We will similarly say that coder c is part of a plurality tie if the number of categories at the maximum is 2 or more, or $|\max_k v(\boldsymbol{y}_m)| > 1$ and $y^c d \in \max_k v(\boldsymbol{y}_d)$. We can similarly define the proportion of times that coder c is part of a plurality tie, $t^c = \text{mean}_d\left(I(y_d^c \in \max_k v(\boldsymbol{y}_d), |\max_k v(\boldsymbol{y}_d)| > 1\right)$.

Before stating our result, we first generalize our assumption of maximum joint validity for our coders given their agreement. Specifically, we will say assume that our coders have the maximum average validity given the agreement across coders. Formally, we will assume that $\dot{\epsilon} = \text{mean}_c\left(\epsilon^c\right)$ is as large as possible, given the coding decisions $\boldsymbol{Y}$. This is a direct generalization from the case where $C = 2$ our Assumption 1, so restate the assumption here:

Assumption 5. ***Wisdom of the Coders (General)*** *We will suppose that* $\dot{\epsilon} = \text{mean}_c\left(\epsilon^c\right)$ *is at a maximum, given* $\boldsymbol{Y}$.

Proposition 5 establishes the bound on coder validity with a larger set of coders.

Proposition 5. *Suppose that there are C coders who code D documents and Assumptions 1 and 5 hold. Then,*

$$\epsilon^c \quad \in \quad [p^c, p^c + t^c]$$

where p^c is the proportion of times c is among the clear plurality of coders and t^c is the proportion of times c is among the pluralities in a plurality tie.

Proof. We first consider the lower bound, given maximum average validity. The assumption that our coders have the maximum average validity given the level of agreement implies that in instances where all the coders agree that the coders are correct. The coders can have two different patterns of disagreement. There can

either be a clear plurality of voters for a category, or there will be a plurality tie. Assume first there is a clear plurality of coders. Then our assumption of maximum joint validity implies that those coders are correct. (If not, then average validity could be improved by making them correct without affecting the agreement). This implies that at worst $\epsilon^c = p^c$. To determine the upper bound suppose that coder c is in a plurality tie t^c of the time. Our assumption of maximum average validity and the coder's agreement provides no information about which group in the plurality tie is correct. Therefore, coder c's accuracy is highest if each instance she is involved in a plurality tie she is correct. This implies that maximum value of $\epsilon^c = p^c + t^c$. Thus $\epsilon^c \in [p^c, p^c + t^c]$. □

Proposition 5 shows that with additional coders we can obtain a more informative upper bound on a coder's decision. The extra information arises because our assumption of maximum accuracy given the agreement level among coders implies that any instance where a plurality of voters agree, they must be correct. This implication is very similar to *Wisdom of the Crowds* results. Proposition 5 generalizes the upper bound from Proposition 1. If $C = 2$, then $p^c = a^{12}$ and $t^c = (1 - a^{12})$. If we assume that coder 1 is always correct when they disagree with coder 2 then the upper bound on coder 1's accuracy is $\epsilon^1 = 1$.

While there is more information from coders that could be used to incorporate information into our algorithm, our empirical case studies have shown us that this creates an overly restrictive set of assumptions that often implies an empty interval. Instead, we use the results in Proposition 5 to apply a modified version of our algorithm to each pair of coders. Suppose that we have a collection of coders C coders and that each coder c and coder j's validity lies in the interval from Proposition 5. For each pair of coders c and j we can define the new pairs of evaluation matrices, $\tilde{E}^c, \tilde{E}^j \in \mathbb{E}^{cj}$ such that $\left(\tilde{E}^c\right)^{-1} \bar{y}^c \in \Delta^{K-1}$, $\left(\tilde{E}^j\right)^{-1} \in \Delta^{K-1}$ and $\left(\tilde{E}^c\right)^{-1} \bar{y}^c = \tilde{E}_j^{-1} \bar{y}_j$, and the diagonal elements satisfy the interval from Proposition 5 and $\epsilon^c + \epsilon^j \leq p^c + p^j + t^c + t^j - t^{cj}$ where t^{cj} is the proportion of the time coder c and coder j are in the plurality tie together. We can search over all pairs of coders c, j to define our new interval,[3]

$$\bar{\pi_k}^{\text{int}} = \left[\min_{i,j} \min_{\tilde{E}_i \in \mathbb{E}_{ij}} \tilde{E}_i^{-1} \bar{y}_i|_k, \max_{i,j} \max_{\tilde{E}_i \in \mathbb{E}_{ij}} \tilde{E}_i^{-1} \bar{y}_i|_k \right] \tag{6.1}$$

