Analyzing Intercoder Agreement

19

In qualitative research projects, questions often arise about the intersubjectivity of the analysis. Given the same interview passage, for example, does my fellow researcher see the same topics addressed as I do, and do they draw the same conclusions? To what extent do we agree on our understanding of categories? With these questions we are entering the field of quality criteria, which should not be neglected in qualitative research. In category-based approaches, the focus is placed on the question to what extent two people identify the same topics, aspects, and phenomena in the data and assign these to the same categories. It is quite possible for two people to agree in terms of content, but assign different categories to a phenomenon, because the category definitions have not yet been clearly formulated. MAXQDA offers numerous (partly interactive) functions, which enable systematic analysis, improvement, and verification of the agreement between coders. Problematic categories, misleading instructions, and blurred category definitions can be identified to improve the quality of analysis step-by-step.

In This Chapter

- Getting to know the objectives and areas of application of agreement testing
- Understanding the procedures for testing intercoder agreement in MAXQDA
- Conducting document-level and segment-level agreement tests
- Calculating percentage agreement and setting useful thresholds
- Taking agreement by chance into account

U. Kuckartz, S. Rädiker, *Analyzing Qualitative Data with MAXQDA*,
https://doi.org/10.1007/978-3-030-15671-8_19

Objectives and Areas of Application for Analyzing Agreement

The extent to which different researchers produce the same result is an important quality criterion of empirical studies. Hence, qualitative research involves the question of intercoder agreement: "To what extent do different coders agree when assigning categories to texts, images and videos?" Checking and improving intercoder agreement is closely linked to the research tradition of qualitative content analysis and is considered a central factor of quality (Kuckartz, 2014; Mayring, 2014; Schreier, 2012). In qualitative methods that follow a more interpretative approach, the question of intercoder agreement rarely arises, or not at all, since categories play a subordinate role or even none at all as tools for analysis. Although research projects following the grounded theory approach involve intensive work with codes and categories, it is not common to perform any intercoder agreement tests in this context. The focus in grounded theory is explicitly on the continuous development of concepts and categories and not their application to the material according to precisely defined rules.

So, what exactly is the benefit of an agreement analysis and for what purpose is it used? Well, there are many answers to this: just as new employees in the control departments of Facebook have to learn the rules for assessing contributions to delete inappropriate ones, it is also the task of new coders in a research project to understand the guidelines for applying categories. A test of agreement with a model example of code assignments reveals to what extent a coder's training has been successful. Intercoder analysis can be used not only to determine the effects of training for coders but also to sharpen the category system and coding instructions. With the help of intercoder analysis, problems with individual categories and their definitions can be identified and reduced, problems concerning the delimitation of categories can be traced, and, in addition, coders can be identified whose work differs systematically from those of the others. Thus, intercoder analysis can be used equally as a tool for checking, improving, and ensuring the quality of coding processes.

If several people are to code the same data for the purposes of an agreement analysis, questions about how best to organize the workflow inevitably arise. In principle, you can distinguish between three separate variations:

Coding at Two Different Times Here the data is first coded by one person, and then their code assignments are checked by one or more people subsequently. For example, as a first step a student assistant codes the texts, the project manager then looks through their code assignments, makes corrections, and discusses doubtful cases with the assistant or in the research team. This procedure is only suitable in the situation where well-developed category definitions already exist. When working alone, sometimes it can also be helpful to repeat the coding process at a later stage; 2–4 weeks later, the same person checks their own coding work again, or the same person codes the data again without looking at the coding work done previously.

Simultaneous Collaborative Coding In the case of concurrent coding, the data can be discussed throughout the whole research team or in pairs of two, and suitable categories can be collectively assigned to data segments. This approach is particularly suitable for initial examinations of the data and the development or initial testing of coding frames. It is, however, exposed to the risk of the influence of research team hierarchies or the dominance of particularly extroverted personalities.

Simultaneous Independent Coding The most frequently used method involves researchers code the data independently of each other and then comparing their results. Independent coding is generally mandatory for the calculation of the percent agreement and for chance-corrected agreement coefficients such as Cohen's kappa or Krippendorff's alpha (see below). In our opinion, however, it is important and desirable for qualitative research that more than just coefficients of agreement are calculated and published. Rather, they should, together with the places where inconsistencies have occurred, form the basis for a systematic discussion of the inconsistencies and the consequences for the category system and coding instructions. Based on this aspiration for qualitative-oriented analyses, we consistently prefer the term "intercoder agreement" to that of "intercoder reliability" in this chapter. Reliability is one of the three classical quality criteria of quantitative research; it stands for the claim of accuracy and replicability of measurements and is mainly located within the context of quantitative content analysis. In addition, the transferability of classical quality criteria to qualitative research must be critically questioned (Kuckartz, 2014, pp. 151–155).[1]

It is striking that simultaneous independent coding is particularly suitable for carrying out *systematic* agreement analyses. The first two analysis variations above represent further ways of increasing the coding quality and can be profitably combined with the third.

