Statistical Power in Content Analysis Designs: How Effect Size, Sample Size and Coding Accuracy Jointly Affect Hypothesis Testing – A Monte Carlo Simulation Approach.

Stefan Geiß

CCR 3 (1): 61–89
DOI: 10.5117/CCR2021.1.003.GEIß

Abstract

This study uses Monte Carlo simulation techniques to estimate the minimum required levels of intercoder reliability in content analysis data for testing correlational hypotheses, depending on sample size, effect size and coder behavior under uncertainty. The ensuing procedure is analogous to power calculations for experimental designs. In most widespread sample size/effect size settings, the rule-of-thumb that chance-adjusted agreement should be ≥.80 or ≥.667 corresponds to the simulation results, resulting in acceptable α and β error rates. However, this simulation allows making precise power calculations that can consider the specifics of each study's context, moving beyond one-size-fits-all recommendations. Studies with low sample sizes and/or low expected effect sizes may need coder agreement above .800 to test a hypothesis with sufficient statistical power. In studies with high sample sizes and/or high expected effect sizes, coder agreement below .667 may suffice. Such calculations can help in both evaluating and in designing studies. Particularly in pre-registered research, higher sample sizes may be used to compensate for low expected effect sizes and/or borderline coding reliability (e.g. when constructs are hard to measure). I supply equations, easy-to-use tables and R functions to facilitate use of this framework, along with example code as online appendix.

Keywords: Content analysis; Power analysis; Sample size; Effect size; Intercoder reliability; Intercoder agreement; Hypothesis testing; Monte Carlo simulation;

Krippendorff (2016) calls for analyses of the effects of lacking coder agreement on subsequent analytical procedures. This study addresses this call. Simulation studies are well-suited to explore how lack of coder agreement[1] affects statistical inferences. The probably most common *inference* in content analysis studies concerns whether relationships between content variables found in a sample exist in the population or not. I call this type of hypotheses "relationship exists hypothesis" (REH). This study analyzes the capacity of empirical studies to reliably discover correlations of varying strengths. Reliable discovery of a correlation is conceptualized as joint low type I and type II error rates ($\alpha<.05$ and $\beta<.05$). This means two things: (1) If we find a significant correlation in a single study, we can be relatively certain that it also applies to the population the study is sampled from (low type I error rate). (2) If we find no correlation, there is most likely no correlation in the underlying population either that is equal to or greater than the specified strength (low type II error rate).

Currently, analyzing coding agreement is the most important tool for assessing the quality of content analysis data (Feng, 2014; Lombard et al., 2002). To that end, communication researchers rely on fixed benchmarks or threshold values (coefficient-specific or general ones) to judge the reliability of content analysis data (e.g. Krippendorff, 2004a; Landis & Koch, 1977). They are easy to apply and provide a general benchmark which levels of reliability are conventionally reached and should be aspired; however, there is a certain degree of arbitrariness involved in choosing benchmark values (Krippendorff, 2004a; Landis & Koch, 1977). And the different critical values and interpretation guidelines are partly contradictory (Altman, 1991; Fleiss et al., 2003; Krippendorff, 2004a; Landis & Koch, 1977). The threshold values reflect scholars' practical experience and their intuition. Developing additional criteria that complement the way we assess the usefulness of content analysis data is needed. Otherwise we run the risks of systematically (1) using data unsuitable for testing a hypothesis; or (2) dumping data even though it is suitable for testing a hypothesis.

Adding effect size and sample size into the equation. That is even more important because we know that coding reliability (as in indicator of coding accuracy) is only one of several factors that influences whether tests of REHs can be relied upon. Analogous to statistical power calculations for experiments (Cohen, 1988), the sample size and the effect size (which can only be estimated) are highly important factors that should be taken into consideration beyond the accuracy of the measurement. This study adds a rationale to how we currently design, report and evaluate quantitative content analysis data: To target sufficient statistical power, we should view (expected) coding accuracy (and coding difficulty), expected effect sizes and sample sizes in conjunction.

Rather than having to rely solely on critical values, this study's simulation results enable rigid computations that will make it possible to estimate what kinds of inferences (for REHs) are possible at a particular constellation of effect size, sample size, and coder agreement. Thereby, it is possible to answer the question to what extent data allow the inferences researchers want to make. This additional anchoring of inter-coder reliability coefficients regarding their consequences for hypothesis testing should also help in demonstrating the consequences of lacking coding accuracy—and help in persuading researchers of the value of testing, monitoring, reporting and discussing inter-coder reliability to combat "disuse, misuse, and abuse" (Feng, 2014) of coder reliability testing.

Coder decision making under uncertainty. There is one additional complication, however: As it is to some extent contested what "non-recognition" means in statistical terms (Feng, 2013; Krippendorff, 2012; Zhao et al., 2012), a fourth factor is the mode of how coders choose codes if they do not recognize the true value of a text to any degree. One can use various models of coder behavior under uncertainty or ignorance, and the simulation will include two distinct models: equality-distribution (ED) and marginal-distribution (MD) chance coding.

Planning your study. Effect sizes cannot be changed by the researchers because they are an inherent characteristic of the real-world relationships one wants to investigate. This leaves content analysts with two possibilities to improve the statistical power of their study: (1) improve the coder agreement and (2) increase the sample size. From the perspective of statistical power, the choice which possibility to pursue does not matter.[2] Coding more material may sometimes be the cheaper solution compared to investing into additional coder training and more rounds of inter-coder agreement testing with uncertain outcomes. Additionally, scaling up sample size can equip communication researchers with the instruments to study phenomena that are farther away from the pole of "manifest" content (Berelson, 1971) and more prone to subjectivity where coders' perceptions and conceptions pollute the measurement (Kepplinger, 1989; Krippendorff, 2017). In such "difficult coding tasks" (Feng & Zhao, 2016), coder agreement can hardly be improved beyond a certain level. And in fact, as long as coder agreement is significantly above chance (Feng, 2013), one can try to improve power to the desired level by boosting the sample size. This resonates with the assertion that besides "1" meaning perfect reliability (and, I would add, "0" meaning that similar agreement could have resulted from chance), "there are no magical numbers" (Krippendorff, 2004b, p. 429). It would be a mistake to effectively ban "difficult coding tasks" from content analytic research. If the statistical power is sufficient for the intended test, even data gathered at substandard levels of coder reliability should be considered as long as systematic errors can be ruled out. To that end, we need rigid methodological research that can tell us under which conditions data can be used to test REHs with joint low α and β error probabilities.

An example may be helpful here: We want to study whether messages with high "argument strength" also tend to feature more "emotional appeals". We anticipate that these are "difficult coding tasks" where coder agreement might be as low as α_K=.50. However, previous research also suggests that there is a strong correlation between argument strength and emotional appeals of approximately R=.40. We can then choose a sample size that allows for reliable testing of the hypothesis (α<.05; β<.05) at the anticipated effect size and coder agreement. Such expectations can be preregistered (van 't Veer & Giner-Sorolla, 2016) to make sure that expectations were formulated a priori rather than a posteriori.

The simulation study. The current study reports a simulation of "content analyses" that explicitly varies those four factors (sample size, effect size, coding accuracy, coder decision-making under uncertainty). Based on the simulation results it estimates equations for finding minimum coding accuracy requirements when specifying acceptable α and β error probabilities, sample size, coder behavior under uncertainty, and the expected size of effect. Furthermore, it provides equations, a downloadable calculator (R script) as well as a set of tables for calculating sample sizes according to expected/obtained effect sizes and expected/obtained coder agreement. It is therefore a valuable resource for scholars evaluating, reporting and designing content analyses that involve REHs.

Inter-coder reliability interpretation guidelines

Random errors, individual systematic errors, and joint systematic errors

Coder agreement is the dominant indicator for assessing the validity or accuracy of coding. High agreement (as a version of reliability testing) is a prerequisite for high validity, but high agreement alone cannot guarantee high validity (Krippendorff, 2004a). Agreement coefficients can only uncover those errors that result in disagreement, which is not always the case. To understand this, it is helpful to discern three types of errors, i.e. deviations between a measurement and the true value: random errors, individual systematic errors, and joint systematic errors. In random errors, the extent and direction of the errors is unpredictable. In both types of systematic errors, the extent and direction is predictable. Unlike random errors, systematic errors can potentially produce data patterns that are artefacts, making them particularly dangerous. Coder agreement will decrease if coders make random errors or individual systematic errors, but it will not decrease if all coders make the same kind of (joint) systematic errors (Table A1). In a sense, the current practice of using coder agreement to estimate data quality (coding accuracy) presupposes the absence of joint systematic errors.

What does the assumption of random errors mean in practical terms for content analysis? Is it a realistic or a baseless assumption? Realistically, it is likely that many coding errors are in fact systematic rather than random; coders establish their personal heuristics and routines that may systematically affect how they code content (as in Table A1). However, content analyses that employ multiple coders can effectively reduce the consequences of individual coders' systematic errors by randomly distributing the material among the coders. Then, individual systematic errors can only produce artefacts if multiple coders consistently make the same systematic errors ("joint systematic error"); for practical considerations, individual systematic errors can be treated as random errors because they are randomly distributed throughout the material.