6.1 Analyzing the Credit Claiming Statements in House Press Releases

Grimmer, Westwood and Messing (2014) analyze the rate of credit claiming across members of the US House. Grimmer, Westwood and Messing (2014) use a collection of 789 triple-coded press releases as a training set for a supervised learning procedure. Here, we analyze only the hand coded press releases to better understand the uncertainty from hand coding. We work with a modified version of their

[3]Equation 6.1 will provide a conservative estimate of coder uncertainty, because it fails to include all the information from the multiply coded data that might further restrict the range of validity that we search over.

Table 4: Reanalyzing the Rate of Credit Claiming in a Sample of US House Press Releases

Coder Number	Advertising	Credit Claiming	Other
1	0.29	0.21	0.51
2	0.23	0.25	0.51
3	0.32	0.30	0.38
Bounds			
Minimum	0.20	0.20	0.51
Maximum	0.26	0.24	0.56

data, analyzing three categories of press releases: advertising, credit claiming for money, and all other press releases. The first three rows in Table **??** presents the proportion of press releases in each of the categories from each of the coders. The relatively close alignment across the categories reflects the high rate of agreement among the coders. All three coders agreed 68.1% of the press releases; coder 1 and 2, while coder 3 disagreed in 9.3% of the press releases; coder 1 and coder 3 agreed while coder 2 disagreed in 7.2% of the press releases; coder 2 and coder 3 agreed, while coder 1 disagreed in 13.7% of the press releases; and all three coders disagreed in only 1.6% of the press releases. We use this agreement rate as input to obtain bounds on the proportion in each category, following Equation 6.1.

7 Propogating Uncertainty from Coder Disagreement to Other Parameter Estimates

The proportion of documents that lie in a set of categories is often an intermediate outcome of interest. The true quantity of interest is often the effect of some intervention on the prevalence of that document or how the prevalence of some category of interest affects some other outcome of interest. In this section we describe a straightforward method for incorporating our uncertainty as a result of our coders' disagreement into the next stage of an analysis. For simplicity we will focus on a linear regression with a regression coefficient as the quantity of interest, but our method extends to other parametric models and more elaborate quantities of interest in a straightforward way. Proposition 7 in the Appendix shows this straightforward generalization.

Suppose that our team of coders classifies documents from S sources with the true proportion in category k from source s given by $\bar{\pi}_k{}^p$. A source might be a politician, a survey respondent, a country's leaders, a year a document was written, or any other covariate stratum. Rather than observe a K component vector $\bar{\boldsymbol{y}}^c$ for each coder c, we now obtain a $K \times S$ matrix of proportions for each coder, $\bar{\boldsymbol{Y}}^c$. We denote the s^{th} row of this matrix for the c^{th} coder as $\bar{\boldsymbol{y}}_s^c$. This row provides the proportion of observations in each category for a particular source of interest.

We will suppose that our interest is in understanding the relationship between a vector of covariates for each source $\boldsymbol{X}_s$—which we collect into $\boldsymbol{X}$—and its relationship with the prevalence of category k across the S sources. We will suppose the

following data generating process:

$$\bar{\pi_k}^s = \boldsymbol{\beta}\boldsymbol{X}_s + \nu_s$$

where ν_p is independent and identically distributed errors, with $E[\nu_s|\boldsymbol{X}_s] = 0$. Our quantity of interest is $\boldsymbol{\beta}$ or the conditional relationship between covariates prevalence of category k across sources. If we observe $\bar{\pi_k}^s$ for each source s then the usual regression estimator will be an unbiased and consistent estimator of $\boldsymbol{\beta}$.