MAXQDA offers specially developed functions that support the determination of intercoder agreement, the control of disagreements, and the improvement of agreement. We will describe these functions with regard to different coders and therefore use the term "intercoder" throughout this chapter. However, these methods can also be used in the case of repeated coding by one and the same person, which is always useful if you want to analyze the stability of your own coding work, the so-called intracoder agreement.

Before you can start analyzing intercoder agreement in MAXQDA, you must first clarify for which documents (and in which order) the analysis is to be carried out. If the amount of data to be analyzed is small, all the documents can be coded by a second person. This is the case, for example, if you have conducted ten half-hour interviews that have been thematically coded using a simple category system. Most

[1] Krippendorff consistently differentiates between agreement and reliability in the context of content analysis (in his case rather classically oriented): "To be clear, agreement is what we measure; reliability is what we wish to infer from it. In content analysis, reproducibility is arguably the most important interpretation of reliability" (2004, p. 414).

of the time, however, the data will be more extensive than that, in which case you will require a sample. It is sometimes suggested in the literature on the subject that a certain percentage of the data should be coded by a second person. Such a percentage rate of perhaps 10% may provide an initial indication, but due to potentially very different amounts of data and the diversity of conditions between projects, further criteria should definitely be included in the decision:

- The expected number of coded segments—for example, it makes no sense to limit yourself to very few documents if you only expect a few of the available categories to be used in the selected documents.
- The diversity of cases—the sample should include a broad spectrum of available data. A well-considered selection of documents according to the principle of maximum contrast (e.g., short vs. long texts or interviews with storytelling vs. short answering interviewees) or a random selection is recommended.
- The stage of development of the category system—especially when applying a newly developed coding frame for the first time, the intercoder agreement analysis should be started relatively early in order to be able to detect deficiencies in the coding frame.
- The available resources—intercoder analyses take time. Projects are often conducted under time pressure, and there are not always people available who are willing and able to do a second round of coding. However, when performing a qualitative content analysis, you should never go completely without an intercoder check. Sometimes the amount of effort involved is overestimated, but the motto "a little is better than nothing at all" clearly applies here.

In general, it is best to check the coding consistency with a manageable data set at an early stage to avoid finding out late in the project that the coding instructions were incomplete or misleading. To be able to report improvements regarding the coding process, you need to perform multiple checks, i.e., start with two very different documents, discuss any inconsistencies that have occurred, and then continue with two further documents.

Before starting the analysis in MAXQDA, it makes sense not only to think about the selection of documents but also the codes you want to include. When analyzing intercoder agreement, it usually makes no sense to check all the codes at the same time. First of all, codes such as "Interesting text passage," "Suitable quotation," and potentially also the code "Other" are often excluded from the analysis. Then, the analyst who has carried out the coding in two steps—by first applying broad-brush themes before subsequently differentiating them—should proceed with the agreement analysis in two equivalent stages: firstly for the broad-brush segments, followed by further analysis of the subcode assignments. In addition, the types of the assigned codes must be taken into account. Simple factual codes (e.g., whether a person claims to be a supporter of a party or not) should only be mixed with sophisticated codes in the context of a complex argumentation analysis if you are not interested in the calculation of an overall value of agreement.

The Procedure for Analyzing Intercoder Agreement in MAXQDA

MAXQDA allows you to determine the agreement between two coders for selected documents. To perform the analysis in MAXQDA, the documents need to exist twice in the project—once coded by person 1 in one document group or set, and once coded by person 2 in another document group or set. In addition, a coding frame and corresponding instructions for the coders must have been defined in advance. What is the best way to organize the agreement analysis process? The following steps illustrate an appropriate procedure:

