COMMUNICATION METHODS AND MEASURES
2017, VOL. 11, NO. 2, 87–104
http://dx.doi.org/10.1080/19312458.2017.1305103

Routledge
Taylor & Francis Group

ARTICLES

# Correcting Measurement Error in Content Analysis

Marko Bachl[a] and Michael Scharkow[b]

[a]Department of Communication, University of Hohenheim, Stuttgart, Germany; [b]Department of Culture & Communication, Zeppelin University, Friedrichshafen, Germany

**ABSTRACT**

Conducting and reporting reliability tests has become a standard practice in content analytical research. However, the consequences of measurement error in coding data are rarely discussed or taken into consideration in subsequent analyses. In this article, we demonstrate how misclassification in content analysis leads to biased estimates and introduce matrix back-calculation as a simple remedy. Using Monte Carlo simulation, we investigate how different ways of collecting information about the misclassification process influence the effectiveness of error correction under varying conditions. The results show that error correction with an adequate set-up can often substantially reduce bias. We conclude with an illustrative example, extensions of the procedure, and some recommendations.

It is a truism in any empirical research that data are rarely perfectly reliable and accurate: Respondents do not always answer the same questions in the same fashion, observers fail to recognize certain events, measurement instruments work with varying accuracy. Content analysis is especially sensitive to measurement error, because data generation relies on the "consensual reading" (Krippendorff, 2004a, p. 212) of semantically or visually ambiguous messages by different coders and/or at different occasions. Content analysts have been aware of this challenge for decades, and many textbooks contain chapters on ensuring and testing coding quality (Krippendorff, 2004a; Neuendorf, 2002). Reporting of intercoder reliability tests is encouraged or even required for publication in many journals and has become a standard practice for most communication researchers (Lombard, Snyder-Duch, & Bracken, 2002; Lovejoy, Watson, Lacy, & Riffe, 2014; Riffe & Freitag, 1997). *Communication Methods and Measures* in particular has served as an important outlet for this laudable development (Hayes & Krippendorff, 2007; Krippendorff, 2008, 2011; Krippendorff & Craggs, 2016; Lovejoy et al., 2014).

The current best practice to deal with measurement error in content analysis is to improve the codebook instructions and train the coders until intercoder reliability meets an acceptable threshold in a test with an adequate sample of coding units. This practice establishes the *reproducibility* of the coding process: The chance-corrected intercoder reliability coefficient indicates how likely the data is to replicate with the same instrument but employing other coders (Krippendorff, 2004a). This is an important quality indicator in its own right and the methods introduced in this article do not relieve researchers from establishing reproducibility. However, after concluding that intercoder reliability is sufficient, usually the data are analyzed, the results of the study are reported, and conclusions are drawn *as if no measurement error had occurred*. This assumption is almost always false and can easily lead to erroneous conclusions. Intercoder reliability is only a necessary, yet not a sufficient condition for accuracy, i.e., that the

**CONTACT** Marko Bachl marko.bachl@uni-hohenheim.de Department of Communication, University of Hohenheim, Fruwirthstraße 46, Stuttgart 70599, Germany.

coded categories reflect the "true" values.[1] More precisely, the observed agreement among coders indicates merely the upper bound of the accuracy of a coding process. If, for example, two coders agree in 70% of their codings, a maximum of 85% of their codings can be correct.[2] Consequently, since the necessary condition of perfect intercoder reliability is almost never fulfilled, coding processes are very rarely perfectly accurate.

Disregarding measurement error in data analysis is disadvantageous in principle (Buonaccorsi, 2010), and at least two arguments come to mind why it can be particularly problematic in content analysis. First, because intercoder reliability is often disclosed, we know for a fact that even under the most optimistic assumptions measurement error of at least modest magnitude must be present in many studies. It seems paradoxical on the one hand to collect and report such information, but, on the other hand, not to use it. Second, measurement error of categorical variables—also known as misclassification—can lead to bias of substantial magnitude and in unpredictable direction even for simple univariate quantities, such as proportions. This property is well known in the methodological literature (e.g., Bross, 1954; Kuha & Skinner, 1997; Marshall, 1990; Schwartz, 1985), and several corrections to mitigate the bias exist. Many, if not most of the variables in content analyses are categorical and are therefore affected by the consequences of misclassification. To the best of our knowledge, however, the problem and possible solutions have not yet been discussed with regard to the specific characteristics of content analysis in communication research.

We will do just that. In the first part of the article, we introduce the concept of misclassification and a notation for its description. We show how misclassification leads to biased estimates even of simple quantities such as proportions, and we demonstrate why this bias can rarely be neglected. We then present matrix back-calculation as a simple solution for reducing bias in proportion estimates. The second part of the article is concerned with the implementation of the back-calculation procedure in content analytical studies. Since the correction requires information about the misclassification process, we present three approaches to gather such information similar to a reliability test and discuss the assumptions behind the approaches. In the third part, we investigate the utility and relative effectiveness of the approaches with a Monte Carlo simulation. We conclude with an illustrative example, extensions of the procedure, and some recommendations.

## Misclassification and its consequences in content analysis

### *Binary variable, equal error rates*

Consider a coding process of a binary variable that records whether the topic of a news story falls into the domain of politics. The most simple model of misclassification, the "equal error rate model" (EERM) (Schwartz, 1985, p. 441), assumes that it is equally difficult to correctly identify news items that do and do not cover politics. The EERM has only one parameter: $\theta$ is the probability of a correct classification, or accuracy, and, consequently, $1-\theta$ is the probability of an error.[3] The expected *observed* (or coded) proportion of politics in the news $p^*_{\text{politics}}$ is the sum of the proportion of correctly identified political stories and the proportion of misclassified non-political stories. Let us assume that a proportion of $p_{\text{politics}} = .3$ of all news items truly cover politics, and news stories are classified correctly with an accuracy of $\theta = .8$. Under these assumptions, we expect to observe

$$p^*_{\text{politics}} = \theta * p_{\text{politics}} + (1-\theta) * \left(1 - p_{\text{politics}}\right) = .8 * .3 + .2 * .7 = .38. \quad (1)$$

[1]We will refrain from joining the philosophical debate whether the "true" value of a variable for a coding unit exists and, if so, can be determined without error. Similarly, we note that even the most accurate coding process does not necessarily guarantee validity (see Krippendorff, 2004a, Ch. 13, for a detailed discussion). Instead, we take a pragmatic stand on the issue. The true value is defined as the category that has to be coded following the researchers who have developed the measurement instrument.

[2]See the subsequent section on the *maximum possible accuracy assumption* for details.

[3]We follow the notation of Kuha and Skinner (1997) throughout the article.

Several observations about Equation (1) are noteworthy: *First*, the observed proportion $p^*_{\text{politics}}$ is biased substantially despite a seemingly acceptable error rate of .2. We would report over one quarter more political stories than there actually are. Second, the bias occurs despite the EERM, which is the closest conceptual equivalent to random normal measurement error of continuous variables. So even if we assume that coders randomly confuse political and non-political stories, the errors do not cancel each other out. This happens because, third, the bias is not only a function of the accuracy $\theta$, but also of the true proportion $p_{\text{politics}}$. If we define bias as the difference of observed and true proportions, $p^*_{\text{politics}} - p_{\text{politics}}$, put in Equation 1 for $p^*_{\text{politics}}$, and rearrange to

$$p^*_{\text{politics}} - p_{\text{politics}} = (\theta - 1) * \left(2 * p_{\text{politics}} - 1\right) = .08, \tag{2}$$

it becomes clear that the magnitude of bias depends on $\theta$, and that the magnitude and the direction of bias depends on $p_{\text{politics}}$. The absolute bias increases with decreasing ability of the coders to assign the correct category. The absolute bias also becomes stronger when the true proportion of political stories deviates farther from $p_{\text{politics}} = .5$, and the direction of bias changes dependent on whether the true proportion is above or below this threshold. This property has substantive consequences: Under the EERM and assuming the same accuracy, a researcher who investigates the occurrence of political news in a population in which they are quite rare (say, MTV news), will quite likely get an overestimate. In contrast, a study of CNN news might underestimate the share of political stories.