The Monte Carlo simulation at the core of this study will suppress systematic errors and include only random coding errors. Thereby, coder agreement becomes an unbiased indicator of coding accuracy (the inverse of coding error). I will discuss the limitations of excluding systematic errors in the discussion.

Overview over guidelines and thresholds

Krippendorff (2004a) suggests that Krippendorff's α's (α_K's) as high as .800 are necessary for trusting the coding, while α_K's between .800 and .667 may suffice for drawing preliminary conclusions. Values lower than .667 are characterized as generally inacceptable. Riffe, Lacy, and Fico (1998) agree to the limits proposed by Krippendorff (2004a) but state that in exploratory and ground-breaking research lower values may satisfice. Landis and Koch

(1977) have characterized Cohen's or Conger's $\varkappa$'s ($\varkappa_C$'s) between .81 and 1.00 as "almost perfect", between .61 and .80 as "substantial", between .41 and .60 as "moderate", between .21 and .40 as "fair" and between .01 and .20 as "slight" agreement; in a liberal interpretation, one might conclude from their labels that use of data material collected at $\varkappa_C$ >.4 or even $\varkappa_C$ >.2 is permissible, relatively independent of the circumstances. Altman (1991) relabels the limits mentioned by Landis and Koch (1977) ("very good", "good", "moderate", "fair", "poor"). Fleiss, Levin, and Paik (2003) rate agreement as "excellent" if Fleiss' $\varkappa$ ($\varkappa_F$)>.75, as "fair" to "good" if $.75 \geq \varkappa_F \geq .40$, and as "poor" if $\varkappa_F < .40$; according to this, using data collected at $\varkappa_F$ as low as .40 could be acceptable. Under most conditions $\varkappa_C$, $\varkappa_F$, and α_K give similar estimates of agreement as they are closely related coefficients (Feng, 2013; Hayes & Krippendorff, 2007; Zhao et al., 2012). This makes the different standards mentioned in the literature even more striking.

Justifications for reliability interpretation guidelines

Justifications for reliability guidelines are either missing or unsatisfactory. Krippendorff (2004a) calls for "suitable experiments" to "verify" (p. 241) his suggestions; he vaguely describes one experiment, concluding that it produced data that "nobody in their right mind would draw conclusions from […]" (Krippendorff, 2004a, p. 241)—which still resulted in $\alpha_K = .44$. However, the experiment and the argument appear ill-suited to further justify a minimum requirement of .667 for α_K. In the end, the recommendations by Fleiss et al. (2003) and by Krippendorff (2004a) are based on the intuition and experience of the respective authors. This is even more obvious for Landis and Koch (1977) who just split the range of above-chance values into five equally-spaced parts, stating: "Although these divisions are clearly arbitrary, they provide useful 'benchmarks' […]" (p. 165, emphasis in original).

The recommendation of both Krippendorff (2004a) and Gwet (2014) to recognize the probabilistic nature of reliability coefficients is helpful. Still, the anchor values and their labels may reflect extensive practical experience and intuition (and therefore be valuable), but lack sufficient evidence. In the case of testing REH's, these rules of thumb can be supplemented by how coding accuracy would impact the power of statistical tests.

"Relationship Exists Hypotheses" (REHs)

There are different types of rationales involved in research that is based on coding results. This paper deals only with content data for *testing REHs*, which covers a broad array of applications of content analysis data. A great number of studies uses content analysis data test hypotheses (or answer research questions) using statistical tests as to whether two or more variables are correlated. These are *REHs*. In the case of REHs, the clear-cut distinction between significant and non-significant results makes the definition of "thresholds" in coding reliability more than just a rule-of-thumb or heuristic: If we specify α (type I, false positive) and β (type II, false negative) error probabilities, the trinity of (true) effect size, (true) coding accuracy, and sample size should determine whether the effect can be detected or not. REH tests are commonly carried out using classical inferential statistics that control the α error probability (if errors can be treated as random). Beyond that, estimates of effect size may be presented. However, they usually only serve to contextualize the test result, i.e. show how strong or weak the effect is once its existence has been established. I will concisely address the applicability of the simulation results to other contexts in the discussion section.

Factors affecting content analysis hypothesis tests

Effect Size

The term *effect size* refers to bivariate or multivariate relationships and describes how closely two or more variables are associated and to what degree they are predictable, given full knowledge of the other variables. Examples include R and R^2, η^2, Cohen's d or Cramér's V. I use Pearson's R statistics as a measure of effect sizes because most readers will be familiar with R; it presupposes interval- or ratio-level data.

Probability Theory suggests that inferential biases from measurement errors—if they can be conceptualized in terms of noise or random errors—generally lead to underestimation of effect sizes by attenuating data patterns (Gustafson, 2004). This will lead to conservative statistical inferences, increasing the likelihood of choosing the null (H_0) rather than the alternative hypothesis (H_1). *H-1: Decreasing accuracy of coding will increase the likelihood of false negative hypothesis test results and deflated effect size estimates. Decreasing accuracy of coding will not increase the likelihood of false positive hypothesis test results and of inflated effect size estimates.*

How *effect size* affects the chance to discover a data pattern has been incorporated in so-called power analyses (Cohen, 1988). It deals with the probability of producing false negative findings (type II or β errors). The statistical power of an analysis (1-β), i.e. its capability of discovering the incorrectness of the null hypothesis (given it is incorrect), depends on the sample size, the researcher-defined level of significance (maximum acceptable α error probability) and the true *effect size* that needs to be estimated. A larger sample, a laxer level of significance, and higher *effect size* will all lead to increases in statistical power. Detecting an effect that is strong should be possible even in the face of substantial noise (Gustafson, 2004) caused by low data quality, or smaller samples. In contrast, detecting an effect that is weak requires higher levels of data quality and/or larger samples. *H-2: The greater the true effect size, the lower the minimum coding accuracy necessary for finding a significant effect* (*or: for rejecting the null hypothesis*).

This being the first investigation of this kind, there are no attempts in the literature to formulate functions or provide tables describing what level of coding accuracy is required at a given level of significance and effect size of a given strength. Establishing such relationships would be helpful in designing and evaluating content analyses, however. Therefore, I ask: *RQ1: What level of accuracy of coding is necessary for correctly identifying an existing effect as statistically significant? How does it depend on the* effect size*, the* sample size *and the* coder behavior under uncertainty*?*

Coding Accuracy and Coder Agreement

Inter-coder agreement minimum requirements (or “cutoffs”) could be misinterpreted as if they were discrete points at which data quality suddenly drops towards zero; by using them to evaluate the publication-worthiness of studies without considering additional factors (such as sample size), reviewers and editors would implicitly subscribe to such a simplistic view. Krippendorff (2004) as well as Landis and Koch (1977) attest that the values they mention are chosen arbitrarily. But if inter-coder agreement gradually rather than discretely affects data quality and cutoffs are to some degree arbitrary, an approach that allows context-sensitive cutoffs would be better-suited to judging whether data allow making the inferences a study wants to make.

Reliability problems are a manifestation of measurement errors and cause inferential biases (explicitly in Fico et al., 2008). Its effects are hardly predictable if measurement error is systematic. Distortions induced by random measurement error are more predictable. Gustafson (2004) illustrates that random measurement error gradually rather than discretely affects estimates of correlations.

Gustafsson’s simulation shows that the negative effect of measurement errors on estimates is not strictly linear, but follows a sigmoid (S-shaped) pattern where an increase in measurement error is more consequential when the level of measurement error is moderate (Gustafson, 2004). The area where the sigmoid function slopes upward may justify a discrete cutoff-point to some extent if the sigmoid function has a very steep slope. Still, even if there is a marked increase in data quality in a narrow band, this cutoff point will not be the same for each study, but depend on sample size and effect size. Figure A1 illustrates different possible shapes of sigmoid curves where the point and abruptness of the "take off" can vary substantially.

H-3: (*a*) *Declining coding accuracy will gradually rather than discretely affect precision of correlation estimates and REH tests.* (*b*) *Decreasing coding accuracy will affect estimates of correlations according to a sigmoid pattern.*

Sample Size

Sample size is a factor whose effect figures most clearly when imagining conducting the same study multiple times with the same sample size; then, one changes the sample size, again conducts multiple studies, and compares the results. The overarching patterns in such simulations is that average estimates (here: *R*) are not affected by sample size, but that the spread of results (*SE*) is the greater the lower the sample size is ($\widehat{\mathrm{SE}}=\frac{1-R^2}{\sqrt{n-1}}$) (Chan & Chan, 2004). This means that with larger sample sizes, one can better rely on the results of a single study compared to a study with smaller sample sizes, ceteris paribus. Since communication researchers typically conduct a single study, greater sample size facilitates inferences from data at given levels of validity and reliability. *H-4: The larger a sample is, the lower the minimum level of coder accuracy necessary to test an REH with $\alpha<.05$ and $\beta<.05$ error rates.*

Simulating coder behavior

In a simulation study of inter-coder agreement, it is necessary to explicitly model exactly how coders decide if they do not or only partially recognize the true value of a unit they are supposed to code. The current practices of correcting for chance agreement can serve as a reference point for creating simple but useful models of coder behavior under uncertainty for the purposes of the simulation. I want to compare two ways of modeling coders' pre-knowledge that they use to make a decision even if they do not recognize the true value they are supposed to measure.