- Step 1—Create a project with all the relevant documents and the complete code system. Specify the coding instructions for individual categories in the code memos. If the units that both people should code have been defined in advance, they can be tagged with a code called "Segments to be coded." If the data material has a uniform structure, it may be sufficient to instruct both coders to always code the entire paragraph or the entire answer to a question.
- Step 2—Activate write protections for the code system and document texts in MAXQDA's User Management system in the case of each coder to protect the project from unwanted changes (see Chap. 18). The selection of documents that are to be coded by a second person can easily be made visible, for example, by assigning a certain color to the document in the "Document System" window.
- Step 3—Provide a copy of the master project file to both coders.
- Step 4—Both people then code the selected documents and add their abbreviations or names to the name of a document group (or set) to be able to identify their code assignments later.
- Step 5—Merge the two project files into one, which will then contain two copies of all the documents to be compared, using the ***Home*** > ***Merge Projects*** function described in Chap. 18. Select the option ***Don't import already existing documents*** here, so that only documents coded by the second person are added. Once the import has been completed, the document names will indicate who has coded which document.

Once a project contains the selected documents that have been coded by both people, the analysis of the intercoder agreement can be done *for two document groups or sets*. To start the procedure, first select the codes you want to include in the analysis by activating them and then open the function via ***Analysis*** > ***Intercoder Agreement***. The dialog box that appears (Fig. 19.1) allows you to specify the two document groups or sets you want to analyze and to distinguish between three types of agreement.

MAXQDA can check the selected documents for consistent coding on three different levels, where the first two consistency types refer to the "document level" and the third to the "segment level":

- ***Code existence in the document***—A match is counted if both coders have assigned the same code to the document. It does not matter in this case whether one person assigned the code three times and the other only once. The location of the code in the document is also irrelevant, as long as the code exists somewhere

in the document. Disagreement regarding a given code therefore only occurs if one person has assigned the code *once or several times* and the other person has *not assigned it at all* in the document. This level of agreement check is interesting, for example, for categories that refer to the entire document. If a code "Previous rehab experience" is to be assigned in a study with rehabilitation patients, it may not matter where and how often the code was assigned in the document—the main point is that both coders *have* assigned it.

- ***Code frequency in the document***—A match is counted if both coders have assigned the same code in the document the same number of times. If one person has assigned the code "self-confidence" three times in the document and the other only twice, there is no agreement for this code. Again, the locations of the coded segments in the document do not play a role here.
- ***Min. code overlap between segments [%]***—A match is counted if both coders have assigned the same code to a given data segment. The segments do not have to be 100% identical in their position; you can set a tolerance range.

To correctly analyze intercoder agreement in MAXQDA, the compared documents need to be identical. If this is not the case, MAXQDA displays a warning message including a reference to the first location where the documents differ. In this case, we recommend abandoning the procedure to first examine the differences between each document to avoid producing incorrect results.

The agreement analysis can be carried out for all document types, i.e., for texts, PDFs, tables and images, as well as audio and video files.