Finally fourth, Equation (2) also reveals that there are only two conditions in which there would be no bias. If $\theta = 1$, no misclassification occurs, and all estimates of $p^*_{\text{politics}}$ are of course unbiased. If exactly half of the news items cover politics, $p_{\text{politics}} = .5$, the expected observed proportion is unbiased, $p^*_{\text{politics}} = .5$, regardless of the coding accuracy, under the equal error rate model. Consequently, a researcher in our example would have to make two interdependent assumptions above the EERM to conclude that the consequences of misclassification are negligible. He or she must assume that the coding is sufficiently accurate given the deviation of the true proportion from .5. If political news are very rare or very frequent, high accuracy is necessary to prevent strong bias. However, because the true occurrence is generally unknown (if we already knew it, there would be no need for the study), this assumption is rarely, if ever, defensible.

In summary, we conclude from the four observations that, even if the EERM held, it is almost never defensible to assume that misclassification can be ignored. Doing so will severely bias estimates of simple univariate quantities and consequently invalidate the conclusions drawn from the empirical results.

### *Binary variable, unequal error rates*

The EERM can be relaxed to allow categories of unequal difficulty and, for multicategorical variables, systematic confusion of certain categories. Unequal difficulty for different categories is often a consequence of coding instructions. For example, a common strategy to improve the precision of codebook instructions is to list several indicators for a category and advise the coders only to assign a category if at least one indicator is clearly present. In the politics example, such instructions might lead to a high probability of $\theta_{\text{non-p.}|\text{non-p.}} = .9$ for correctly identifying a non-political story as such. In return, the probability of correctly recognizing political stories will often be lower, say $\theta_{\text{pol.}|\text{pol.}} = .7$, because the coders look carefully for a limited set of indicators and will miss some political stories that do not contain any of them. Again assuming a proportion of $p_{\text{politics}} = .3$ and adapting Equation (1), we expect to observe

$$\begin{aligned} p^*_{\text{politics}} &= \theta_{\text{pol.}|\text{pol.}} * p_{\text{politics}} + \left(1 - \theta_{\text{non-p.}|\text{non-p.}}\right) * \left(1 - p_{\text{politics}}\right) = \\ &= .7 * .3 + .1 * .7 = .28. \end{aligned} \tag{3}$$

We note that the bias as a function of true proportions and classification probabilities turns out to be smaller in this specific setting. However, if we assume the same coding process but the inverse

true proportion $p_{\text{politics}} = .7$, we are expected to observe $p^*_{\text{politics}} = .52$, which is a substantial underestimate. As a last example, we plug an equal distribution of $p_{\text{politics}} = .5$ in Equation (3), which yields an expected observed proportion of $p^*_{\text{politics}} = .4$. The observed proportion for an equal distribution of the true values is no longer unbiased under the unequal error rate model. It follows from Equation (3) that the observed proportion of political stories in the presence of misclassification for both categories (i.e., $\theta_{\text{pol.}|\text{pol.}} < 1$ and $\theta_{\text{non-p.}|\text{non-p.}} < 1$) is unbiased if and only if

$$\frac{p_{\text{politics}}}{1 - p_{\text{politics}}} = \frac{1 - \theta_{\text{non-p.}|\text{non-p.}}}{1 - \theta_{\text{pol.}|\text{pol.}}}. \tag{4}$$

Assuming the error rates from our example, no bias occurs for $p_{\text{politics}} = .25$. In practice, a researcher will never know the true occurrence of political news, so this equilibrium of misclassification is not only unlikely, but also unknowable.

Misclassification can be described more concisely in matrix notation. The misclassification matrix (MCM) for the coding process of a binary variable is a $2 \times 2$ matrix that contains the conditional probabilities that a category is coded given the true value of the coding unit. By convention, the true values are presented in the columns, and the probabilities in each column sum to 1. The MCM for the politics example is

$$\Theta_{\text{politics}} = \begin{pmatrix} \theta_{\text{pol.}|\text{pol.}} & \theta_{\text{pol.}|\text{non-p.}} \\ \theta_{\text{non-p.}|\text{pol.}} & \theta_{\text{non-p.}|\text{non-p.}} \end{pmatrix} = \begin{pmatrix} .7 & .1 \\ .3 & .9 \end{pmatrix}, \tag{5}$$

where the first column contains the probabilities of assigning the politics category, $\theta_{\text{pol.}|\text{pol.}}$, or the non-politics category, $\theta_{\text{non-p.}|\text{pol.}}$, respectively, if the news item truly covers politics. The diagonal entries are the probabilities of correct classifications, and the off-diagonal contains the error rates.

Matrix notation also simplifies the equations to compute the expected observed proportions. Equation (3) translates to

$$P^*_{\text{politics}} = \Theta_{\text{politics}} \times P_{\text{politics}} = \begin{pmatrix} .7 & .1 \\ .3 & .9 \end{pmatrix} \times \begin{pmatrix} .3 \\ .7 \end{pmatrix} = \begin{pmatrix} .28 \\ .72 \end{pmatrix}, \tag{6}$$

where $P^*_{\text{politics}}$ is the vector of the observed proportions of political and non-political stories, and $P_{\text{politics}}$ is the vector of the true proportions.

## Multicategorical variables

The previous sections generalize directly to multicategorical variables. We thus limit the following section to a general introduction of the notation. Let $A$ be a multicategorical variable with $k$ categories $1, \ldots, k$. The MCM $\Theta_A$ is a $k \times k$ matrix,

$$\Theta_A = \begin{pmatrix} \theta_{1|1} & \theta_{1|2} & \theta_{1|\ldots} & \theta_{1|k} \\ \theta_{2|1} & \theta_{2|2} & \theta_{2|\ldots} & \theta_{2|k} \\ \theta_{\ldots|1} & \theta_{\ldots|2} & \theta_{\ldots|\ldots} & \theta_{\ldots|k} \\ \theta_{k|1} & \theta_{k|2} & \theta_{k|\ldots} & \theta_{k|k} \end{pmatrix}. \tag{7}$$

The most simple MCM is the equal difficulties, equal error rates model (EDEERM). Similar to the EERM for binary variables, all diagonal entries contain the same probability, $\theta_{k|k}$, and all off-diagonal entries contain $\left(1 - \theta_{k|k}\right)/(k-1)$. Under the EDEERM, all observed proportions will be biased towards a value of $1/k$, and the bias will be stronger for proportions which deviate farther from $1/k$. More complex error structures are possible (and likely to occur in practice). The only constraint is that the probabilities in each column must sum to 1. The stronger the true misclassification process deviates from the simple EDEERM, the less predictable the bias is in direction and magnitude. The expected observed proportions are again calculated using the matrix notation in Equation (6), where

$\Theta_A$ is the $k \times k$ misclassification matrix, and $P_A$ and $P_A^*$ are vectors of length $k$ which contain the true and the misclassified category proportions.

### *Correcting misclassified proportions via matrix back-calculation*

A simple solution to correct for the consequences of misclassification exists if the misclassification process is known. The true proportions can simply be back-calculated from the observed proportions and the misclassification probabilities. As a first example, we solve Equation (1) from the politics example for $p_{\text{politics}}$ to back-calculate to the true proportion of political stories:

$$p_{\text{politics}} = \frac{p^*_{\text{politics}} + \theta - 1}{2 * \theta - 1} = \frac{.38 + .8 - 1}{2 * .8 - 1} = .3. \tag{8}$$

It becomes clear from the denominator that the equation only has a solution if $\theta \neq .5$. This intuitively makes sense, because an error rate of $\theta = .5$ under the EERM for a binary variable basically implies that the coders just flip a coin to determine whether a news item covers politics. The observed proportion of political stories is then merely the outcome of an unconditional random process without reference to the true values, which consequently cannot be recovered by back-calculation.