Equality-distribution (ED) guessing. In the first version, coders are relatively ignorant: they know nothing about a variable except how many scale points the measurement scale has (as described in the code book). When they do not recognize the true value and have to guess, they would simply pick any of the scale points on the measurement scale, with equal probability. I call this: equal-distribution (ED) guessing. Each scale point has the same chance of showing up (equal distribution), similar to a (non-loaded) dice. This certainly oversimplifies coder behavior because coders could e.g. have systematic preferences for particular scale points such as the center of the scale. Zhao et al. (2012) interpret the ways in which coder agreement coefficients adjust for chance agreement as implicit models of coder behavior when guessing. In this view, the scale-based guessing modeling of coder behavior is related to how Brennan and Prediger's $\varkappa$ (and equivalent coefficients) adjusts for chance agreement. In Bayesian terms, the measurement scale serves as an information-poor prior, the ED prior.

Marginal-distribution (MD) guessing. In the second version, coders have a strong preconception of how frequent each value on the scale will show up; this may represent their (good) intuition, experience with coding, or learning during the coder training or during the coding process. When guessing, they would prefer some values and choose them more often. In fact, the simulation assumes that coders have a perfect sense of how frequent which values are in the population studied. When not recognizing the true value, the probability of picking the respective value equals that value's relative frequency in the true marginal distribution (known in the simulation). This is to some extent unrealistic because coders cannot have perfect knowledge of the true marginal distribution, and obtaining a good estimate of the marginal distribution is one of the key goals of many content analyses. Often, we will overestimate the information coders have and can utilize in their guessing. This can be thought of as a "worst-case" scenario that is a useful comparator: what would happen if coders had such correct pre-conceptions? I call this: marginal-distribution (MD) guessing. According to Zhao et al. (2012), the marginals-based mode of coder guessing corresponds to the model of coder guessing behavior implied by how Fleiss $\varkappa$ and Krippendorff's α (and equivalent coefficients) adjust for chance agreement. In Bayesian terms, the true marginal distribution serves as an information-rich prior, the MD prior. To be sure, many other priors would be possible, e.g. using the mode of the distribution as a prior—which would also be information-rich. This study concentrates on ED and MD as simple and widespread notions of coder behavior under uncertainty.

Modeling coder behavior versus adjusting for chance agreement. Usually, intercoder agreement coefficients have no explicit model of coder behavior. Attempts to interpret them in terms of a "latent" or "implicit" model of coder behavior (Feng, 2013; Zhao et al., 2012) have received severe criticism (Krippendorff, 2012, 2016) on grounds that statistical control of chance agreement does not imply a model of coder behavior. Krippendorff (2012) views the procedure as a powerful statistical control independent of any assumptions about coder behavior, while Zhao et al. (2012) maintain that the statistical procedures imply a model of coders' behavior under uncertainty. This debate, fierce as it has been, is of little practical consequence for the simulation I have been running, except for the fact that it is useful to include multiple models of coder guessing behavior for comparison.

RQ2: How do different models of coder guessing under uncertainty affect the simulated precision of estimates and the testing of REHs?

Method

Number of simulation runs and software implementation

The simulation uses (1) two *modes of coder behavior* under uncertainty (MD, ED); (2) 21 *levels of coding accuracy* (0.00, 0.05, 0.10, ... 0.90, 0.95, 1.00); (3) 12 *sample sizes* (10, 20, 30, 40, 50, 100, 200, 300, 400, 500, 700, 1000; sum of all sample sizes=3350); and (4) ten *effect sizes* (correlations) (true value: .75, .50, .40, .30, .40, .30, .20, .15, .10, .05; realization value: .70, .46, .32, .29, .36, .25, .19, .18, .09, .07) between five variables. The true values were the target values that the random number generation was supposed to reach. The deviation between "true values" and "realization values" of correlations stems from the fact that the data were generated in a random process which induces some "noise" into the realizations. Additional deviations may be induced if the predefined correlation matrix is "impossible" in the sense that setting a target correlation between variable A and B decreases the degrees of freedom for reaching the target correlation between A and C and between B and C. All variables were, for simplicity's sake, generated as random numbers from a normal distribution which was mapped on a typical 7-point scale (1–7). An analogous binary coding task simulation displayed similar results (Online Appendix).

Two coders' decisions were simulated, and each scenario was calculated 1000 times. So there is 2×21×12 = 504 scenarios, 504×10 = 5040 relationships, 5040×1000=5 040 000 correlation estimates, 3350×2×21×5×2 = 1 407 000 coding decisions per run, and 1 407 000 × 1000=1 407 000 000 simulated decisions. Results (i.e. distributions of correlation estimates) are distributions within one scenario and are based on 1000 simulation runs that would equate 1,000 empirical studies with the respective sample size. All computations were run in *R* (R Core Team, 2015). Random data with prespecified correlations were generated done using the packages *mvtnorm* (Genz et al., 2015) and *GenOrd* (Barbiero & Ferrari, 2015).

Varying levels of coding reliability

The coding reliability Q of both coders is perfect at Q=1.00 (perfect recognition, full substantial agreement between both coders) and is arbitrary at Q=0.00 (pure "guessing", only "chance agreement" between both coders). The intermediate levels of Q are characterized by a mixture of recognition and guessing (according to the statistical model of guessing / chance agreement). For example, Q=.30 would mean 30% recognition and 70% guessing. Mixtures of guessing and recognizing were modeled in two ways: (a) by picking either the true or the guessed (or randomly drawn) value with the probability pre-defined by accuracy (coding reliability) levels in each decision or (b) by computing a weighted mean of guessed and true value in each decision, and round it to the next scale-point. Both mixing procedures led to very similar results such that this study reports only the data generated according to method (b).

Varying conceptions of chance agreement

MD guessing is implemented by listing the true values of all items to be coded (true marginal distribution). A coder who guesses picks one of these values and assigns it to the unit. This is drawing with replacement. ED guessing is implemented by listing the available values in a limited scale (e.g. a seven-point scale). A coder who guesses randomly picks one of these values, with equal probability for each of the different values, and assigns it to the unit. This is also drawing with replacement.

Measures

Correlations between of interval-scaled variables are estimated with Pearson's product-moment correlation formula. The null hypothesis (t-test of correlation) is ϱ=0. *R* is the estimate of that true correlation. Each correlation coefficient is replicated 1,000 times (1,000 simulation runs) and "confidence ranges" are the .025 and .975 quantiles of the distribution of the simulated estimates.

Reliable Detection of Effects: Type I and Type II Errors

Each simulation scenario has 1,000 replications for each correlation. I report the median value along with the range after cutting off the replications with the 25 highest and the 25 lowest values. This range equals a 95% confidence interval.

Type II errors. If the H_1 is true (and one can only make type II errors by definition), the type II error probability β falls below 5% as soon as the lower limit of this confidence range surpasses the critical value for a significant (.05, .01, .001) finding in a single study. So an effect (described by the H_1) validly becomes detectable once the *lower limit* of the confidence range climbs above the critical value. This is desirable, and coding reliability should be high enough to reliably detect effects in this way.

Type I errors. If the H_0 is true (and one can only make type I errors by definition), the type I error probability α increases above 5% as soon as the upper limit of the confidence range surpasses the critical value for a significant finding in a single study. So an inexistent effect (described by the H_1) is falsely “discovered” in more than 5% of studies once the *upper limit* of the confidence range surpasses the critical value. Such findings would be highly problematic.

Results

The results are clear-cut regarding the mechanisms hypothesized: Higher effect sizes, higher sample sizes, and higher coding accuracy all reduce the bias in correlation estimates, and the bias is generally negative. Figure 1 shows the correlation coefficients obtained, the shaded area designates the range of correlation estimates, cutting off the upper and lower 2.5% of the distribution, resulting in a 95% confidence region.