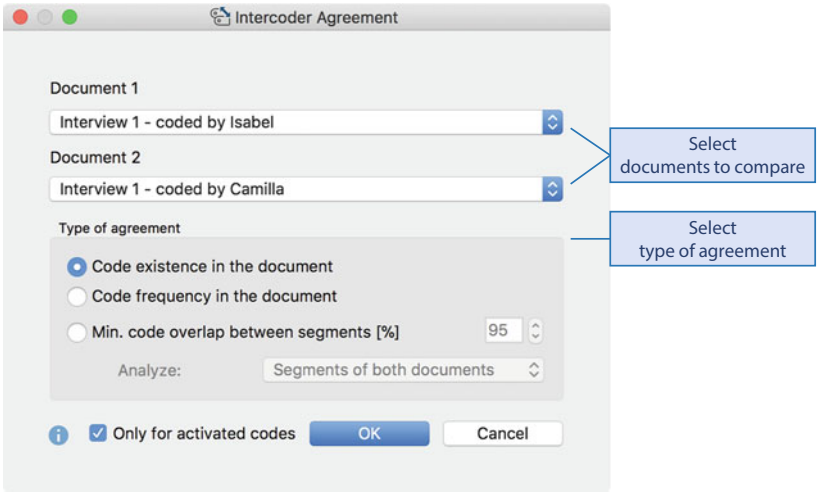


Fig. 19.1 Options dialog box for intercoder agreement analyses

Document-Level Intercoder Agreement

You can see an example of the result of an intercoder analysis in Fig. 19.2. Camilla and Isabel independently coded interview 1, and then an intercoder check was performed using the second option above, that is, where a match is counted when the code has been assigned the same number of times. To compare their code assignments, MAXQDA uses the Code Matrix Browser, in which the two documents are displayed side by side in two columns. In Fig. 19.2, the numbers display was switched on to provide an accurate comparison: for example, Isabel and Camilla have both assigned the code "Interests" twice, which means there is an intercoder agreement for this code. In the title bar, MAXQDA displays the percent agreement between the two coders in relation to all codes displayed in the Code Matrix Browser. The percentage value is calculated by the number of codes for which there is intercoder agreement divided by the number of all analyzed codes. This is the same as the number of rows with concordances divided by the total number of rows when all codes are expanded. The 60% agreement shown in the title of the example is calculated as follows: for three codes ("Interests," "Money and Financial Issues," "Religion and Spirituality") the two coders agree, and for the other two codes, they do not. The number of matching codes divided by the total number of codes is $3/(3+2) = 60\%$. This value indicates that the two coders agree on 60% of the codes, and they differ accordingly on 40% of the codes. Incidentally, the analysis variation "code existence in the document" would result in a percentage of agreement of 100% between these two coders, since both completely agree that the same four codes apply to this document and one does not.

▶ **Please Note** By default, only the activated codes are displayed in a linear list without their parent codes. When selecting the option ***Display Codes with Hierarchy*** in the toolbar, non-activated parent codes that are necessary for the correct display of the code tree may also be listed. The calculation of the percent agreement ignores these codes.

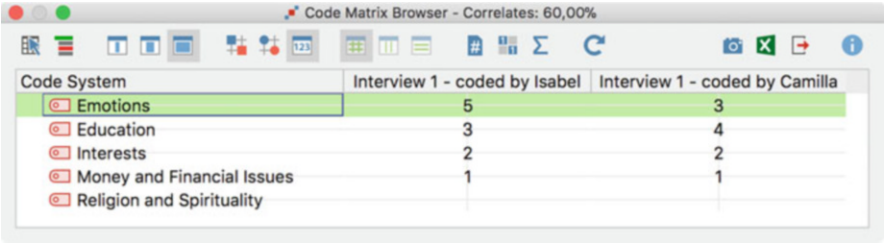


Fig. 19.2 Results table for an intercoder agreement analysis at the document level

Usually, the interesting aspects of an intercoder analysis are the disagreements, which can reveal problems concerning individual codes, the coding instructions, or the approaches of individual coders. MAXQDA supports the analysis of disagreements through its interactive display of the results: double-clicking on a cell lists all the corresponding coded segments in the "Retrieved Segments" window. Hence, in the example, you can see which segment Camilla has coded with "Emotions," but Isabel has not.

Segment-Level Intercoder Agreement

Although a document-level analysis provides initial indications of systematic differences in the use of categories, in most cases a segment-specific agreement check will be necessary. You can do this by selecting the third type of agreement in the options dialog box (***Min. code overlap between segments [%]***, Fig. 19.1). In practice, coders often not only assign categories to predefined segments but in fact define the segments that need to be coded as they go. In other words, if the latter approach is taken, it can often happen that one person codes an extra character or word compared to the other, and it is just as if not more likely that two coders will differ by 1 or 2 seconds when coding the same scene in a video. In order to ignore these minor, inconsequential differences in segment boundaries during the intercoder analysis, you can set a minimum overlap of two coded segments to be compared in the options dialog box as a percentage value. A code overlap of 100% means that the segment boundaries must match precisely to be counted as identical segments in MAXQDA.

► **Please Note** The percentage set in the options dialog box should not be confused with the percent agreement that is displayed as the result of the analysis. The options dialog box only lets you set the minimum overlap for which two coded segments are considered to be identical.

You must specify a minimum overlap every time you run this analysis. A value of 100% should be set if the coders worked on predefined segments. This would be the case if, for example, the coding instructions were to code each paragraph of a text, or if all the segments to be coded were previously tagged with a specific code and were then assigned to thematic codes. In most cases, however, you would be advised to start with a minimum overlap of about 95% as a test and reduce this value step-by-step if it results in an inordinate number of *insignificant* disagreements.