Back-calculation for unequal error rates and multicategorical variables can be expressed more compact in matrix notation. We thus refer to the procedure as *matrix back-calculation*. Solving Equation (6) for $P_{\text{politics}}$ yields

$$P_{\text{politics}} = \Theta^{-1}_{\text{politics}} \times P^*_{\text{politics}} = \begin{pmatrix} .7 & .1 \\ .3 & .9 \end{pmatrix}^{-1} \times \begin{pmatrix} .28 \\ .72 \end{pmatrix} = \begin{pmatrix} 1.5 & -\frac{1}{6} \\ -0.5 & \frac{7}{6} \end{pmatrix} \times \begin{pmatrix} .28 \\ .72 \end{pmatrix} = \begin{pmatrix} .3 \\ .7 \end{pmatrix}, \tag{9}$$

where $\Theta^{-1}_{\text{politics}}$ is the inverse of the misclassification matrix. As for Equation (8), a solution does not exist in the case of unconditional random misclassification, i.e., all entries of the misclassification matrix are $1/k$.

Figure 1 concisely summarizes the key points from the first section. The data come from the simulation which is described below. For now, we omit the technical details and focus on its illustrative purpose. The continuous lines show the overall deviation of the uncorrected observed proportions from the true proportions, quantified by the Root Mean Squared Error (RMSE) of all category proportions. The underlying misclassification process assumes unequal difficulties and

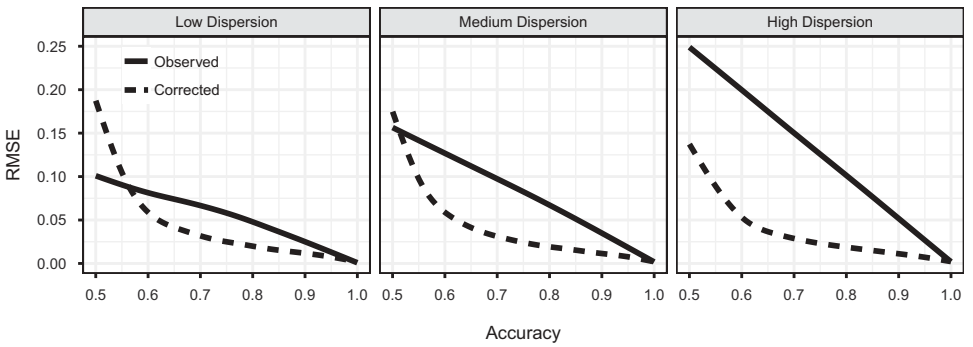


**Figure 1.** Bias of the observed proportions and performance of the correction.

*Notes*. The figure shows the overall deviation from the true proportions, quantified by the Root Mean Squared Error (RMSE) of all categories, on the y axis as function of the true accuracy of the coding process on the x axis. The facets distinguish between low, medium, and high dispersion, where dispersion is quantified by the deviation from an equal distribution. The figure is based on simulated data under the condition of fairly accurate knowledge of the MCM from extensive preferred standard tests with 8 coders and 100 test units. Details on the simulation are documented in the Methods section below.

error rates. Three patterns become obvious: (1) less accuracy leads to stronger bias; (2) more dispersed true proportions lead to stronger bias; and (3) accuracy and true distribution interact, such that the bias induced by a similar level of inaccuracy is stronger for more dispersed true proportions. The broken lines show the results from the correction procedure with fairly accurate knowledge of the MCM. It is evident that the correction has the potential to mitigate the bias substantially if there is sufficient information about the misclassification. In the next section, we show how such information can be obtained.

## Approximations of the misclassification matrix

### *Approximation with test-standard designs*

#### *Preferred standard*

The most straightforward way to approximate the MCM is a comparison of the misclassified categories with an external standard which is assumed to represent the correct values. These values are measured with a "preferred procedure" (Kuha & Skinner, 1997, p. 646) which is of high quality, but too expensive or otherwise infeasible to implement in the whole study. In the biomedical sciences, preferred standards are, for example, collected from medical records (e.g., for medicine use, Marshall, 1990) or with diagnostic tools (e.g., a biochemical carbon monoxide test for smoking behavior, Bauman & Koch, 1983).

In content analytical studies, the test of accuracy, which compares the classifications of the coders to a standard, provides a direct estimate of the MCM. Krippendorff (2004a, p. 216) suggests that classifications of "acknowledged experts" or of "panels of experienced content analysts" can be used as preferred standard.[4] This approach is, for example, utilized in the ongoing Manifesto Project (Volkens, 2007) and in an external re-analysis of its data quality (Mikhaylov, Laver, & Benoit, 2012). Another example of a media content analysis is provided by Schmid-Petri and colleagues (2015). It should be noted that the approach is not undisputed. Krippendorff (2004a) warns that preferring one specific interpretation of a coding unit over others may be epistemologically problematic, particularly for complex theoretical constructs.

While we agree with the philosophical argument, we still think that it is reasonable to correct according to preferred codings. They are selected to reflect the beliefs of the researchers, who, in the end, will analyze the collected data and draw substantive conclusions. This does of course not mean that the preferred standard always identifies the truth—this is a question of validity, not of accuracy (Krippendorff, 2004a, Ch. 13). Correcting with reference to such a preferred standard makes it at least more likely that the data agree with the researchers' interpretations. Consider a common example: A researcher wants to quantify the share of political coverage in the evening news, because she wants to test the hypothesis that political news have become less important over time. She has a very strict understanding of political news, such that only news which cover policy, politics, and polity qualify as political. The coders, however, may not be as politically educated as the researcher and implicitly use a more simple rule: a news story is political if any politician is mentioned. Now consider that the evening news recently covered politicians more often in non-political contexts (for example, the new dog of the president), but report less on policy issues. According to the preferred understanding of the researcher, the share of political news became in fact lower. But according to the simple heuristic of the coders, the share might be stable over time, because they code politicians in non-political contexts as political. Now consider that all coders apply the "politician heuristic" equally and a intercoder reliability test therefore shows good results. The researcher would—

[4]An anonymous reviewer was more pessimistic about the availability of preferred standards: "In practice, accuracy is a form of reliability that is rarely measurable because standards are hard to come by." We agree that standards in the sense of a truth criterion are in fact rare. We argue, however, that a preferred standard in the sense of a correct application of the measurement instructions can, and should, be established. This is implicitly done during coder training, when the researchers explain the instrument to the coders. It is then only a small step to explicitly establish a preferred standard by classifying a set of test units.

following the current practice—accept the data as reliable and reject her hypothesis. Only if she compares some test codings against her preferred standard, she will be able to detect (and subsequently correct) the coders' bias.

### *Majority standard*

If the collection of a preferred standard is infeasible or rejected because of epistemological arguments, a standard can be established by surveying multiple codings of a coding unit for the most frequent category. Put simply, if a majority of independent coders agree on the classification of a coding unit, then their category is treated as the standard value for this coding unit. In the case of a tie, the standard category is randomly selected from the most frequent categories. The majority standard removes the need for an external preferred standard, but it comes with a necessary precondition and an additional assumption. Establishing a majority classification of course requires at least three coders for each coding unit of the test-standard comparison. Moreover, the procedure assumes that the majority of coders always recognizes the correct category. The likelihood of the assumption being fulfilled depends on the number of categories, the number of multiple codings per unit, and, most importantly, on the accuracy itself.