Effect size

In line with H-1, random errors by coders skewed the estimates of effect size towards zero, leading to a lower probability of refuting the null hypothesis. If coder errors can be viewed as random measurement error, the errors will be conservative and unfairly favor the null hypothesis over the alternative hypotheses rather than the other way around. In effect, the rate of type II errors increases. The risk of type I errors remains stable (Figure 2): In case of the nonexistent correlation (random number generation resulted in R=0.07 even though the target correlation behind the random number

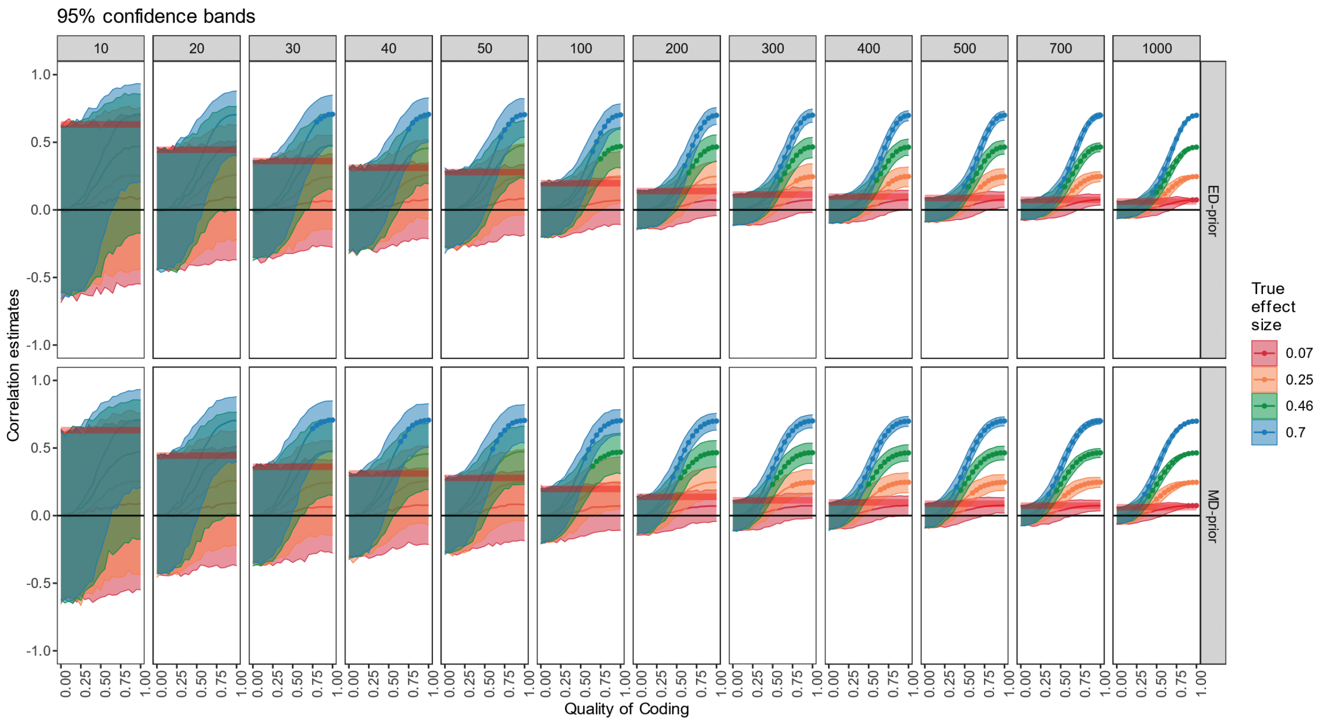


Figure 1: 95% of simulated study results fall into the shaded area. Dots represent the median estimate of the scenarios where the 97.5% of all simulation results were above the threshold for a statistically significant correlation (p<.05). These scenarios satisfy α<.05 and β<.05 under a true alternative hypothesis (i.e. they allow for reliably detecting a correlation).

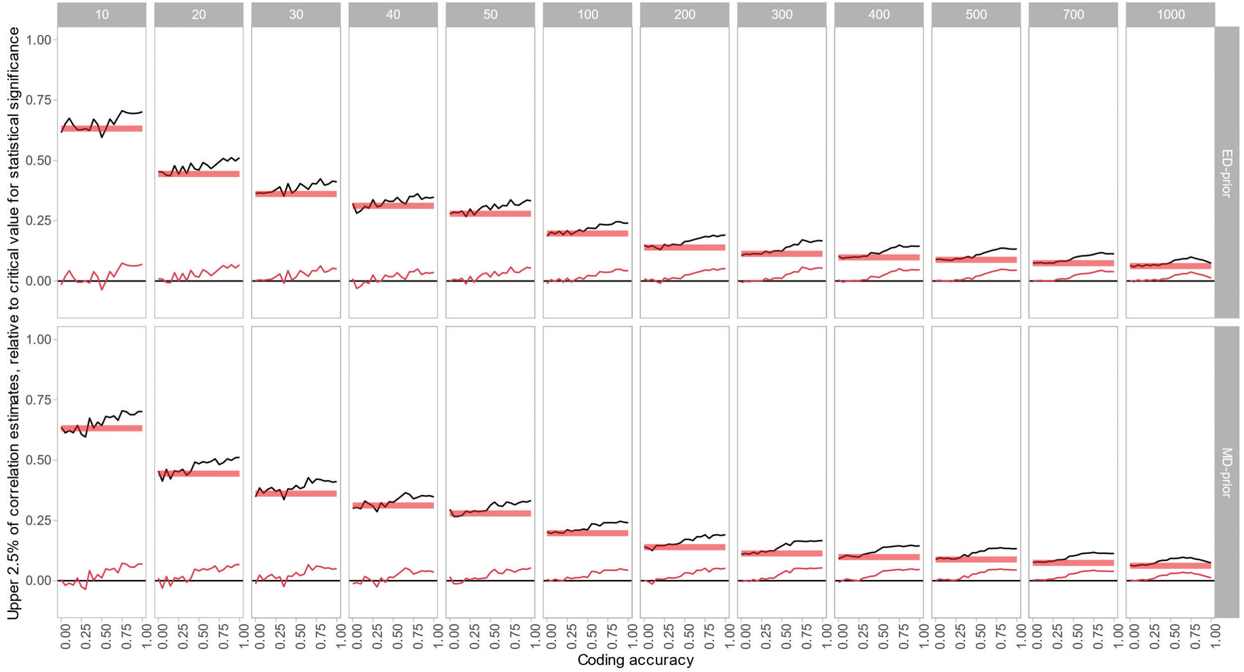


Figure 2: Probability of type I or α errors are represented by the y-distance between the black line and the red bar (or between the red line and the zero line). If the line is above the bar, the type I error rate will be above 5% (two-sided); if it is on the line, it is exactly 5% (two-sided); if it is below, the type I error rate is less than 5% (two-sided). Overall, there are no indications of systematically exceeded risks of making type I errors. A slight increase at very high coding accuracy traces back to relatively high "true correlation". Example: at sample sizes of 10 with an ED prior and a coding accuracy of .75 or higher (top left panel), the 2.5% highest correlations are at around R=0.700 and slightly above the critical value of .632

generation was set to 0), the highest 2.5% of estimates in the simulation varied around the significance limits for a study of the respective sample size. If measurement errors can be treated as random and proper statistical tests are selected, content analysts do not have to worry about potentially spurious positive findings caused by coding reliability problems. The greater risk is potentially spurious non-findings: Decisions for the null hypothesis will remain ambiguous under low coding reliability; but decisions for the alternative hypothesis despite low coding reliability are likely to mean something. This finding is valid for all effect sizes, sample sizes and for both coder guessing models.

H-2 is also supported (Figure 1). For a visual inspection, let us keep the sample size constant and look at the scenario with n=300. The effect size of .70 (“very strong”) is reliably detected ($\alpha<.05$ and $\beta<.05$) at accuracy=.45 (MD-coding) and .60 (ED-coding). The effect of 0.46 (“strong”) is detected at .50 (MD) and .65 (ED). The effect of 0.25 (“moderate”) is detected at .75 (MD) and .80 (ED). The effect of .07 (“nonexistent”) never reached the level of reliable detection in this scenario. The pattern is similar at all sample sizes, such that we can conclude that larger effects are easier to detect and therefore even detectable at relatively high levels of noise (random error) induced by lack of coding accuracy.

Coding accuracy scores produced here almost perfectly translate to inter-coder agreement scores obtained from the simulated data. For the MD scenario, one can think of the coding accuracy levels (e.g. 0.75) as α_K, $\varkappa_F$, or $\varkappa_C$ coefficients of 0.75. The relationship is not 1 to 1, however: α_K, $\varkappa_F$, and $\varkappa_C$ values are lower than actual accuracy when accuracy is low or moderate; the difference is moderate, but systematic (Figure A2). For the ED scenarios, accuracy levels almost perfectly translate to $\varkappa_{BP}$ coefficients (Figure A2). Under ED, α_K, $\varkappa_C$ and $\varkappa_F$ are too pessimistic. Under MD, $\varkappa_{BP}$ is too optimistic.