After starting an intercoder agreement analysis at the segment level, MAXQDA first processes the code assignments of the first document and then the ones of the second document. Each coded segment is checked to see if the other person has

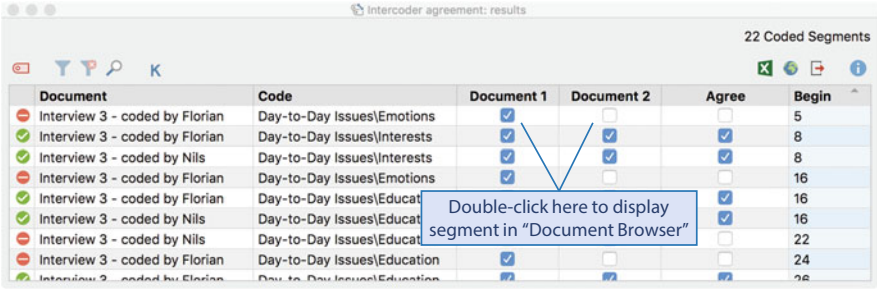


Fig. 19.3 Results table for an intercoder agreement analysis at the segment level (segment table)

assigned the same code to this segment. Assuming one person has coded 10 segments in the document and the other person has coded 12 segments, then 10 + 12 = 22 segments are checked for agreement. Figure 19.3 shows part of the results table MAXQDA produces for this analysis. Each row contains the result of the check of a single segment, where green check marks represent agreements and stop signs represent disagreements. The following information is provided in the first row: in paragraph 5, Florian (Document 1) assigned the code "Emotions," but Nils (Document 2) did not. Accordingly, the check mark is missing in the "Agree" column, and a stop sign is displayed in the first column.

▶ **Tip** By clicking on the column heading "Agree", the table can be sorted for that column, so that all disagreements are displayed at the top of the table.

In qualitative research in general, especially at the beginning of the analysis process, all disagreements are investigated and their causes identified. Did the second person miss something? Should two codes be more clearly separated or even merged? Did the two people assign the same code but diverge on where to place the segment boundaries? The interactivity of the result table supports you in identifying these and similar problems. Double-clicking in either the "Document 1" or "Document 2" columns opens the respective documents at the corresponding location in the "Document Browser" and lets you examine the differences between the coders.

The question of how to deal with disagreements swiftly follows. As we emphasized above, qualitatively oriented research projects aim to use these differences as an occasion for discussions concerning code assignments, the coding frame, and the document segments. To reach an agreement on problematic coded segments, it is sometimes helpful to include the whole case as contextual information or to investigate other coded segments under the same category. As a rule, it is worth logging this problem-solving and consolidating process as well as the arguments and points of view put forward, because these discussions can often result in valuable information. Not infrequently you can make very interesting discoveries relevant to

your research project and analysis process as a whole while solving the problem of defining and demarcating categories—even if these are only hypotheses that need to be tested on further data.

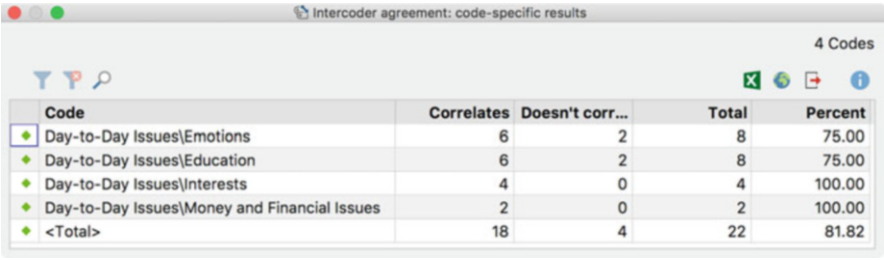
To optimize the coded segments in the MAXQDA project file, one of the two documents should be defined as the "master document." In this document, the code assignments can be improved where necessary, so that it contains the optimized version. Once you have finished the intercoder analysis, you can delete the second version of each document to continue working in the optimized, merged project.

Code-Specific Results Table

MAXQDA not only displays the segment results table but also a so-called code-specific results table. The latter shows not only the percentage of agreement across all codes but also for each code individually (Fig. 19.4). The percentage of agreement results from the proportion of agreements in the evaluated code. In Fig. 19.4, in the row "<Total>", with 18 agreements and a total of 22 coded segments, the overall percentage is 81.82%.

The percent agreement provides valuable information for the identification of problematic codes, but you should evaluate its size with a good degree of caution. Particular attention should be paid to the absolute number of segments evaluated per code in the "Total" column and the number of disagreements. For example, if each coder created five segments, a single disagreement would already reduce the percentage of agreement by 20 percentage points. And with two segments (as with the code "Money and Financial Issues" in Fig. 19.4) the table would only include the values 0% and 100%. Moreover, in addition to the absolute number, the total number of segments per coder should also be considered. If it is very unequal, there will usually be a systematic difference between the coders, for instance, because one of them coded the data in smaller, broken-down parts than the other, which will lead to a small percentage of agreement.