As we have argued throughout this paper, it is unlikely that our coders have performed without error. The consequences of coder error for our regression can be substantial. If our coders code with error, then we can write the observed vector of proportions for category k $\bar{y}_k$ as $\bar{y}_k = \pi_k + \delta_k$, where δ_k is the vector of average bias across coders[4] from coding for each source. Unless $\delta_k^s = 0$ for each source s the regression estimator will no longer be consistent or unbiased because:

$$\begin{aligned}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\bar{\boldsymbol{y}} &= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\bar{\boldsymbol{\pi}} + \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{\delta} \\ &= \boldsymbol{\beta} + \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{\delta} \qquad (7.1)\end{aligned}$$

The second term remains so long as our coders code with error. While the regression will be biased and inconsistent, we can develop an interval estimator that will contain the true value of the regression coefficient of interest. To do, we first have to make an additional assumption about the errors our coders make.

When focused on the proportion of documents in each category, it is true by construction that there is one matrix that relates our coders' decisions to the true proportions. When we examine many sources, however, it is possible that coders make different sorts of errors from different sources. We will denote the errors that coder c makes on documents from source s as $\boldsymbol{E}_s^c$. The specific errors that coder c makes on source s might be the result of coders' biases toward the source—such as liberal college students coding documents from Republicans—or the result of some sources using more ambiguous language—which might occur with language from historical sources.

In many instances, however, coders are likely to make the same errors across sources. This might be particularly likely if coders are randomly assigned to documents that are from a relatively similar set of sources—say official documents from members of Congress—and unnecessary information is removed to limit the possibility that coders' biases will affect their decisions.

To develop our interval estimator for the coders' proportions, we will make the additional assumption that our coders make the same errors across sources. Specifically we will assume that $\boldsymbol{E}_s^c = \boldsymbol{E}^c$ for all sources s.

[4]This use of the term bias is different than what can be found sometimes in the literature on handcoding. Di Eugenio and Glass (2004) for instance use this term to indicate the degree to which the coders disagree. Another example from Byrt, Bishop and Carlin (1993): If Observers A and B differ in their assessment of the frequency of occurrence of a condition in a study group, we say that there is a bias between the observers. (page 424).

Assumption 6. ***Source Evaluation Matrix Stability Assumption*** *For all sources s, we assume that $\boldsymbol{E}_s^c = \boldsymbol{E}^c$ for all coders c.*

In other words, our coders can make their own specific errors, based on their confusion about the coding rules or their attention to the documents, but across sources, s, the coders make the same errors—they aren't modifying their errors according to the source of the document they are coding. This assumption, in addition to likely being true, will substantially simplify the development of an interval estimator for quantities of interest. Under this assumption, then, implies that for the true evaluation matrix $\boldsymbol{E}^c$, $\left(\boldsymbol{E}^{-1}\right)^c \bar{\boldsymbol{y}}_s^c = \bar{\boldsymbol{\pi}}^s$ for all sources s.

To construct the interval estimator, suppose our quantity of interest is the tth coefficient β_t. We will assume again that our two coders have the maximum joint validity given the level of agreement. We will further restrict the set of evaluation matrices to be just those that, when inverted and applied to the estimated proportions $\left(\boldsymbol{E}^c\right)^{-1} \bar{\boldsymbol{y}}^c$ the result lies in the simplex and where $\left(\boldsymbol{E}^1\right)^{-1} \bar{\boldsymbol{y}}^1 = \left(\boldsymbol{E}^2\right)^{-1} \bar{\boldsymbol{y}}^2$. Define $\mathbb{E}$ as the set of pair matrices that satisfy these conditions. Then we can define an interval estimator for β_t^{int} as,

$$\beta_t^{\text{int}} = \left[\min_{(\boldsymbol{E}^1) \in \mathbb{E}} \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{X}' \left(\left(\boldsymbol{E}^1\right)^{-1} \bar{\boldsymbol{y}}^1 \right) |_t, \max_{(\boldsymbol{E}^1) \in \mathbb{E}} \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{X}' \left(\left(\boldsymbol{E}^1\right)^{-1} \bar{\boldsymbol{y}}^1 \right) |_t \right]$$

Where, again $|_t$ selects the t^{th} element from the vector.