Coding reliability

As predicted by H-3, the “inner” 95% of the simulation results develop gradually according to a sigmoid pattern (Figure 1): Moving rightward along the x-axis (i.e. coding accuracy is improving) has no or little impact at first; then, an acceleration sets in where growing coding accuracy strongly improves the accuracy of the estimates. The slope grows steeper. As coding accuracy further increases, saturation sets in because the median estimate approaches the true value; the curve flattens. Figure 1 illustrate these S-curves at different effect sizes, sample sizes, and models of coding. What we observe there is that the “takeoff” starts the earlier the greater the effect size is, and that in case of MD-coding, the “takeoff” starts earlier than in the case of ED-coding. Earlier “takeoff” goes hand in hand with earlier “saturation”.

Sample size

H-4 receives empirical support. Visually, this is clear from the dotted medians in Figure 1, which represents scenarios in which a correlation is reliably detected ($\alpha<.05$ and $\beta<.05$). At similar levels of coding accuracy and effect size, sample size strongly affects the capacity of reliably discovering effects. For instance, the strong effect size (R=0.70) under MD-coding is reliably detected at Q=0.75 for n=30. In smaller samples, even this strong effect is not reliably detectable at any level of accuracy; for n=50, Q≥.60 is needed; for n=200, Q≥.45 is needed; at n=1000, Q≥.35 may suffice.

Coder behavior under uncertainty

At the same sample size and level of coding accuracy, MD-coding allows for detecting smaller effect sizes than ED-coding (Figure 1). Practically, this means that if we use coder agreement coefficients that are good indicators of coding accuracy under ED-coding (e.g. $\varkappa_{BP}$), we need relatively high coder agreement to make a specific inference (if sample size and effect size are constant). If we use coder agreement coefficients that are good indicators of coding accuracy under MD-coding (e.g. $\varkappa_{F}$ or α_{K}), lower coder agreement is satisficing for the same inference.

That fits the impression that, usually, coding agreement scores (reaching between –1 and +1) are “higher” or “more favorable” with “ED-like” coefficients (like $\varkappa_{BP}$) compared to α_{K} and other coefficients that inspired the MD-coding simulation. So, while showing seemingly more favorable assessments, using $\varkappa_{BP}$ also warrants higher levels of agreement to reach the same inferential capacities compared to a score obtained, e.g., in the α_{K} framework. So α_{K} only seems “stricter” and $\varkappa_{BP}$ seems more “relaxed” if one does not consider that the “key” to interpreting the scores should be “stricter” for $\varkappa_{BP}$ than for α_{K}. This makes sense, because MD-coding uses an information-rich prior that mixes more “true” information into chance coding than ED-coding does. The coefficients $\varkappa_{BP}$ and α_{K} are just different and warrant different interpretations despite them being mapped on similar scales.

Interactions between sample size, effect size, and coding accuracy

The lower limits of the confidence region of correlation estimates (i.e. the 0.025 quantile) are most crucial because as soon as they surpass the critical value for a correlation to be regarded statistically significant (e.g. at α<.05), the statistical power of the scenario surpasses 0.95—the scenario allows reliable detection of correlations. Nonlinear effects were expected, and the data were probed as to which degree of polynomials should be considered to explain the simulation results.

Sample size (6th degree polynomial; this function describes the logarithmic impact of sample size) explained 49.4% of variation. Coding accuracy (3rd degree polynomial; this polynomial serves to approximate the S-curves) explained 25.2% of variation. Effect size (1st degree polynomial; this reflects the linear impact of effect sizes) explains 12.5% of variance. Effect size and coding accuracy powerfully interact (adding 9.9% of explained variance), whereas sample size's effect was completely independent of the other two factors. Overall, the polynomials and their interaction explained 97.7% of variation in lower limits of the confidence interval (see Table A2 for details).

From simulation results to equations

The expected median correlation—as a function of effect size [abs(ρ)], and coding accuracy [Q] (but not he sample size [n])—is estimated using the following sigmoid function:

$$(1)\quad Mdn\,(r) = \frac{1 \cdot abs\,(\rho)}{1 + e^{(-(b+d\cdot[abs(\rho)]))\cdot([Q]-c)}}$$

The range of the 95% confidence bands around that median value was estimated separately for the upper and the lower bound, as a function of coding accuracy [Q] and sample size [n] (but not the effect size [abs(ρ)]):

$$(2)\quad \mathrm{LCB(r)}_{95\%} = (\mathrm{b} + 1 - [\mathrm{Q}]) \cdot \frac{\mathrm{a}}{\sqrt{[\mathrm{n}]}} + \frac{\mathrm{d} \cdot [\mathrm{Q}]^2}{[\mathrm{n}]} + \mathrm{c} \cdot [\mathrm{n}] \cdot [\mathrm{Q}]^2$$

$$(3)\quad \mathrm{UCB(r)}_{95\%} = (\mathrm{b} + 1 - [\mathrm{Q}]) \cdot \frac{\mathrm{a}}{\sqrt{[\mathrm{n}]}} + \frac{\mathrm{d} \cdot [\mathrm{Q}]^2}{[\mathrm{n}]} + \mathrm{c} \cdot [\mathrm{n}] \cdot [\mathrm{Q}]^2$$

These procedures approximate the median simulation results with R^2=.985/.998 (ED-coding: b=8.895,c=0.576,d=–0.119; MD-coding: b=8.904, c=0.495, d=–0.074), the lower confidence bound with r^2=.985/.993 (ED-coding: a=-0.250, b=7.178, c=0.00005, d=–1.087; MD-coding: a=-0.270, b=6.546, c=0.00006, d=–1.193) and the upper bound with r^2=.982/.984 (ED-coding: a=0.336, b=5.007, c=–0.00004, d=–0.851; MD-coding: a=0.410,

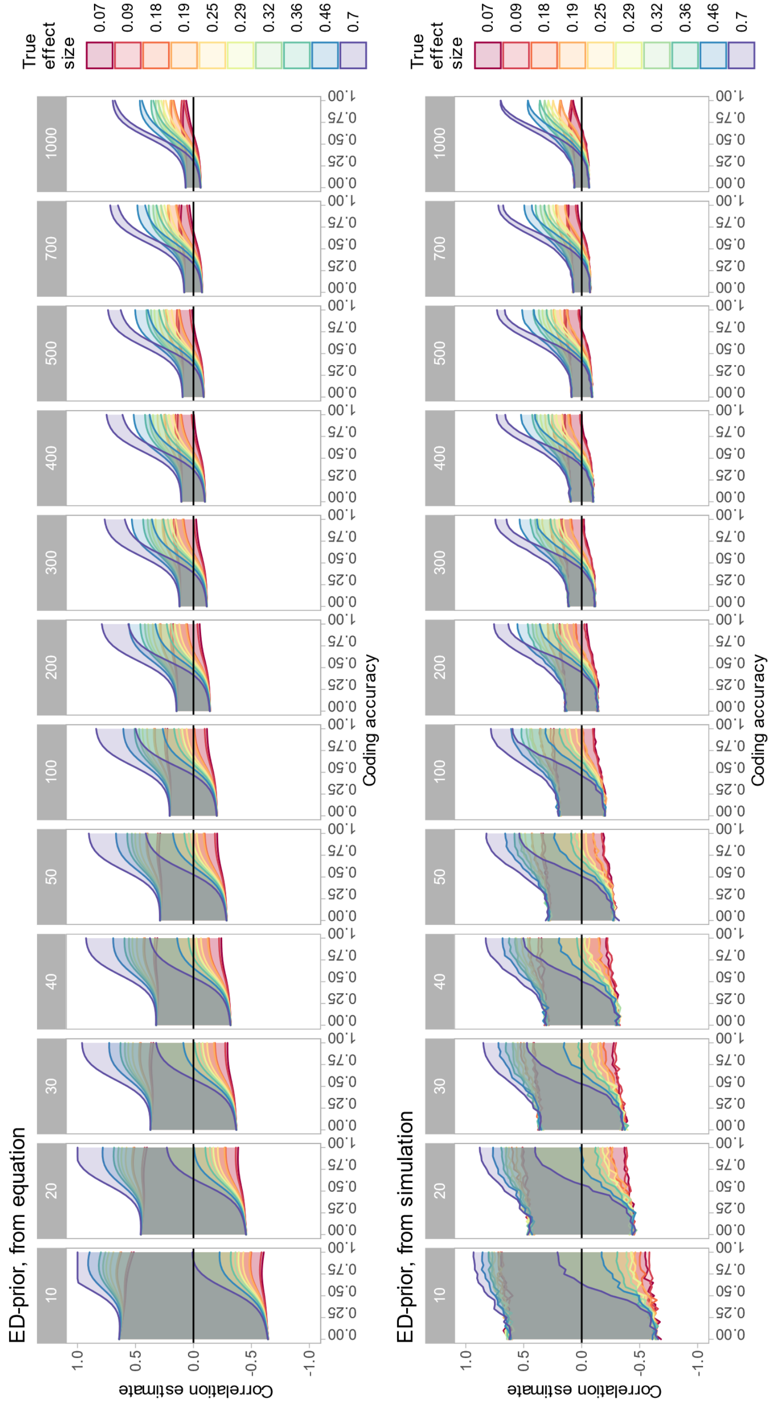


Figure 3: Comparison of modeled and simulated 95% confidence bands (ED-coding).