If you include the percent agreement for a qualitative study in a report or publication, you should always also state which segments caused the remaining differences and what relevance these have to the study. It is not particularly helpful to present only the "overall percentage" for all your codes. It might, instead, be better to provide at least the minimum and the maximum percentages regarding all tested codes or, better still, to include all the information in the code-specific results table in



Code	Correlates	Doesn't corr...	Total	Percent
Day-to-Day Issues\Emotions	6	2	8	75.00
Day-to-Day Issues\Education	6	2	8	75.00
Day-to-Day Issues\Interests	4	0	4	100.00
Day-to-Day Issues\Money and Financial Issues	2	0	2	100.00
<Total>	18	4	22	81.82

Fig. 19.4 Results table for an intercoder agreement analysis at the segment level (code table)

the publication. This is advisable because, for the reasons mentioned above, the absolute number of evaluated codes and disagreements should also be stated. During the analysis it is therefore necessary to document problematic codes and discrepancies to be able to include them in reports and publications later.

The question remains as to what is to be regarded as a low and what as a high percentage of agreement. Unfortunately, this question cannot be answered with established thresholds, because the percent agreement between coders not only depends on the number of absolute coded segments as described, but also on other factors. These include, in particular, the number and variance of different (sub) categories as well as the degree of difficulty of the coding process itself (it is more difficult to code an interviewee's argument, e.g., than it is to assign factual or thematic codes). A simple inversion of the conclusion will usually help you evaluate whether the percentage is too low: if a code has 80% agreement, for example, this means that 20% of the code assignments differ. Usually this level of disagreement would not be considered ideal, and you would need to set stricter thresholds. However, how you determine suitable values should ultimately always be related to the content of your data.

Coding Units vs. Coded Segments

When conducting an intercoder analysis using MAXQDA, it is important to distinguish between coding units and coded segments. In many approaches for the analysis of intercoder agreement, it is assumed that a single code is assigned to each coding unit. This is the case, for example, if physicians rate an x-ray image for the presence of a disease or if entire newspaper articles are rated low, medium, or high on a scale of "latent racism." Then exactly one code has been assigned by each coder to each coding unit.

This is not the case in many analyses performed with MAXQDA. It is quite possible that the two coders have assigned different numbers of codes to the same passages in the document. Furthermore, it is common for the coding units not to have been predefined, and the coders quite often decide about which passages to code on their own, which may result in overlapping coded segments or coded segments contained within other segments. To perform an intercoder analysis in such cases, MAXQDA follows the same process described above: first the codes of coder 1, then those of coder 2, are checked for agreement with the other person, respectively. Figure 19.5 shows some examples that illustrate how many matches result from this procedure in different cases. The first row, in which both coders assigned only "Code A" to the same segment, indicates that two agreements have been counted here. The case where the coders assigned one code to the segment each, but these codes are not the same, results in two disagreements, as shown in the second row.

When you open the ***Intercoder Agreement*** function, a drop-down menu labeled ***Analyze:*** is visible in the options dialog box that opens (Fig. 19.1), where you can instruct MAXQDA to analyze only the coded segments in a single document, i.e., the document coded by only one of the coders, to assess their level of agreement with the other coder. This setting can be used where both coders have coded predefined

			Coded segments	Agreements	Disagreements	Percent
A	Text segment	A	2	2	0	100%
A	Text segment	B	2	0	2	0%
B, A	Text segment	A, B	4	4	0	100%
B, A	Text segment	A	3	2	1	67%
B, A	Text segment	C	3	0	3	0%
B, A	Text segment	C, D	4	0	4	0%

Fig. 19.5 Number of agreements, in different constellations, between two coders

segments with a single code. In this case, the number of coded segments evaluated corresponds to the number of coding units in the results table.

Calculating Chance-Corrected Agreement Coefficients like Kappa

"My supervisor told me to calculate kappa. How do I do this in MAXQDA?" "I need kappa for my analysis, don't I?" These questions are not only on the minds of the many doctoral candidates we advise in workshops, they are also regularly asked in the discussion forum on the MAXQDA website. This is often based on the desire to legitimize one's own qualitative approach to a scientific community that primarily adheres to the tradition of quantitative research. Here kappa is a known coefficient, which is why supervisors often demand it. A coefficient like kappa quantifies the quality of qualitative analysis as a figure and translates the work into a comprehensible (and familiar) form. Calculation of such measures helps to keep you connected to some scientific communities and increases the chance of publication in certain journals.