Figure 2 illustrates these dependencies under the EDEERM. The majority decision works better with more coders, more categories, and more accurate coders. The last-mentioned influence can turn out to be problematic in the context of error correction: incorrect majority decisions introduce noise to the standard, consequently to the estimate of the MCM, and finally to the corrected estimates. As we have shown above, less accurate codings lead to more biased proportion estimates and a greater need for correction. Because they also lead to worse estimates of the MCM, a correction based on the majority standard can be expected to perform worse with lower accuracy. Finally, we note that the relationships with the quality of the majority standard are less straightforward if the coding process deviates from the EDEERM. The more the difficulties of the categories differ, and the more systematic some categories are confused with others, the less predictable is the quality of the majority

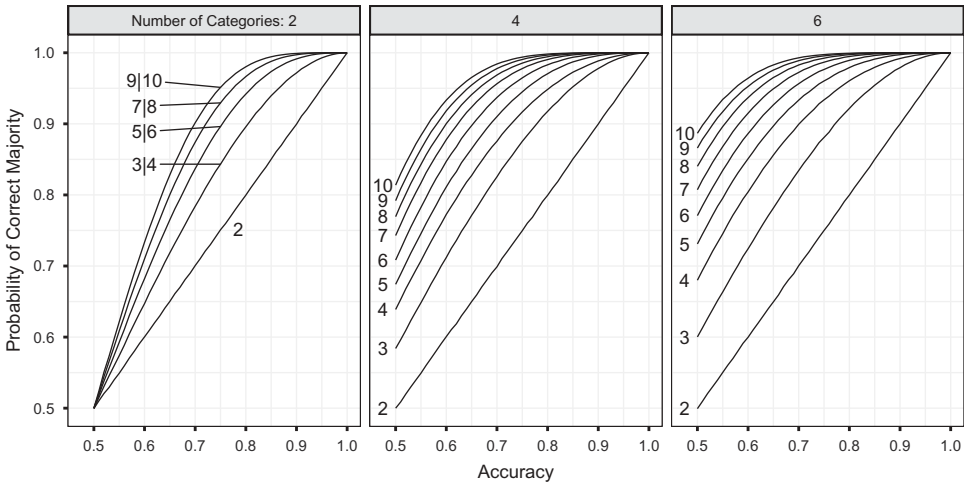


**Figure 2.** Probability of the coder majority identifying the correct category.

*Notes*. The figure shows the probability that the majority of coders identifies the correct category of any given coding unit under the equal difficulty, equal error rates model for all coders. The x axis shows the true accuracy of the single coders. The facets present the relationship for variables with 2, 4, or 6 categories. The lines represent the number of coders. For a variable with two categories, the probabilities for three and four, five and six, …, coders are identical.

standard. These complexities have to be considered if majority decisions are used to establish the standard categories.

We want to acknowledge the critical opinion of an anonymous reviewer on the theoretical suitability of the majority standard. In the reviewer's words: "Although observing clear majorities of matching categories undoubtedly increases content analysts' confidence in trusting their data, correcting data by majority vote fails to improve the actual reliability of the corrected data. Without a standard, there is no way of knowing whether a majority of coders is correct." We agree that the assumption of correct majority decisions is untestable without an external standard (if an external standard was available, there would be no need for a majority standard), and that the assumption is often not perfectly met. With many coders and a reasonably optimistic assumption on their individual accuracy, however, the majority is able to quite reliably select the correct category (see Figure 2). This is why crowd-sourcing approaches to content analysis can produce data of acceptable quality from multiple codings per unit by minimally trained coders (Benoit, Conway, Lauderdale, Laver, & Mikhaylov, 2016). In the end, whether or not a researcher is willing to trust a majority standard depends on his or her assumptions about the accuracy of the single coders.

### *Estimation of the MCM*

After a preferred or majority standard is established, the test units of each coder are cross-tabulated with the standard, so that the standard categories are in the columns, and divided by the column marginal frequencies. Following the notation of Equation (7), the estimated MCM of a variable $A$ with $k$ categories for any coder $m$ is (omitting the $m$ index of the single proportions for better readability)

$$\hat{\Theta}_A^m = \begin{pmatrix} \hat{\theta}_{1|1} & \hat{\theta}_{1|2} & \hat{\theta}_{1|\ldots} & \hat{\theta}_{1|k} \\ \hat{\theta}_{2|1} & \hat{\theta}_{2|2} & \hat{\theta}_{2|\ldots} & \hat{\theta}_{2|k} \\ \hat{\theta}_{\ldots|1} & \hat{\theta}_{\ldots|2} & \hat{\theta}_{\ldots|\ldots} & \hat{\theta}_{\ldots|k} \\ \hat{\theta}_{k|1} & \hat{\theta}_{k|2} & \hat{\theta}_{k|\ldots} & \hat{\theta}_{k|k} \end{pmatrix}, \tag{10}$$

where $\hat{\theta}_{1|1}$ is the proportion of test units with the standard category 1 which coder $m$ classified into category 1, $\hat{\theta}_{2|1}$ is the proportion of test units with the standard category 1 which coder $m$ classified into category 2, $\hat{\theta}_{1|2}$ is the proportion of test units with the standard category 2 which coder $m$ classified into category 1, and so on.

Each coder's individual MCM could then, in principle, be used to correct the estimates from his or her codings. However, as the simulation below will show, a sufficiently precise estimate of the MCM requires quite a large number of test-standard comparisons. In practice, most content analytical studies will not be able to afford such extensive tests. The assumption of interchangeable coders allows us to estimate the MCM more efficiently. In order to be reproducible, content analytical data should be created by a coding procedure that creates the same results, "regardless of who enacts that procedure" (Krippendorff, 2004b, p. 414). According to this assumption, all coders' individual MCMs are probabilistic deviations from the true misclassification process. The individual matrices can thus be averaged into a more precise approximation of the MCM of the whole coding process,

$$\hat{\Theta}_A = \frac{1}{m} \sum^{m} \hat{\Theta}_A^m, \tag{11}$$

assuming an equal number of test-standard comparisons of each coder for simplicity of notation.

### Approximation with the maximum possible accuracy

In some situations, neither a preferred nor a majority standard may be available. This includes, for example, studies with only two coders and no preferred standard, or secondary analyses for which no test data, but information on intercoder agreement can be retrieved. In other studies, the scale of the coder test is just too small to obtain a sufficiently precise estimate of the full MCM, but it may still provide some information on the overall accuracy. The lack of empirical information is compensated by two additional assumptions. First, we assume that the EDEERM applies. Second, we assume that the coding process produces data of the maximum possible accuracy.

It is well known that the agreement between coders is not directly related to the accuracy of a coding process, because without a known standard we cannot distinguish whether the coders agree on the correct or on an incorrect category for any given coding unit (Krippendorff, 2004b). However, it is possible to derive the upper bound for the accuracy of a coding process that could have produced the observed intercoder agreement (Grimmer, King, & Superti, 2015). Consider a coding process with $m = 2$ coders who agree in their classification of 80% of the coding units in a intercoder test. The most optimistic assumption is that at least one coder assigns the correct category to each coding unit. Consequently, all agreements must be correct (80%), as must be half of the disagreements (10%). More generally, the maximum possible accuracy for $m = 2$ coders given the intercoder agreement is obtained by

$$\hat{\theta}_{k|k}^{\max} = \frac{1 + \text{Agreement}}{2} = \frac{1 + .8}{2} = .9. \tag{12}$$

The generalization to $m > 2$ coders is straightforward and resembles the idea of the majority standard in the previous section. The maximum possible accuracy for a set of test units which are each classified by multiple coders is the number of codings that agree with a majority code of the respective test unit, divided by the total number of codings. Table 1 and Equation (13) illustrate the logic for a simple example with $n_{\text{Test}} = 3$ test units and $m = 3$ coders. The absolute majority category "A" must be assumed to be correct for test units 1 and 2 in order to maximize the accuracy. Only one coding of test unit 3 can be correct. The maximum possible accuracy for the coding process of the three coders given the observed intercoder agreement data in Table 1 is

$$\hat{\theta}_{k|k}^{\max} = \frac{\text{Number of codes with one majority category}}{\text{Total number of codings}} = \frac{3 + 2 + 1}{9} = .67. \tag{13}$$

The estimated MCM for a variable $A$ with $k$ categories, $\hat{\Theta}_A$, contains $\hat{\theta}_{k|k}^{\max}$ in all diagonal cells and $\left(1 - \hat{\theta}_{k|k}^{\max}\right)/(k-1)$ in all off-diagonal cells, under the EDEERM.

Both assumptions which are required for this approximation will most likely not be perfectly met in practice. Particularly the assumption of the maximum possible accuracy might be contestable as overly optimistic. It is, nevertheless, considerably more realistic than the current practice of first acknowledging that intercoder agreement was not perfect, but then proceeding as if accuracy was perfect. In contrast to ignoring disagreement, the maximum possible accuracy formalizes the most optimistic upper bound based on the available empirical information.