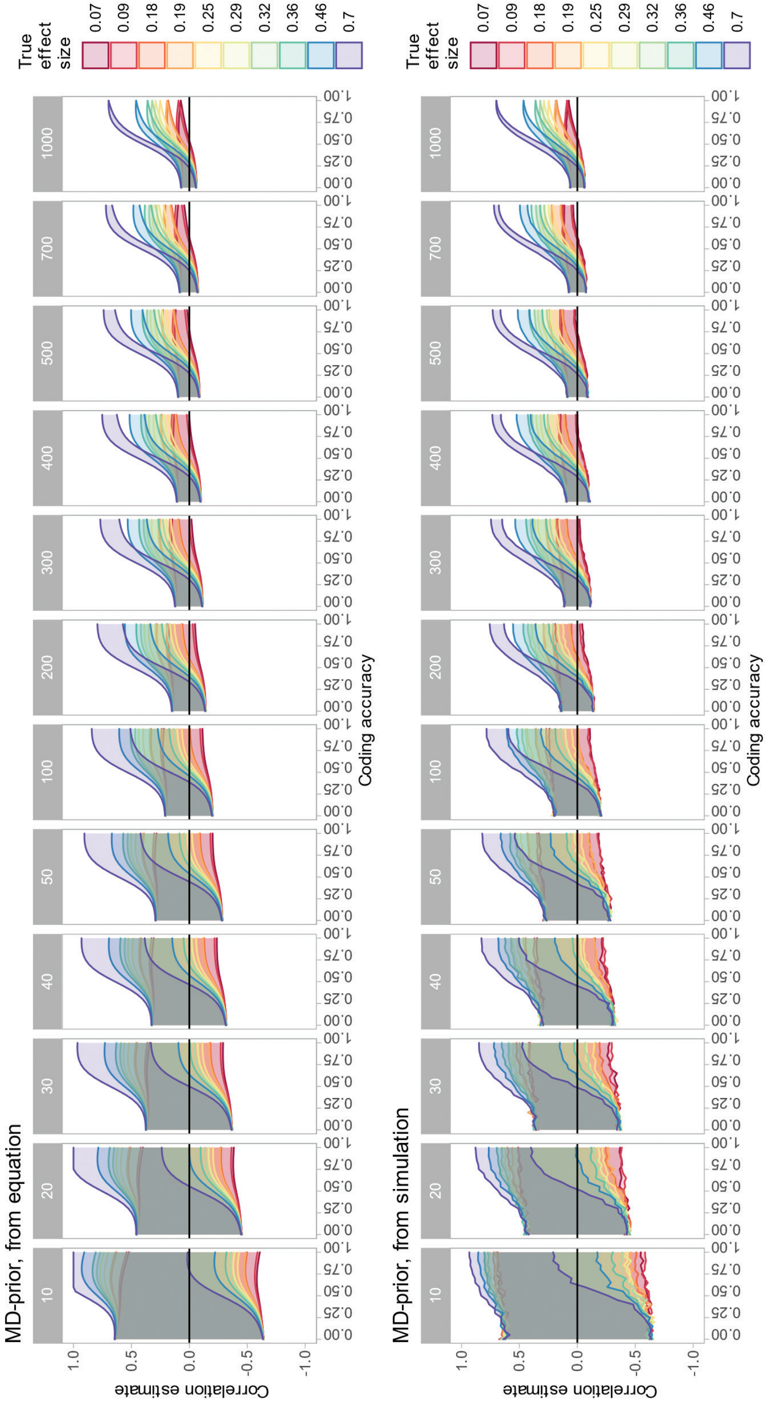
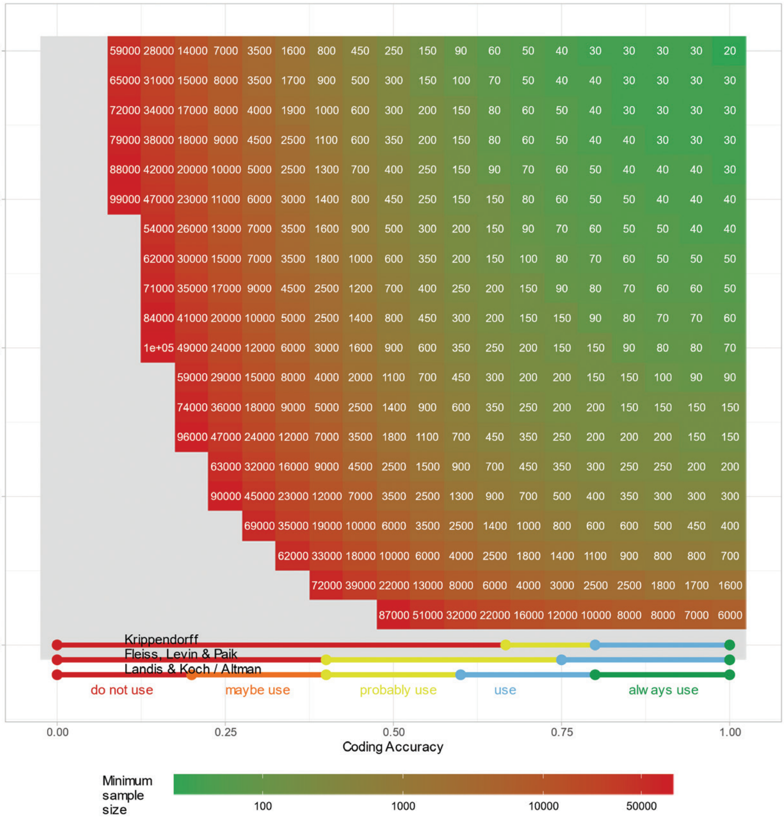


Figure 4: Comparison of modeled and simulated 95% confidence bands (MD-coding).

b=3.914,c=−0.00005,d=−0.552). Figures 3 (ED) and 4 (MD) illustrate how closely the equations reproduce the simulation results. We can safely use these equations to estimate the percentiles of the sampling distribution of correlations created by content analysis studies with a specified coding accuracy, sample size, and effect size.

Figures 5 (ED) and 6 (MD), based on the equations, specify which sample sizes are needed to reliably detect an effect at α<.05 and β<.05. With estimates of the true effect size ρ (based on the sample correlation R), and the true coding accuracy Q (based on inter-coder reliability indexes), one can look up the minimum necessary sample size. As a comparison, it also displays the thresholds proposed by Krippendorff (2004a), Landis and Koch (1977), Altman (1991), and Fleiss et al. (2003).



59000	28000	14000	7000	3500	1600	800	450	250	150	90	60	50	40	30	30	30	30	20
65000	31000	15000	8000	3500	1700	900	500	300	150	100	70	50	40	40	30	30	30	30
72000	34000	17000	8000	4000	1900	1000	600	300	200	150	80	60	50	40	30	30	30	30
79000	38000	18000	9000	4500	2500	1100	600	350	200	150	80	60	50	40	40	30	30	30
88000	42000	20000	10000	5000	2500	1300	700	400	250	150	90	70	60	50	40	40	40	30
99000	47000	23000	11000	6000	3000	1400	800	450	250	150	150	80	60	50	50	40	40	40
	54000	26000	13000	7000	3500	1600	900	500	300	200	150	90	70	60	50	50	40	40
	62000	30000	15000	7000	3500	1800	1000	600	350	200	150	100	80	70	60	50	50	50
	71000	35000	17000	9000	4500	2500	1200	700	400	250	200	150	90	80	70	60	60	50
	84000	41000	20000	10000	5000	2500	1400	800	450	300	200	150	150	90	80	70	70	60
	1e+05	49000	24000	12000	6000	3000	1600	900	600	350	250	200	150	150	90	80	80	70
		59000	29000	15000	8000	4000	2000	1100	700	450	300	200	200	150	150	100	90	90
		74000	36000	18000	9000	5000	2500	1400	900	600	350	250	200	200	150	150	150	150
		96000	47000	24000	12000	7000	3500	1800	1100	700	450	350	250	200	200	200	150	150
			63000	32000	16000	9000	4500	2500	1500	900	700	450	350	300	250	250	200	200
			90000	45000	23000	12000	7000	3500	2500	1300	900	700	500	400	350	300	300	300
				69000	35000	19000	10000	6000	3500	2500	1400	1000	800	600	600	500	450	400
					62000	33000	18000	10000	6000	4000	2500	1800	1400	1100	900	800	800	700
							72000	39000	22000	13000	8000	6000	4000	3000	2500	2500	1800	1700	1600
								87000	51000	32000	22000	16000	12000	10000	8000	8000	7000	6000

Figure 5: ED-prior. Minimum Sample Sizes to Detect an Effect with β<.05 at α<.05, as a function of the true effect size and coding accuracy. Inter-coder reliability coefficients are estimates of coding accuracy.