Even if an excessive emphasis on a chance-corrected coefficient at the expense of other important quality criteria for qualitative research should be called into question, the calculation of chance-corrected coefficients does have its place. When determining the percent agreement, the question arises as to how likely it is that agreement could have arisen by chance. To answer this question, chance-corrected coefficients were developed that subtract possible random matches from the raw agreement. Their central idea involves determining to what extent human coders can code a text or video, with an existing set of categories, better than a randomly working machine.

The following basic formula is often used to calculate chance-corrected coefficients: $(Po - Pc)/(1 - Pc)$, where Po is the observed percentage of agreement and Pc is the expected agreement by chance. Following this method of calculation, the resulting value indicates how much the agreement of the coders exceeds the agreement by chance. Coefficients such as Kappa (Cohen, 1960), Pi (Scott, 1955), and Alpha (Krippendorff, 1970) differ primarily in how the expected agreement by

Table 19.1 "Code by code" table; the agreements are on the diagonal line (a, e, i)

	Coder 2			
Coder 1	Cat. 1	Cat. 2	Cat. 3	SUM
Cat. 1	a	b	c	a + b + c
Cat. 2	d	e	f	d + e + f
Cat. 3	g	h	i	g + h + i
SUM	a + d + g	b + e + h	c + f + i	N

chance is computed. Usually a matrix "categories by categories," as shown schematically in Table 19.1, is used as the basis for calculating these coefficients. The cells indicate how often the respective categories have been assigned to a coding unit by the two coders. The primary diagonal contains the cells with the agreements between the two coders (the cells a, e, i). The agreement by chance is computed using the marginal sums. For the following explanations, it is important to emphasize that generating a table of this kinds usually requires that the coders have assigned only one category to each segment.

Here we will limit ourselves to the calculation of chance-corrected coefficients for agreements at the segment level, since they are seldom determined at the document level and cannot be performed automatically in MAXQDA. To request the calculation of a chance-corrected coefficient at the segment level in MAXQDA, click on the kappa symbol in the results table (Fig. 19.3) after performing the intercoder analysis. MAXQDA then generates a set of results as shown in Fig. 19.6.

What information is included in the results MAXQDA produces and how can it be interpreted? In practice in qualitative research projects, you will rarely be able to

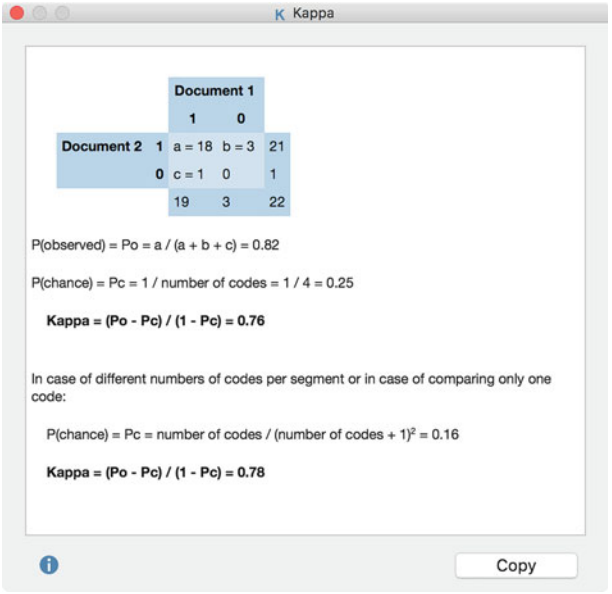


Fig. 19.6 Results window for calculating kappa according to Brennan and Prediger (1981)

create a matrix as shown in Table 19.1, since coders often assign more than one code to a segment, and, furthermore, it also happens that one person assigns one or more codes to a segment where the other person assigns no category at all. This is the reason why MAXQDA goes through the coded segments of one coder, and then the segments of the other, as described above, and counts them as matches if the other person has assigned the same code to the same segment. In order to determine a chance-corrected coefficient for this procedure, MAXQDA generates a 2 × 2 table as shown in Fig. 19.6. The upper left cell (a) indicates how often the two coders have assigned the same code to a segment. The upper right cell (b) and the lower left cell (c) indicate for how many segments the two coders differ in their assignment. The fourth cell (d) is always zero, because due to the method used, there are no predefined segments that were not coded by either coder.