**Table 1.** Maximum possible number of accurate codings.

| Test Unit | Coder 1 | Coder 2 | Coder 3 | Majority Code | Max. # Accurate |
|---|---|---|---|---|---|
| 1 | A | A | A | A | 3 |
| 2 | A | A | B | A | 2 |
| 3 | A | B | C | A\|B\|C | 1 |

### Correcting proportion estimates

After an estimate of the MCM has been obtained by means of a preferred or majority standard comparison or by using the maximum possible accuracy assumption, the correction of the category proportions from the misclassified variable is straightforward. The inverse of the estimated MCM, $\hat{\Theta}_A^{-1}$, is plugged in Equation (9) for $\Theta_A^{-1}$, and a corrected estimate $\hat{P}_A$ can be obtained from the vector of the proportions, $P_A^*$, which is estimated from the misclassified variable

$$\hat{P}_A = \hat{\Theta}_A^{-1} \times P_A^*. \tag{14}$$

In the next section, we investigate the effectiveness of matrix back-calculation in correcting proportion estimates from misclassified variables under varying conditions. The research question is simple: Which of the three approximations of the MCM yields the best results given the information that is available to a researcher?

## Evaluation study

### Method

We conducted a Monte Carlo simulation in order to test the effectiveness of the three approximate MCMs for matrix back-calculation of proportion estimates.[5] As we have shown in the first section, matrix back-calculation will asymptotically uncover the true proportions by definition if the true MCM is known. The simulation is therefore not a test of the correction method *per se*, but a test of how well the three approximations of the MCM work under varying conditions. The correction of simple univariate proportion estimates via matrix back-calculation is the most simple use case of misclassification correction methods and a suitable example for the present evaluation. The results concerning the approximation of the MCM are also valid for the correction of more complex bi- and multivariate estimates and inferential statistics. Such extensions are introduced toward the end of this article.

The quality of a MCM approximation depends on the amount of available information about the misclassification process. In the context of a content analysis, the approximation is affected by (a) the number of coders, (b) the sample size of the test-standard comparison, and (c) the accuracy of the coding process. For (a) and (b), we selected numbers of coders and test sample sizes which are typically achievable in medium-sized and large research projects and which are also recommended as requirements for meaningful intercoder reliability tests (Krippendorff, 2011). Additionally, the bias of the naive estimate as well as the overall effectiveness of the correction procedure are strongly related to the dispersion of the true category proportions (see Figure 1, above), which we accounted for in the simulation. Finally, we also varied the number of categories in order to show the applicability to multicategorical variables with any number of categories. All discrete and continuous input parameters of the simulation are summarized in Table 2.

Overall, three discrete input parameters of the simulation varied systematically in a 3 ($k$ number of categories) × 3 ($m$ number of coders) × 2 ($n_{\text{Test}}$ test sample size) fully crossed design. This resulted in a total of 18 conditions. In order to adequately cover the sample space for the category proportions, we first randomly sampled 10,000 combinations of category

**Table 2.** Input parameters of the Monte Carlo simulation.

| Parameter | Values |
|---|---|
| Number of categories, $k$ | 2, 4, 6 |
| Vector of category proportions, $P$ | $\mathcal{U}(k, [0, 1])$ with $\sum P = 1$ |
| Coding accuracy, mean of the diagonal of $\Theta$ | $\mathcal{U}(1, [.5, 1])$ |
| Number of coders, $m$ | 2, 3, 8 |
| Test sample size, $n_{\text{Test}}$ | 30, 100 |

[5]The simulation was implemented in R (R Core Team, 2016), and full replication code is available at OSF: https://dx.doi.org/10.17605/OSF.IO/E7M9Z.

**Table 3.** Boundaries of the dispersion levels.

| k | Level | Lower bound | Upper bound |
|---|---|---|---|
| 2 | low | [0.50, 0.50] | [0.33, 0.67] |
| 2 | moderate | [0.33, 0.67] | [0.17, 0.83] |
| 2 | high | [0.17, 0.83] | [0.00, 1.00] |
| 4 | low | [0.25, 0.25, 0.25, 0.25] | [0.01, 0.29, 0.33, 0.36] |
| 4 | moderate | [0.08, 0.19, 0.26, 0.47] | [0.02, 0.06, 0.19, 0.72] |
| 4 | high | [0.02, 0.08, 0.17, 0.73] | [0.00, 0.01, 0.02, 0.98] |
| 6 | low | [0.16, 0.16, 0.16, 0.17, 0.18, 0.18] | [0.01, 0.05, 0.19, 0.22, 0.26, 0.27] |
| 6 | moderate | [0.04, 0.05, 0.12, 0.24, 0.26, 0.30] | [0.00, 0.01, 0.03, 0.14, 0.24, 0.57] |
| 6 | high | [0.01, 0.05, 0.09, 0.10, 0.15, 0.61] | [0.01, 0.02, 0.02, 0.02, 0.09, 0.84] |

*Note*. Differences from $\sum P = 1$ are rounding errors.

proportions within each condition from a multinomial distribution. We then computed the dispersion for each set of proportions, defined as the square root of the sum of the squared deviations from an equal distribution. We grouped the replications in three subsets of equal dispersion ranges (low, moderate, and high dispersion) for each $k$. Table 3 gives some example distributions at the boundaries of the dispersion levels for each $k$. Finally, we drew a stratified sample of 1,000 replications for every condition and dispersion level, and generated the accuracy for each simulation run by sampling from a uniform distribution, $\mathcal{U}(1, [.5, 1])$. The final data set of the Monte Carlo simulation therefore consists of 54,000 runs. A single simulation run included the following steps.

(1) A variable with the true category assignments was created, given the number of categories $k$ and the vector of proportions $P$. The category proportions were used to simulate the true data for $n = 1,000$ cases plus a test sample of a given size $n_{\text{Test}}$.
(2) A true MCM for the coding process, $\Theta$, was generated. The diagonal entries for the probability of a correct classification were allowed to vary between categories and constrained only to have a mean given by the pre-set coding accuracy. The off-diagonal error rates were unevenly distributed and constrained only to sum to 1 in each column together with the diagonal entry.
(3) Using the probabilities from the true MCM, the true content data was misclassified, yielding the *observed* content data.
(4) For every coder $m$, the test data was also misclassified, yielding $m$ codings for each test unit. Using the test data, the *observed* MCMs $\hat{\Theta}$ for the three approaches were estimated as described in the previous section.
(5) Finally, the proportion estimates $P^*$ were computed from the observed content data, and subsequently corrected via matrix back-calculation according to Equation (14), using each of the three MCM approximations.

We are primarily interested in the question whether matrix back-calculation using the three MCM approximations succeeds in producing proportion estimates which are closer to the true proportions than the simple observed proportions. For each simulation run, the Root Mean Squared Error (RMSE) of the corrected proportions was subtracted from the RMSE of the observed proportions. The result was divided by the RMSE of the observed proportions. The outcome quantifies the relative RMSE reduction (positive values, corresponding to a better estimate) or the relative RMSE increase (negative values, corresponding to a worse estimate), respectively.

## Results

The presentation of the results is structured according to the information which is available to researchers who want to choose an adequate correction procedure for their study. The simulation

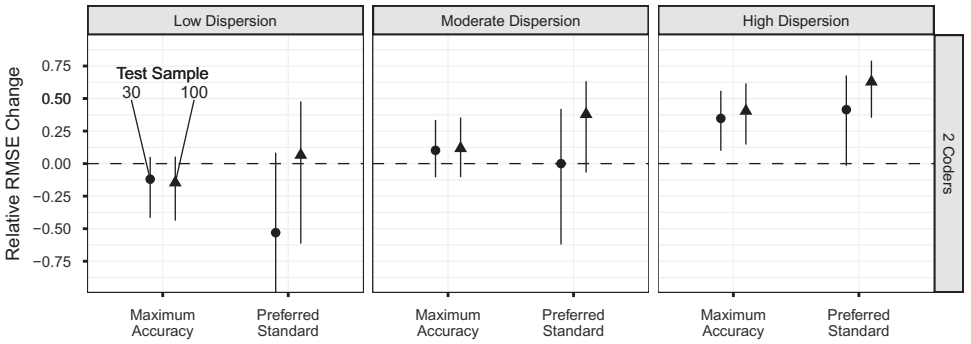


**Figure 3.** Performance of matrix back-calculation with different approximations of the misclassification matrix (two coders).