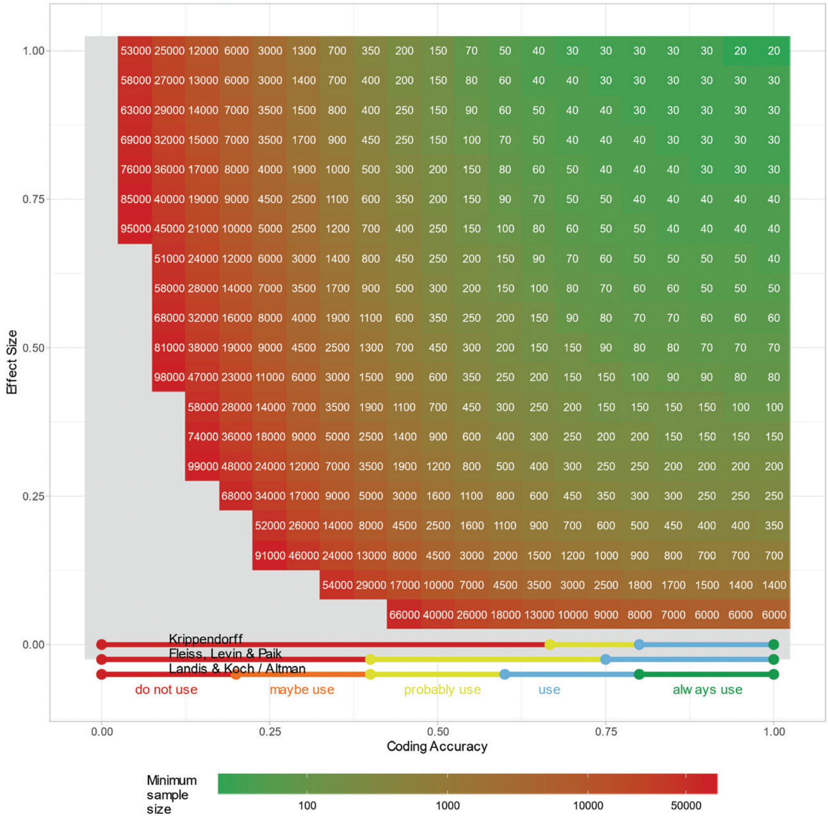


Figure 6: MD-prior. Minimum Sample Sizes to Detect an Effect with β<.05 at α<.05, as a function of the true effect size and coding accuracy. Inter-coder reliability coefficients are estimates of coding accuracy.

If coding accuracy is above .80, one has sufficient statistical power in most research settings. Sample sizes of 500 (ED) or 450 (MD) are enough to detect correlations as low as .20; and sample sizes of 100 are enough to detect effect sizes of .45. So, that reliability tests should have values of .800 is a reasonable starting point, though very low effect sizes (e.g. ρ=.05) necessitate extremely large sample sizes to allow reliable discovery.

Discussion

Observations and interpretations

Within the boundaries of the approach presented here, several important conclusions can be drawn immediately.

Krippendorff's (2004) *interpretation guideline is useful in almost all settings. With very small effect sizes, even Krippendorff's supposedly "strict" guidelines prove too liberal. In very large samples coupled with large effect sizes,* Fleiss et al.'s (2003) *more liberal guidelines can be applied.*

In most studies, Krippendorff's suggestion to use data with intercoder agreement $\alpha \geq .80$ or, more risky, $\alpha \geq .667$, converges with this study's simulation results in many rather typical study settings. The interpretations put forth by Landis and Koch (and adopted by Altman) appear too liberal for effective hypothesis testing. Reliability between .20 and .40 is mostly worthless even if sample sizes and effect sizes are very high. Fleiss' interpretations are applicable for studies that couple large sample sizes with large effect sizes.

Of more "liberal" coefficients and "stricter" cutoff limits

Coefficient-specific interpretation guidelines should be applied; the more information a coefficient considers, the more inferences are possible even at lower values of the coefficient. If appropriate interpretation guidelines are used, the conception of chance agreement does not fundamentally affect inferences.

The comparison between two scenarios of coder behavior under uncertainty (MD-coding and ED-coding) shows that the different conceptions of chance agreement are probably not as consequential as the fierce debate in the literature (Feng, 2013; Feng & Zhao, 2016; Krippendorff, 2012, 2016; Zhao et al., 2012) suggests. It is true that α_K values will usually be lower than $\varkappa_{BP}$ values; the reason is that more information (about the underlying distribution) is interpreted in terms of (potential) chance agreement rather than substantial agreement. If one obtains $\alpha_K = .50$, one can draw better inferences than if $\varkappa_{BP} = .50$, for example, ceteris paribus.

The joint impact of coding accuracy, sample size, effect size and MD/ED-coding

The probably most important findings are the exact curves that describe how correlation estimates (and their relation to the true value) respond to decreasing coding accuracy—depending on models of coding behavior under uncertainty, sample size, and effect size.

- *Correlation estimates converge towards the true value with increasing coding accuracy. Its onset and speed varies with other properties.*

- *The convergence is not linear, but S-shaped.*
- *The onset point for the convergence process varies with effect size: the larger the effect size, the earlier (i.e. at lower levels of coding accuracy) does the convergence start. The deceleration also sets in at lower levels of coding accuracy. The approximation is "stretched out" more.*
- *MD-coding leads to earlier onset of convergence compared to ED-coding. This also leads to faster "narrowing" of confidence bands.*
- *Confidence bands are slimmer with greater sample sizes and with greater coding accuracy.*

Coding accuracy does not affect type I error rate. Type I error rates do not increase with lower accuracy if error is random rather than systematic. This means that the main mistakes one must consider under these conditions are twofold. One: Independent of any cutoffs, we will usually underestimate effect sizes if coder agreement is not perfect, and each ever-so-small decrease in coding accuracy (reflecting in lacking agreement) will further depress the effect sizes found. Two: We are more likely to overlook correlations that in fact exist but that are not strong enough to show as a data signal because there is too much noise and the statistical power is too low. Hence, not finding a hypothesized correlation at low levels of coder agreement, small sample sizes, and potentially small effects render the data useless. The data are not suited to distinguish between small correlations and non-correlations. In contrast, if coder agreement is high enough, sample size is decent, and effect size is at least moderate, such a null finding can more safely be interpreted in terms of rejecting the hypothesis.

This means that effect size plays an important part in what one can find in a content analysis—while theory sections and rationales leading up to hypotheses rarely delve into the topic of effect sizes in content analyses. When designing content analyses, we should definitively take expected effect sizes into account, and maybe also experience from previous studies or pretests regarding the level of coding accuracy that can be obtained in a measurement. This might be a way to improve hypothesis testing for hard-to-measure constructs by choosing an appropriate sample size.

A notice of caution

The simulation results are bound to the assumption that errors in coding are random errors rather than systematic errors. Furthermore, the simulation presupposes that content analysts try to test hypotheses or relationships using classical inferential statistics. For studies with a more descriptive focus or used in data combination with survey results (Scharkow & Bachl, 2017; Schuck et al., 2015), for example, the arguments cannot be applied fully. Let us examine how the simulation results can be useful in other common applications:

Descriptive and exploratory analyses of content analysis data. In short, with some limitations, the simulation results can be informative for many other study contexts that involve content analysis data. The relationship between effect size, sample size, and coding accuracy applies to all kinds of content analysis data collected; the only complication is that in exploratory and descriptive studies, there is no discrete test result for which a certain level of statistical power can serve as criterion. Generally, coding accuracy will always reduce the bias of point estimates according to a sigmoid function. Greater sample sizes will decrease the variability of effect size estimates across studies. One can, however, define the lowest effect size for which a conclusive hypothesis (at $\alpha<.05$ and $\beta<.05$) test is possible.

Mixing content analysis data with data from other sources. The simulation has focused on two content analytic measurements that are analyzed for correlations. But what about combinations between one content analytic measurement and measurements collected with other methods? This case is somewhat problematic because inter-coder agreement data are not immediately compatible with other modes of assessing data quality. Still, the analyses presented here can be applied to the content analytic measure under the assumption that the other measure (collected with another method) is at least measured at the same level of accuracy as the content analysis measure. We would then independently check the other measure's accuracy (e.g. using internal consistency measures in surveys).

Semi-automated content analyses. Despite the rise of automated content analysis procedures, manual content analyses are still vital in the field. But assessing the performance of human coders is also important in semi-automated methods which have gained substantially in relevance and popularity. Here, human coders generate (a) training data to feed into machine-learning algorithms and (b) validation data to assess the performance of the semi-automated coding (Song et al., 2020). Here, it is pivotal that high coder agreement in the human coding is established before using it as a benchmark (Song et al., 2020). If that is established, one can compare the criterion data created by human coders with those generated by the trained algorithm to assess the effectiveness of the training in replicating human coding. This also means that the calculations presented here are important for designing semi-automated content analysis studies; in particular, the scalability of semi-automated data gathering can allow for very large sample sizes that can to some degree compensate for substandard coding accuracy. The equations presented here can help in making plausible choices.