Proposition 6 shows that, β_t will be contained in β_t^{int}.

Proposition 6. *Suppose that Assumptions 1, 2, and 6 hold for all categories and all coders. Suppose that using the true proportions $\bar{\boldsymbol{\pi}}$ and covariates $\boldsymbol{X}$ yields the regression coefficient is β_t. Then $\beta_t \in \beta_t^{int}$.*

Proof. If coders have maximum average validity, then the true evaluation matrices $(\boldsymbol{E}^1, \boldsymbol{E}^2) \in \mathbb{E}$. This implies that $\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{X}' \left(\boldsymbol{E}^1\right)^{-1} \bar{\boldsymbol{y}}^1 = \boldsymbol{\beta}$ and $\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{X}' \left(\boldsymbol{E}^2\right)^{-1} \bar{\boldsymbol{y}}^2 = \boldsymbol{\beta}$. Therefore $\beta_t \in \beta_t^{\text{int}}$ □

In Appendix C we show that our interval estimator generalizes for any estimator that would be consistent if the true proportions were used as either a dependent or independent variables.

7.1 Analyzing the Taunting Rate Over Time

Grimmer, King, and Superti (2015) (GKS) analyze the rate US Senators engage in taunting: explicit, public, and negative attacks on the other party or its members. To measure the rate of taunting GKS engages in a massive hand coding effort, employing a team of coders to label thousands of Senate speeches. GKS ensured that 10% of all the speeches were double coded, to measure the agreement of the coders. In this section we use a subset of the hand-coded data to examine how the rate of taunting has varied over time and to demonstrate the need to adjust for coder disagreement.

To apply the algorithm from the previous section, we aggregate their more granular coding into three categories: taunting, policy discussion, and other types of Senate speech. We then focused on the proportion in each category from the two most prolific coders in the data set. This yielded 379 double coded Senate speeches and the two coders had an overall agreement rate across the three categories of 0.83. Altogether the two coders classified and 11,136 total speeches coded.[5] Using the complete set of codes we measured the proportion of speeches from each senator in each of three categories for the two coders.

Using the agreement rate between the coders and the measures of the proportion in each category, we used the procedure in the previous section to measure the average taunting rate from the 101st to the 109th Congress. The thick-black line in Figure 5 shows the intervals on the average taunting rate in each Congress from our bounding procedure, while the dashed line demonstrate the average taunting rate if we fail to propogate coder uncertainty.