*Notes*. The figure summarizes the results of 18,000 simulation runs with $m = 2$ coders. A detailed description of the figure is given within the text. The y axis is cropped at the lower end in order to enhance readability. The lowest quartile for the preferred standard correction in the low dispersion, $n_{Test} = 30$ condition ends at −1.63.

outcomes are summarized across the true accuracy, which is unknown to the researcher. Accuracy is of course a major predictor of the bias of the observed proportions (see Figure 1, above), but it is not related to other factors in the simulation by design, and consequently the results for the other factors are not confounded by the aggregation. Likewise, we present the findings only for three general levels of dispersion of the true category proportions (low, moderate, high). Researchers cannot *a priori* know the true proportions, and the observed proportions are biased by unknown misclassification error. Yet both prior knowledge and the observed results should enable a reasonable guess whether the variable of interest has rather low, moderate, or high dispersion. Finally, because the number of categories, $k$, did not affect the results above and beyond the other factors, we summarized the simulation outcomes across the levels of $k$. All results hold up in separate analyses for variables with 2, 4, or 6 categories.

Both result figures (Figures 3 and 4) are constructed similarly, with the MCM approximations on the x axis and the relative change of the RMSE on the y axis. The plot columns show the results for low, moderate, and high dispersion of the true proportions, and the plot rows show the number of coders. The shapes represent the size of the test sample. All plots show median values of 3,000 runs of the simulation. If the median value is above 0 (the broken horizontal line), the corrected estimates were less biased than the observed proportions in over half of the simulation runs. This is the point where a researcher would, on the long run, report less biased estimates if he or she always used the respective correction procedure. The vertical lines show the interquartile ranges (IQR). If the IQR is completely above 0, three quarters of the simulations show an improvement of the proportion estimates by matrix back-calculation correction. The IQR additionally gives an impression of the variability of the corrections' performance. Wide intervals indicate high variability, that is, a great range of RMSE reduction or increase occurred during the simulation runs. Narrow intervals indicate that the correction performed more consistently.

Figure 3 presents the simulation results for studies with $m = 2$ coders, which allow only for MCM approximations by maximum accuracy or by preferred standard comparison. Studies with only two coders provide relatively little empirical information on the misclassification process. If the true category proportions are distributed evenly (left panel), the bias of the uncorrected estimate is expected to quite small, and matrix back-calculation does more harm than good. In the condition with the least amount of information, $m = 2$ coders and $n_{Test} = 30$ test units, the preferred standard MCM is based on only 60 test-standard comparisons. The estimated matrix is neither accurate nor precise and leads to more strongly biased results than the simple observed proportions. The preferred standard method performs better yet still insufficiently for moderately dispersed proportions, again with high variability reflecting the uncertainty in the estimated MCM. In this low-

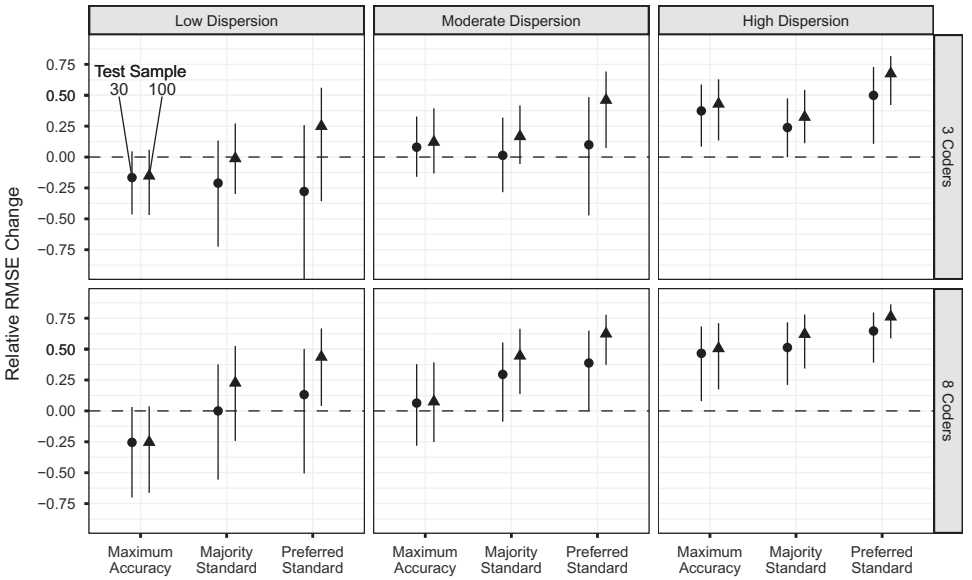


**Figure 4.** Performance of matrix back-calculation with different approximations of the misclassification matrix (more than two coders).

*Notes*. The figure summarizes the results of 36,000 simulation runs with $m > 2$ coders. A detailed description of the figure is given within the text. The y axis is cropped at the lower end in order to enhance readability. The lowest quartile for the preferred standard correction in the low dispersion, $n_{\text{Test}} = 30$, $m = 3$ condition ends at −0.91.

information setting, matrix back-calculation correction is only recommendable for variables which are believed to be very unevenly distributed. An increase of the test units to $n_{\text{Test}} = 100$ substantially enhances the effectiveness of the preferred standard procedure. It works sufficiently well for moderately and very well for highly dispersed category proportions, in the latter case almost always providing less biased results than the uncorrected estimates.

Correcting with a MCM derived from the maximum possible accuracy assumption leads to substantial improvements for moderate and particularly for high dispersion levels. It outperforms the preferred standard procedure if only little empirical information is available and comes second only in the high dispersion, $n_{\text{Test}} = 100$ condition. The results are generally less variable, because all available information is used to estimate just one parameter of the MCM, the maximum possible accuracy. The efficiency comes at the cost of bias, because in the simulation (like in reality), both assumptions are almost never perfectly met. Additional information does not increase performance asymptotically, because the assumptions do not become more correct with more data.

Figure 4 additionally includes the results for the majority standard procedure, which is only possible with at least $m = 3$ coders. A comparison with Figure 3 and across the rows of Figure 4 shows that more coders improve the effectiveness of all approaches, since more information increases the quality of all MCM approximations. The majority standard correction benefits particularly from additional coders, because both a more accurate majority standard (see Figure 2, above) and an increased number of test-standard comparisons become available. If a research project employs $m = 8$ coders, the majority standard procedure is a viable option in almost all conditions. It even shows improvement over the uncorrected estimates with the minimum requirement of $m = 3$ coders for moderately (with $n_{\text{Test}} = 100$) and highly dispersed variables.

A MCM approximation by maximum possible accuracy is nevertheless preferable to the majority standard approach if the number of coders $m$ is small. Yet the maximum accuracy procedure benefits less from additional coders or test units compared to both test-standard approaches. The major drawback of this procedure is that it is not suitable for correcting the proportion estimates of evenly distributed

variables. In the low dispersion condition, the bias of the uncorrected estimates is mainly caused by unequal difficulties of the categories and unequal error rates. The approaches based on test-standard comparisons yield better estimates of these characteristics of the MCM with more empirical information about the coding process. In contrast, the assumed EDEERM of the maximum accuracy approximation makes it impossible to account for such error structures, and it therefore always performs subpar, even with many coders and test units (see the lower left facet of Figure 4).

The preferred standard procedure works best in most conditions, especially if many test-standard comparisons are available ($n_{\text{Test}} = 100$). It even shows acceptable ($m = 3$) or good ($m = 8$) effectiveness for variables with low dispersion. The most relevant drawback of the preferred standard approach is its high variance with lower dispersion and less information (smaller $m$ and $n_{\text{Test}}$). Under most conditions, the corrected estimates are not more strongly biased than the uncorrected estimates on average, but the correction adds more uncertainty.