Recommendations

There are some quite general recommendations that I can suggest based on the findings.

- *Consider coder agreement, sample size, and effect size in conjunction.* If possible with reasonable efficiency, increasing reliability and validity and thereby coder agreement is the traditional and often the most efficient way of providing more power for hypothesis testing; but increasing sample size can be a good addition or a viable alternative, particularly if coding accuracy cannot be improved further but is substantially above chance. If the main interest is not testing hypotheses but finding accurate effect size estimates, improving inter-coder agreement is the only way to go and adjusting sample size is not viable.
- *Make sure to rule out joint systematic errors as well as possible.* Using multiple coders and distributing material among them randomly (or in a way with similar effect) will not pollute particular groups of cases or variables more than others. Coder training may induce systematic errors into the whole group, which may lead to situations with high agreement but poor validity. So be careful in coder training not to make general prescriptions that hurt validity only to safeguard high agreement.
- *Form expectations regarding "true" effect sizes and the level of coding accuracy you can achieve.* Effect size plays an important part as to whether it is possible to test hypothesized correlations or not. In our theorizing, we should be more explicit about what size of correlation we expect, e.g. based on what sizes of correlation previous studies have found. For very small correlations, the recommended minimum coder agreement values (e.g. $\alpha_K \geq .800$) are not strict enough. For very strong correlations, they may be relaxed a little. The formulae, tables and scripts presented in this study provide a better foundation for discussing whether data are

appropriate for testing a hypothesis, but only if we have a rough expectation regarding the effect size.

- *Use expected coding accuracy and effect size when designing your study, and preregister your design.* If one actually anticipates that an effect size will be very low and/or that coding accuracy will be relatively low (hard-to-measure constructs), one can use that to design the study such that the hypothesis can still be tested—by boosting sample size. Such design plans become even more plausible and useful if they are pre-registered, documenting that these expectations were formed a priori. This paper (equations, tables, R script) can and should be used when designing content analyses, and the information should be included in publications as well as preregistrations.
- *Additional criteria for reviewers assessing the appropriateness of content analysis results.* The fixed benchmarks for evaluating inter-coder agreement results remain important. Additionally, one can and should consider the constellation of effect size, sample size and coding accuracy and whether they allow a reliable detection of a correlation. This is relatively uncritical if expectations regarding coding accuracy and effect size are documented in a preregistration. If there is no preregistration, reviewers should be careful as authors may instrumentally think up exceeded effect size expectations to justify insufficient coder agreement.

Outlook

There are additional topics to tackle in the future that this study was not able to solve or focus on in an appropriate manner. Important additions would be to introduce variability of coding accuracies for different variables in the same simulation run. Currently, the study assumes that in a simulation run all variables are measured at the same level of accuracy.

Second, it is important to include different kinds of non-normal distributions of the underlying data. Non-normality of distributions is a common problem in the social sciences that content analysts should be able to consider.

Third, future research should try to incorporate different kinds of systematic coding errors into the simulation. These would have to be regarded as ideal types, such as “individual-specific idiosyncracies” (e.g. one coder tends to give more favorable ratings), “group-specific idiosyncracies” (e.g. left-leaning coders rate left-leaning parties and candidates more favorably) and “common biases” (e.g. all coders rate all politicians too favorably) that blend into an overall mixture of systematic errors.

Fourth, scholars should study the effects of coding accuracy and other factors on study results in more complex analyses such as regressions or mixed models. In the same vein, delving into the case of data linking (Scharkow & Bachl, 2017) appears promising.

Notes

[1] I use “coder agreement“ to subsume inter-coder, intra-coder and researcher-coder agreement. I prefer “agreement” over “reliability” because “agreement” is more specific and less prone to causing confusion. “Coder agreement” and “coding reliability” are treated as synonyms.

[2] Improving coder agreement leads to less-biased point estimates, however.

References

Altman, D. G. (1991). *Practical statistics for medical research*. Chapman and Hall.

Barbiero, A., & Ferrari, P. A. (2015). *GenOrd: Simulation of Discrete Random Variables with Given Correlation Matrix and Marginal Distributions* (1.4.0) [Computer software]. http://CRAN.R-project.org/package=GenOrd

Berelson, B. (1971). *Content analysis in communication research*. Hafner.

Chan, W., & Chan, D. W.-L. (2004). Bootstrap Standard Error and Confidence Intervals for the Correlation Corrected for Range Restriction: A Simulation Study. *Psychological Methods*, *9*(3), 369–385. https://doi.org/10/dqvthj

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed). L. Erlbaum Associates.

Feng, G. C. (2013). Factors affecting intercoder reliability: A Monte Carlo experiment. *Quality & Quantity*, *47*(5), 2959–2982. https://doi.org/10.1007/s11135-012-9745-9

Feng, G. C. (2014). Intercoder reliability indices: Disuse, misuse, and abuse. *Quality & Quantity*, *48*(3), 1803–1815. https://doi.org/10.1007/s11135-013-9956-8

Feng, G. C., & Zhao, X. (2016). Do Not Force Agreement: A Response to. *Methodology*, *12*(4), 145–148. https://doi.org/10/gdqc5m

Fico, F. G., Lacy, S., & Riffe, D. (2008). A Content Analysis Guide for Media Economics Scholars. *Journal of Media Economics*, *21*(2), 114–130. https://doi.org/10.1080/08997760802069994

Fleiss, J. L., Levin, B., & Paik, M. C. (2003). *Statistical methods for rates and proportions* (3rd ed). Wiley.

Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F., & Hothorn, T. (2015). *mvtnorm: Multivariate Normal and t Distributions* (1.0-3) [Computer software]. http://CRAN.R-project.org/package=mvtnorm

Gustafson, P. (2004). *Measurement error and misclassification in statistics and epidemiology: Impacts and Bayesian adjustments*. Chapman & Hall/CRC.

Gwet, K. L. (2014). *Handbook of Inter-Rater Reliability* (3rd ed.). Advanced Analytics.

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. https://doi.org/10/cs2t97

Kepplinger, H. M. (1989). Content Analysis and Reception Analysis. *American Behavioral Scientist*, *33*, 175–182. https://doi.org/10/cczw6m

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

Krippendorff, K. (2004b). Reliability in Content Analysis: Some Common Misconceptions and Recommendations. *Human Communication Research*, *30*(3), 411–433. https://doi.org/10.1111/j.1468-2958.2004.tb00738.x

Krippendorff, K. (2012). Comment: A dissenting view on so-called paradoxes of reliability coefficients. In C. T. Salmon (Ed.), *Communication Yearbook* (Vol. 36, pp. 481–500). Routledge.

Krippendorff, K. (2016). Misunderstanding Reliability. *Methodology*, *12*(4), 139–144. https://doi.org/10.1027/1614-2241/a000119

Krippendorff, K. (2017). Three concepts to retire. *Annals of the International Communication Association*, *41*(1), 92–99. https://doi.org/10/gf659g

Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, *33*(1), 159. https://doi.org/10.2307/2529310

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. https://doi.org/10.1111/j.1468-2958.2002.tb00826.x

R Core Team. (2015). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. http://www.R-project.org/

Riffe, D., Lacy, S., & Fico, F. (1998). *Analyzing media messages: Using quantitative content analysis in research*. Erlbaum.

Scharkow, M., & Bachl, M. (2017). 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*, *34*(3), 323–343. https://doi.org/10/ggbm28

Schuck, A. R. T., Vliegenthart, R., & De Vreese, C. H. (2015). Matching Theory and Data: Why Combining Media Content with Survey Data Matters. *British Journal of Political Science*, 1–9. https://doi.org/10/gdqc3h

Song, H., Tolochko, P., Eberl, J.-M., Eisele, O., Greussing, E., Heidenreich, T., Lind, F., Galyga, S., & Boomgaarden, H. G. (2020). In Validations We Trust? The Impact of Imperfect Human Annotations as a Gold Standard on the Quality of Validation of Automated Content Analysis. *Political Communication*, *37*(4), 550–572. https://doi.org/10.1080/10584609.2020.1723752

van 't Veer, A. E., & Giner-Sorolla, R. (2016). Pre-registration in social psychology—A discussion and suggested template. *Journal of Experimental Social Psychology*, *67*, 2–12. https://doi.org/10/f85xtx

Zhao, X., Liu, J. S., & Deng, K. (2012). Assumptions behind inter-coder reliability indices. In C. T. Salmon (Ed.), *Communication Yearbook* (Vol. 36, pp. 419–480). Routledge.

About the Author

Stefan Geiß, Associate Professor
Department of Sociology and Political Science
Faculty of Social and Educational Norwegian University of Science and Technology (NTNU), Trondheim, Norway
stefan.geiss@ntnu.no