Figure 5: Taunting Rates in the US Senate, Across Congresses

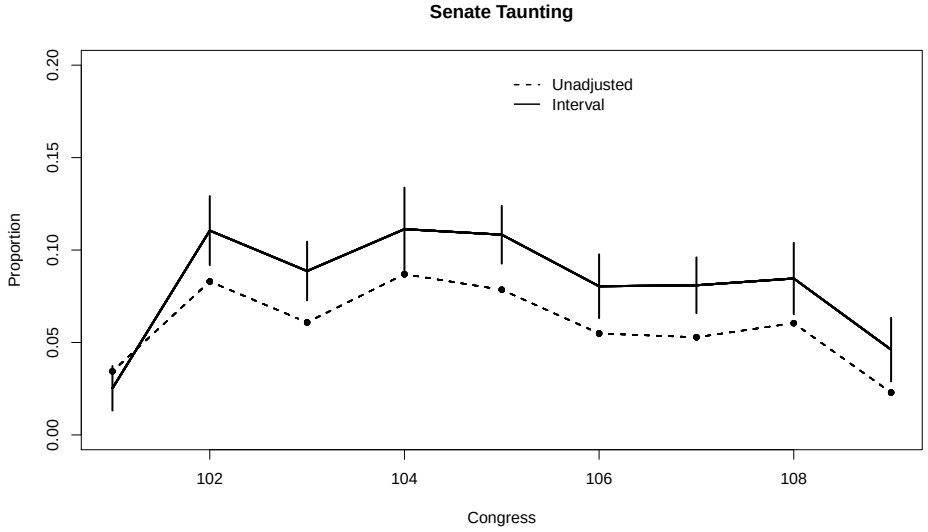


Figure 5 demonstrates that failing to include coder error can lead to a serious underestimation of the taunting rate in the Congress. In the worst cases, the unadjusted taunting rate is more than a full percentage point lower in the *minimum* of the interval from the bounding procedure. Overall, the unadjusted taunting rate underestimates the taunting rate about 0.5 percentage points—or 9 percent underestimate.

Substantively, Figure 5 shows that there has not been a secular increase in taunting in the Senate. Rather, taunting appears highest immediately after the Republican revolution, but subsided somewhat during the Bush administration.

[5]The total number of double-coded speeches is substantially lower than 10% because there were many coders included in the data set and here we focus on only one pair of coders.

8 Conclusion

A Summary of Notation

- Define $\text{mean}_j(a_j) = \sum_{j=1}^{J} a_j / J$ where $J = |\{j : \forall j\}|$.
- indices:

  d ($d = 1, \ldots, D$) document

  s ($s = 1, \ldots, S$) source (of documents)

  k ($k = 1, \ldots, K$) category

  c ($c = 1, \ldots, C$) coder (always as superscript)

- truth:

  $\pi_d \in \{1, \ldots, K\}$ true category for document d (always as subscript)

  $\bar{\pi}_k = \text{mean}_d[I(\pi_d = k)]$ true proportion of documents in category k, with vector $\bar{\boldsymbol{\pi}} = \{\bar{\pi}_k : k = 1, \ldots, K\}$

- measures:

  $y_d^c \in \{1, \ldots, K\}$ coder c's classification of document d

  $\bar{y}_k^c = \text{mean}_d[I(y_d^c = k)]$, proportion of documents classified into category k by coder c, the average over all coders $\bar{y}_k = \text{mean}_c(\bar{y}_k^c)$, and vector $\bar{\boldsymbol{y}} = \{\bar{y}_k : k = 1, \ldots, K\}$

- Validity and Reliability:

  Evaluation Proportion: ϵ_{jk}^c are misclassification proportions, the proportion of times coder c classifies a document into category j among those whose true category is k, with matrix $\boldsymbol{E}^c = \{\epsilon_{jk}^c\}$, the diagonals of which, ϵ_{jj}^c, are the **validities**

  Confusion Matrix element: $m_{jk}^{12} = \text{mean}_d[I(y_d^1 = j, y_d^2 = k)]$, with diagonal elements being the **reliabilities (or agreement proportions)**, $a^{12} = \sum_{k=1}^{K} m_{kk}^{12}$

- Concepts

  Prior, Standard Assumptions: (a) above some (mysterious) threshold of reliability, validity is not effected (i.e., the estimator is not biased); (b) at least one coder is always right

  Constant Category Validity Assumption: $\epsilon_{kk}^c = \epsilon^c$ for all k

  Wisdom of the Coders: Technically, this is a "Maximum Average Coder Validity" assumption, or $\dot{\epsilon}^{12} = \frac{1+a^{12}}{2}$. Equivalently, as we show, this means if a plurality of coders of document d agree, they vote for the truth; if there is a tie, then at least one side votes with the truth.

Coding Generation Process: $\bar{y}_k^c = \sum_{k=1}^K \epsilon_{jk}^c \pi_k$, or in matrix form: $\bar{y}^c = E^c \pi$

Coder Trust Assumption: $\epsilon^1 = \frac{1+a^{12}}{2} + \lambda \frac{1-a^{12}}{2}$, $\epsilon^2 = 1 + a^{12} - \epsilon^1$ for $\lambda \in [-1, 1]$

B Algorithm To Obtain Identification Region

To obtain the intervals using coding decisions and our assumptions, we use a brute force approach to optimization. Specifically, suppose that we observe decisions from coder 1 $\boldsymbol{y}^1$ and coding decisions from coder 2 $\boldsymbol{y}^2$, with agreement rate a^{12}. From Proposition 1 that $\epsilon^1 + \epsilon^2 = 1 + a^{12}$.