## Discussion

Although matrix back-calculation correction of course cannot magically "heal" measurement errors, it is a viable option to correct for the consequences of coding errors in content analysis. The simulation has shown that it can substantially reduce the bias of proportion estimates if the correction is made with an adequate MCM approximation. The gains are highest when the true proportions are strongly dispersed and, consequently, the uncorrected observed estimates are most strongly biased. However, the potential for improvement is rather limited and much information on the misclassification process is needed when the true dispersion is low and the bias is mainly caused by unequal error rates. More information about the coding process, as provided by larger test sample sizes and more coders, leads to a better performance of the correction procedure. Especially the approximations based on test-standard designs, the majority and the preferred standard procedures, benefit from additional information. Only small gains are visible for the MCM with the maximum possible accuracy, because it is derived less from empirical information and more from plausible, but imperfect assumptions.

In sum, the *preferred standard* procedure performed best under most conditions. The approach only produces worse estimates than the uncorrected proportions if very little information on the misclassification is available. However, the preferred standard estimate of the MCM of course requires additional information, that is, a preferred standard measure. If such a measure is available, the procedure is recommended in most circumstances. The *majority standard* procedure is a viable alternative if researchers cannot, or are not willing to, collect a preferred standard, but employ at least three coders. It works well under many conditions, especially as the number of coders increases. Finally, a misclassification matrix which is derived from the *maximum possible accuracy* is an alternative if the true proportions are believed to be at least moderately unequal and one has only little empirical information about the coding process. However, it should *not* be used if it is likely that the variable of interest has low dispersion.

### *Illustrative example*

We illustrate the application of matrix back-calculation correction with a common scenario in communication research—proportion estimates of a categorical variable. For the sake of simplicity, we use a dichotomous variable, but the procedure is identical for variables with more than two categories. In a recent article, Schuck, Boomgaarden, and De Vreese (2013) reported the results of a large cross-national content analysis on the 2009 European elections.[6] Specifically, strategy framing

[6]In the original publication, three dichotomous indicator variables were used to compute an average score for the strategy frame, and an average reliability score. However, for the purpose of this illustration, this mean index is just a rescaled dichotomous variable, which does not change the interpretation of proportions.

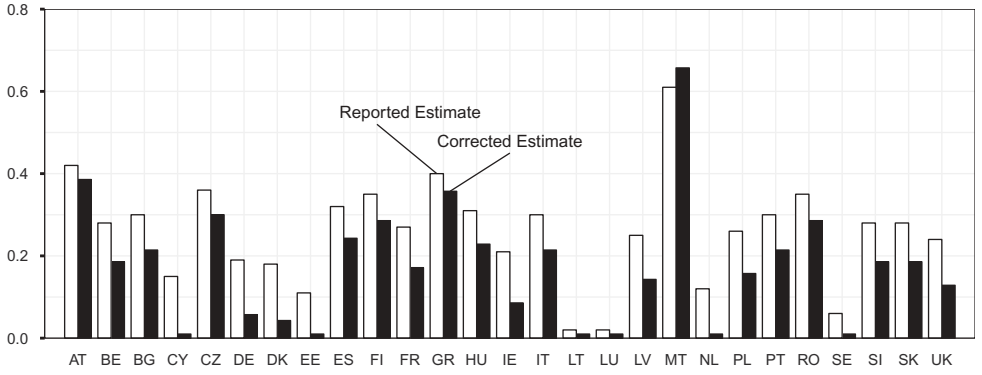


**Figure 5.** Reported and corrected estimates of strategy framing in 27 countries.
*Notes*. The reported estimates are from Figure 1 of Schuck et al. (2013).

in the campaign coverage was coded by 58 coders. The intercoder reliability from a test of 35 randomly selected articles was Krippendorff's $\alpha = .61$.

Across all media outlets in all 27 countries, Schuck et al. (2013) reported an average proportion of .29 for the strategy frame, which means that one can expect 29 "fully present" strategy frames in 100 news articles on the EU election. As we have shown above, imperfect intercoder agreement indicates that the accuracy of the coding process cannot be perfect. Even without further information on the coding process, we can make an educated guess about the maximum possible accuracy. We assume an intercoder agreement of .7, because the reported chance-corrected $\alpha$ coefficient is usually somewhat lower than simple percent agreement. Assuming interchangeable coders and Equation (12), the maximum possible accuracy is .85. Matrix back-calculation with a MCM based on the maximum accuracy assumption yields a corrected proportion estimate of .2, roughly one third or 9 percentage points below the reported proportion. Likewise, we can compute a corrected version of their Figure 1, which displays the proportion of strategy frames in the news coverage of each EU member country (see Figure 5). The corrected proportions are, with the exception of Malta, substantially lower than the observed proportions, sometimes dramatically so: The estimates for Germany and Denmark, for example, are less than half as large as the original proportions. The results illustrate that the bias introduced by misclassification is affected by the dispersion of the categories.

The corrected estimates of course depend on assumptions, namely that the coding process had an accuracy of .85 and equal error rates, and that misclassification occurred consistently across countries. We emphasize that we are not claiming that the example correction uncovered perfectly unbiased estimates. However, our assumptions are explicitly stated and, at the very least, more reasonable than the assumption of perfect accuracy with imperfect intercoder agreement. The correction has consequences for the substantive interpretation of the results: Strategy framing seemed to be less present in the news coverage of the most countries. However, the differences between the countries turned out to be larger than reported. The latter is not only interesting from a descriptive perspective, it is also relevant for subsequent media effects analyses (such as those subsequently conducted by Schuck et al., 2013). More equal observed distributions of media messages which are assumed to exert an effect can lead to substantially attenuated media effect estimates in studies which link a media content analysis to survey data (Scharkow & Bachl, 2016).

## Extensions

In this article, we have covered the consequences of misclassification error for univariate quantities of categorical data, in particular for proportion estimates. We then introduced and evaluated

correction by matrix back-calculation as a promising remedy. Descriptive presentations of univariate quantities are important in many content analytical studies. However, more sophisticated analyses are often conducted with content analytical data, and their results are also impaired by misclassification. Below, we give a concise overview of some extensions. All of them require approximations of the MCM. The three approximations, which were described and evaluated in detail above, can be similarly implemented with the extensions.

### *Inferential statistics for univariate summaries*

If the coding units of a study are a sample from a larger population, researchers might want to report standard errors or confidence intervals for their proportion estimates. Kuha and Skinner (1997, p. 650, Equation 28.15) provide a formula for the variance-covariance matrix of the vector of corrected proportions, which takes into account the uncertainty both in the vector of observed proportions and the estimated misclassification matrix. A double bootstrap procedure, which is applied to the estimation of the misclassification matrix *and* to the correction by matrix back-calculation, is a viable alternative.

### *Cross-tabulations with error-free variables*

If a misclassified variable is cross-tabulated with a variable which is most likely not subject to misclassification (e.g., formal variables such as date, media outlet, or country of origin), matrix back-calculation can simply be applied to the group proportions, on the condition that each subsample is sufficiently large. Figure 5 above provides an example for this approach. The application of matrix back-calculation to subsamples adds another assumption, *nondifferential misclassification*: The MCM is assumed to be independent of the grouping variable, that is, coding is not affected by, for example, the outlet in which an article was published. Most content analyses already build on this assumption, because the established tests of intercoder reliability are usually conducted without regard to grouping variables. If researchers assumed that coding quality differs between, for example, media outlets, they also would have to conduct and report separate intercoder tests for each outlet.

### *Statistical models with one or more misclassified variables*

In two-way tables with at least one misclassified variable, statistical relationships are attenuated and statistical power is reduced (Bross, 1954; Kuha & Skinner, 1997). The consequences of misclassification error for more complex statistical models are far from trivial. Both attenuation and inflation of effect sizes are possible, as are inflated type I or type II error probabilities for statistical tests. Kuha and Skinner (1997) give an overview of traditional model-based correction methods. Another promising alternative is the misclassification simulation and extrapolation method (MC-SIMEX) (Küchenhoff, Lederer, & Lesaffre, 2007; Küchenhoff, Mwalili, & Lesaffre, 2006). First, a sequence of increasingly biased estimates is obtained by fitting the model to the observed and successively misclassified data, which is generated according to a MCM (misclassification simulation step). Second, the estimates are extrapolated backwards to the point where no misclassification error existed (extrapolation step). MC-SIMEX rests on very few assumptions, can be used with any number of misclassified variables, and is straightforward to implement for many statistical models. Lederer and Küchenhoff (2006) provide an easy-to-use implementation for linear and generalized linear models in *R*. A user only has to define the naive statistical model of the observed data and a MCM in order to obtain corrected estimates. First tests assured us that MC-SIMEX works well with the three MCM approximations which were evaluated in this article.

### *Adjusting estimates from automatic coding data*

Matrix back-calculation (as well as MC-SIMEX and other extensions) can also be used to adjust the estimates of data which were generated by automatic coding procedures to the estimates which would be expected if manual coders classified all coding units. The MCM is simply exchanged by the

confusion matrix, in which the categories of the manual coders are treated as the preferred standard. Note that this procedure does not provide corrected estimates in sense of this article, but estimates which are adjusted to the—probably also biased—manual categories. A two-step procedure which first adjusts the estimates from the automatic coding data to manual coders and then corrects these estimates as described in this article seems intuitive at first glance, but further evaluations are needed on the validity of such approaches.

## Conclusion

Misclassification error is inevitably present in content analytical data. The current practice in communication research is paradoxical: We acknowledge measurement imperfection by reporting results from intercoder reliability tests, but then proceed with statistical analysis and substantial interpretation as if no misclassification had occurred. In this article, we have shown that misclassification of categorical variables, which make up the majority of content analytical measures, almost always leads to biased estimates even of simple univariate quantities such as category proportions. The amount and direction of the bias are affected by both the misclassification process and the distribution of the true categories. Consequently, misclassification can never be safely ignored, because both properties are generally unknown to the researcher.

Matrix back-calculation is a simple correction method for univariate quantities which can be used for many typical analyses in content analytical studies. Correction procedures for more complex models are also available. All correction methods require information on the misclassification process. We described three approaches to approximate the MCM and investigated their effectiveness. The Monte Carlo simulation demonstrated the potentials of matrix back-calculation based on the three approximations, with substantial improvement under many conditions. The correction works best if meaningful information on the coding process is available: The MCM approximation benefits from larger test samples, more coders, and—if available—an external preferred standard. Researchers who take measurement seriously have to conduct informative tests of the coding process (see Krippendorff, 2011, for a similar request with regard to intercoder reliability). Benefits from larger tests are manifold, because they not only provide better estimates of the MCM, but also more meaningful estimates of chance-corrected intercoder reliability and, during the process, more intensive coder training. But even with limited information about the misclassification process, a correction based on the maximum accuracy assumption can improve the estimates under many conditions, and it is more realistic than the self-deceit of perfect accuracy with imperfect intercoder agreement.

In our opinion, the current practice of ignoring misclassification error in content analysis can and should be abandoned in favor of methods which propagate misclassification error. Not doing so bears the danger of reporting biased results and drawing invalid conclusions.

## References

Bauman, K. E., & Koch, G. G. (1983). Validity of self-reports and descriptive and analytical conclusions: The case of cigarette smoking by adolescents and their mothers. *American Journal of Epidemiology*, *118*(1), 90–98. doi:10.1093/oxfordjournals.aje.a113620

Benoit, K. R., Conway, D., Lauderdale, B. E., Laver, M., & Mikhaylov, S. (2016). Crowd-sourced text analysis: Reproducible and agile production of political data. *American Political Science Review*, *110*(2), 278–295. doi:10.1017/S0003055416000058

Bross, I. (1954). Misclassification in 2 X 2 tables. *Biometrics*, *10*(4), 478–486. doi:10.2307/3001619

Buonaccorsi, J. P. (2010). *Measurement error: Models, methods, and applications*. Boca Raton, FL: CRC Press.

Grimmer, J., King, G., & Superti, C. (2015). *The unreliability of measures of intercoder reliability, and what to do about it*. Retrieved from http://stanford.edu/~jgrimmer/Handbib.pdf

Hayes, A. F., & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures*, *1*(1), 77–89. doi:10.1080/19312450709336664

Krippendorff, K. (2004a). *Content analysis: An introduction to its methodology* (2nd ed.). Thousand Oaks, CA: Sage.

Krippendorff, K. (2004b). Reliability in content analysis. Some common misconceptions and recommendations. *Human Communication Research*, *30*(3), 411–433. doi:10.1111/j.1468-2958.2004.tb00738.x

Krippendorff, K. (2008). Systematic and random disagreement and the reliability of nominal data. *Communication Methods and Measures*, *2*(4), 323–338. doi:10.1080/19312450802467134

Krippendorff, K. (2011). Agreement and information in the reliability of coding. *Communication Methods and Measures*, *5*(2), 93–112. doi:10.1080/19312458.2011.568376

Krippendorff, K., & Craggs, R. (2016). The reliability of multi-valued coding of data. *Communication Methods and Measures*, *10*(4), 181–198. doi:10.1080/19312458.2016.1228863

Küchenhoff, H., Lederer, W., & Lesaffre, E. (2007). Asymptotic variance estimation for the misclassification SIMEX. *Computational Statistics & Data Analysis*, *51*(12), 6197–6211. doi:10.1016/j.csda.2006.12.045

Küchenhoff, H., Mwalili, S. M., & Lesaffre, E. (2006). A general method for dealing with misclassification in regression: The misclassification SIMEX. *Biometrics*, *62*(1), 85–96. doi:10.1111/j.1541-0420.2005.00396.x

Kuha, J., & Skinner, C. (1997). Categorical data analysis and misclassification. In L. E. Lyberg, P. Biemer, M. Collins, E. D. De Leeuw, C. Dippo, N. Schwarz, & D. Trewin (Eds.), *Survey measurement and process quality* (pp. 633–670). New York, NY: Wiley.

Lederer, W., & Küchenhoff, H. (2006). A short introduction to the SIMEX and MCSIMEX. *R News*, *6*(4), 26–31.

Lombard, M., Snyder-Duch, J., & Bracken, C. C. (2002). Content analysis in mass communication: Assessment and reporting of intercoder reliability. *Human Communication Research*, *28*(4), 587–604. doi:10.1111/j.1468-2958.2002.tb00826.x

Lovejoy, J., Watson, B. R., Lacy, S., & Riffe, D. (2014). Assessing the reporting of reliability in published content analyses: 1985–2010. *Communication Methods and Measures*, *8*(3), 207–221. doi:10.1080/19312458.2014.937528

Marshall, R. J. (1990). Validation study methods for estimating exposure proportions and odds ratios with misclassified data. *Journal of Clinical Epidemiology*, *43*(9), 941–947. doi:10.1016/0895-4356(90)90077-3

Mikhaylov, S., Laver, M., & Benoit, K. R. (2012). Coder reliability and misclassification in the human coding of party manifestos. *Political Analysis*, *20*(1), 78–91. doi:10.1093/pan/mpr047

Neuendorf, K. A. (2002). *The content analysis guidebook*. Thousand Oaks, CA: Sage.

R Core Team. (2016). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Riffe, D., & Freitag, A. (1997). A content analysis of content analyses: Twenty-five years of Journalism Quarterly. *Journalism & Mass Communication Quarterly*, *74*(3), 515–524. doi:10.1177/107769909707400306

Scharkow, M., & Bachl, M. (2016). How measurement error in content analysis and self-reported media use leads to minimal media effect findings in linkage analyses: A simulation study. *Political Communication*, 1–21. doi:10.1080/10584609.2016.1235640

Schmid-Petri, H., Adam, S., Schmucki, I., & Häussler, T. (2015). A changing climate of skepticism: The factors shaping climate change coverage in the US press. *Public Understanding of Science*. doi:10.1177/0963662515612276

Schuck, A. R. T., Boomgaarden, H. G., & De Vreese, C. H. (2013). Cynics all around? The impact of election news on political cynicism in comparative perspective. *Journal of Communication*, *63*(2), 287–311. doi:10.1111/jcom.12023

Schwartz, J. E. (1985). The neglected problem of measurement error in categorical data. *Sociological Methods & Research*, *13*(4), 435–466. doi:10.1177/0049124185013004001

Volkens, A. (2007). Strengths and weaknesses of approaches to measuring policy positions of parties. *Electoral Studies*, *26*(1), 108–120. doi:10.1016/j.electstud.2006.04.003