Using this information we perform a grid search over the agreement values and the evaluation matrices, while encoding constraints. Our algorithm first searches over the possible accuracy levels, given agreement. For each accuracy level we then perform a grid search over the evaluation matrices consistent with that level of accuracy. By iterating over combinations of accuracy and evaluation matrices we are able to characterize the maximum and minimum possible values.

More specifically, we use the following procedure. Suppose we have a K category measure.

- For each ϵ^1 and ϵ^2 such that $\epsilon^1 + \epsilon^2 = 1 + \mathtt{a}^{12}$:
  - For each column k and each $\boldsymbol{z}_k \in \Delta^{K-1}$:
    - Create matrix E^2 such that:

$$\begin{pmatrix} \epsilon^2 & (1-\epsilon^2)z_{12} & \dots & (1-\epsilon^2)z_{1K} \\ (1-\epsilon^2)z_{21} & \epsilon^2 & \dots & (1-\epsilon^2)z_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ (1-\epsilon^2)z_{K1} & (1-\epsilon^2)z_{K2} & \dots & \epsilon^2 \end{pmatrix} \tag{B.1}$$

    - For each column k and each $\boldsymbol{x}_k \in \Delta^{K-2}$ create matrix E^1 such that

$$\begin{pmatrix} \epsilon^1 & (1-\epsilon^1-\tilde{x}_2)x_{12} & \dots & (1-\epsilon^1-\tilde{x}^K)x_{1K} \\ \tilde{x}_1 & \epsilon^1 & \dots & \tilde{x}_K \\ (1-\epsilon^1-\tilde{x}_1)x_{31}) & \tilde{x}_2 & \dots & (1-\epsilon^1-\tilde{x}^K)x_{3K}) \\ \vdots & \vdots & \ddots & \vdots \\ (1-\epsilon^1-\tilde{x}_1)x_{K1} & (1-\epsilon^1)x_{K2} & \dots & \epsilon^1 \end{pmatrix} \tag{B.2}$$

      where each $\tilde{x}_k$ are unknown values, leaving us with K unknown values, which we collect into $\tilde{\boldsymbol{x}}$
    - Solve for $\tilde{x}_k$ such that

$$(\boldsymbol{E}^1)^{-1}\bar{\boldsymbol{y}}^1 - (\boldsymbol{E}^2)^{-1}\bar{\boldsymbol{y}}^2 \;=\; \mathbf{0} \tag{B.3}$$

      using a Newton-Raphson algorithm. This ensures that the proportions are equal.

- If there is a solution use it to calculate $\boldsymbol{E}^1$ and then if $(\boldsymbol{E}^1)^{-1}\bar{\boldsymbol{y}}^1 \in \Delta^{K-1}$ then update the following for each k

$$\begin{aligned} \min_{\mathtt{k}} &= \min(\min_{\mathtt{k}}, [(\boldsymbol{E}^1)^{-1}\bar{\boldsymbol{y}}^1]_k) \\ \max_{\mathtt{k}} &= \max(\max_{\mathtt{k}}, [(\boldsymbol{E}^1)^{-1}\bar{\boldsymbol{y}}^1]_k) \end{aligned}$$

- For each k return $\min_{\mathtt{k}}$ and $\max_{\mathtt{k}}$

Modifying the algorithm for the cases described above is straightforward. To perform the sensitivity analysis we alter the range of validity that we search over. When one coder is trusted more, we restrict the set of matrices that we search over. When we have multiple coders we restrict the range of joint validity that we search over.

The algorithm becomes increasingly complex as the number of categories increase. But for three and four categories the model is easily fit and extended for all of our cases. In many examples one or two categories are of interest and therefore more complex schemes could be reduced to a 2-3 category scheme.

C A General Approach to Using Proportions as Dependent and Independent Variables

In Section 7 we demonstrated that our approach could be applied to a case where a parameter from a linear regression is the quantity of interest. In this section we provide a straightforward extension of Proposition 6 to any consistent estimator where the estimated proportions may be the dependent or independent variable.

For this general case, suppose that the set of potential covariate values for the sources is $\boldsymbol{\mathcal{X}}$. Suppose we have an estimator $g : \Delta^{K-1} \times \boldsymbol{\mathcal{X}} \to \Re$, $g(\boldsymbol{\pi}, \boldsymbol{X})$ for some quantity of interest ϕ. Note that we are being intentionally ambiguous about whether $\boldsymbol{\pi}$ is the independent or dependent variable. We will suppose that $g(\boldsymbol{\pi}, \boldsymbol{X})$ is a consistent estimator of ϕ, so that $g(\boldsymbol{\pi}, \boldsymbol{aX}) \to^{\text{plim}} \phi$. As in Section 7, we also restrict attention to only those evaluation matrices that, when inverted and applied to the hand coded estimate, returns estimates in the simplex and where the inverted evaluation matrices, applied to the coders' estimates, provide the value. As before, define this set of pairs of matrices as $\mathbb{E}$.

Then, we can define an interval estimator for the quantity of interest ϕ^{int} as,

$$\phi^{\text{int}} = \left[\min_{(\boldsymbol{E}^1, \boldsymbol{E}^2) \in \mathbb{E}} g((\boldsymbol{E}^1)^{-1}\bar{\boldsymbol{y}}^1, \boldsymbol{X}), \max_{(\boldsymbol{E}^1, \boldsymbol{E}^2) \in \mathbb{E}} g((\boldsymbol{E}^1)^{-1}\bar{\boldsymbol{y}}^1, \boldsymbol{X}) \right] \tag{C.1}$$

Proposition 7 shows that Equation C.1 will contain the true value of ϕ.

Proposition 7. *Suppose that $g(\boldsymbol{\pi}, \boldsymbol{X})$ is a consistent estimator for ϕ, that coder 1 and coder 2 have maximum joint validity, that the coders' errors are source independent, and that the coders have constant validity across categories. Then $\phi \in \phi^{int}$.*

Proof. The argument parallels the proof of Proposition 7. First note that the true evaluation matrices $\boldsymbol{E}^1, \boldsymbol{E}^2 \in \mathbb{E}$ by Assumptions 1, 2, and 3. This implies that

$\left(\boldsymbol{E}^1\right)^{-1}\bar{\boldsymbol{y}}^c = \left(\boldsymbol{E}^2\right)^{-1}\bar{\boldsymbol{y}}^c = \boldsymbol{\pi}$ and thus $g(\left(\boldsymbol{E}^1\right)^{-1}\bar{\boldsymbol{y}}^c, \boldsymbol{X}) = \phi$ and $g(\left(\boldsymbol{E}^2\right)^{-1}\bar{\boldsymbol{y}}^2, \boldsymbol{X}) = \phi$. Therefore $\phi \in \phi^{\text{int}}$ □

References

Béchard, Jean-Pierre and Denis Grégoire. 2005. "Entrepreneurship education research revisited: The case of higher education." *Academy of Management Learning & Education* 4(1):22–43.

Bortree, Denise Sevick and Trent Seltzer. 2009. "Dialogic strategies and outcomes: An analysis of environmental advocacy groups Facebook profiles." *Public Relations Review* 35(3):317–319.

Byrt, Ted, Janet Bishop and John B Carlin. 1993. "Bias, prevalence and kappa." *Journal of clinical epidemiology* 46(5):423–429.

Cronbach, Lee J. 1951. "Coefficient alpha and the internal structure of tests." *psychometrika* 16(3):297–334.

De Vreese, Claes H, Susan A Banducci, Holli A Semetko and Hajo G Boomgaarden. 2006. "The news coverage of the 2004 European Parliamentary election campaign in 25 countries." *European Union Politics* 7(4):477–504.

Di Eugenio, Barbara and Michael Glass. 2004. "The kappa statistic: A second look." *Computational linguistics* 30(1):95–101.

Druckman, James N, Martin J Kifer and Michael Parkin. 2009. "Campaign communications in US congressional elections." *American Political Science Review* 103(03):343–366.

Druckman, James N, Martin J Kifer and Michael Parkin. 2010. "Timeless strategy meets new medium: Going negative on congressional campaign web sites, 2002–2006." *Political Communication* 27(1):88–103.

Druckman, James N and Michael Parkin. 2005. "The impact of media bias: How editorial slant affects voters." *Journal of Politics* 67(4):1030–1049.

Grimmer, Justin, Sean Westwood and Solomon Messing. 2014. *The Impression of Influence: Legislator Communication, Representation, and Democratic Accountability*. Princeton University Press.

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

Hripcsak, George and Daniel F Heitjan. 2002. "Measuring agreement in medical informatics reliability studies." *Journal of biomedical informatics* 35(2):99–110.

Jamal, Amaney, Robert O Keohane, David Romney and Dustin Tingley. 2014. "Anti-Americanism or anti-interventionism? Evidence from the Arabic Twitter universe." *Perspectives on Politics nd. Forthcoming* .

Kuha, Jouni and Chris Skinner. 1997. *Survey measurement and process quality*. John Wiley & Sons chapter Categorical data analysis and misclassification.

Leccese, Mark. 2009. "Online information sources of political blogs." *Journalism & Mass Communication Quarterly* 86(3):578–593.

Lombard, Matthew, Jennifer Snyder-Duch and C Campanella Bracken. 2002. "Content analysis in mass communication." *Human communication research* 28(4):587–604.

McCall, Leslie. 2013. *The Underserving Rich: American Beliefs about Inequality, Opportunity, and Redistribution*. Cambridge University Press.

Mikhaylov, Slava, Michael Laver and Kenneth Benoit. 2012. "Coder Reliability and Misclassification in the Human Coding of Party Manifestos." *Political Analysis* 20(1):78–91.

Molinari, Francesca. 2008. "Partial Identification of Probability Distributions with Misclassified Data." *Journal of Econometrics* 144(1):81–117.

Nazli Nik Ahmad, Nik and Maliah Sulaiman. 2004. "Environment disclosure in Malaysia annual reports: a legitimacy theory perspective." *International Journal of Commerce and Management* 14(1):44–58.

Riffe, Daniel and Alan Freitag. 1997. "A content analysis of content analyses: Twenty-five years of Journalism Quarterly." *Journalism & Mass Communication Quarterly* 74(3):515–524.

Scott, William A. 1955. "Reliability of content analysis: The case of nominal scale coding." *Public opinion quarterly* .

Socher, Richard et al. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Stewart, Brandon M and Yuri M Zhukov. 2009. "Use of force and civil–military relations in Russia: an automated content analysis." *Small Wars & Insurgencies* 20(2):319–343.

Tai, Kai Sheng, Richard Socher and Christopher Manning. 2015. Improved Semantic Representations From Tree-Structured Long Short-Term Memory Networks. In *ACL*.

Van Gorp, Baldwin. 2005. "Where is the frame? Victims and intruders in the Belgian press coverage of the asylum issue." *European Journal of Communication* 20(4):484–507.

Wacholder, Sholom, Ben Armstrong and Patricia Hartge. 1993. "Validation Studies Using an Alloyed Gold Standard." *American Journal of Epidemiology* 137(11):1251–1258.

Williams, Dmitri, Nicole Martins, Mia Consalvo and James D Ivory. 2009. "The virtual census: Representations of gender, race and age in video games." *New Media & Society* 11(5):815–834.

Zullow, Harold M, Gabriele Oettingen, Christopher Peterson and Martin E Seligman. 1988. "Pessimistic explanatory style in the historical record: CAVing LBJ, presidential candidates, and East versus West Berlin." *American Psychologist* 43(9):673.