Below this table, both the raw agreement Po and the expected agreement by chance Pc are calculated. Po corresponds to the value output in the "<Total>" row of the code-specific table. Since the marginal distributions of the 2 × 2 table are always unequally distributed due to cell d = 0, agreement by chance cannot be calculated as for Cohen's Kappa, Scott's Pi, or Krippendorff's Alpha. Unequal marginal distributions can lead to abstruse and paradoxical values in Cohen's Kappa, which is a frequently articulated point of criticism (e.g., in Feinstein & Cicchetti, 1990; or Gwet, 2008). When calculating Pc, the calculation in MAXQDA therefore follows a concept proposed by Brennan and Prediger (1981).[2] Instead of determining the expected agreement by chance using the marginal distribution, the number of categories is used here. You can see quite easily that the probability of agreement decreases as the number of categories increases. Computationally, Pc is $1/n$, where n corresponds to the number of categories used. This is graphically illustrated in the left table in Fig. 19.7: the number of gray cells with coincidences on the main diagonal corresponds to the category number n, and the total area of the table corresponds to n^2 cells, resulting in a random coincidence of $n/n^2 = 1/n$.

In the frequently occurring case that the coders differ in the number of categories assigned per segment, the expected random agreement can be slightly corrected downward. As the middle table in Fig. 19.7 shows, another category "X" is added in this case, which represents "not coded." Since the number of cells with coincidences (by chance) still corresponds to the category number n, Pc is now calculated with $n/(n+1)^2$. MAXQDA also displays the value calculated in this way. In the example, it is 0.78, slightly greater than the "normal" Kappa value of 0.76. As the table at the far right in Fig. 19.7 illustrates, this way of calculating must also be used if only a single code is evaluated for an agreement analysis.

[2]Krippendorff (2004, p. 417) points out that this proposal was already formulated in the 1950s and was later "reinvented" with slight variations by several authors, including Brennan and Prediger (1981).

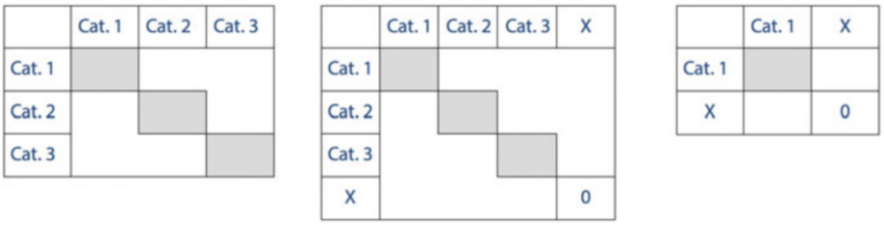


Fig. 19.7 Determining the agreement by chance (gray); "X" stands for "not coded"

▶ **Please Note** MAXQDA does not provide Cohen's Kappa, but kappa according to Brennan and Prediger (1981), who named their coefficient with a Greek kappa with subscript n: κ_n. If you use the results of a MAXQDA calculation in a publication, you should make a reference to Brennan and Prediger to avoid confusion.

Clearly, as the number of categories increases, the agreement by chance calculated in this way will decrease. Let us assume that there is a 90% match. Then the random correction for two categories leads to a Brennan and Prediger's Kappa of 0.80, and for ten categories, Kappa would be as high as 0.89.

But how should the level of Kappa be evaluated? Brennan and Prediger's Kappa can take values between −1.00 and +1.00; a value of 0 corresponds to a parity with chance, and a value of +1.00 corresponds to perfect agreement of the coders—this is as far from agreement by chance as you can get. 1.00 is reached if the percent agreement between the two coders is 100%. The interpretation of the value can be based on the established benchmark notes for Cohen's Kappa: according to Landis and Koch (1977) one can label a result as good ("substantial") from 0.61 and as very good from 0.81 ("almost perfect"). However, any such threshold could be misleading. First, in many cases Cohen's Kappa can never reach the value of 1.00 due to its calculation method, which is why the threshold values for Brennan and Prediger's Kappa might be raised but should never be lowered. Second, as previously explained for the percent agreement, the definition of this threshold and the interpretation of its value should also be justified in relation to content, for example, by explaining which remaining inconsistencies were accepted by the researchers.

We want to conclude this section with two important notes: *firstly*, we think that the calculation and publication of a chance-corrected coefficient should by no means distract qualitative researchers from the process of improving their category system. *Secondly*, it is necessary for the calculation of chance-corrected coefficients that the segments to be coded, i.e., the coding units, are defined a priori. If the coders are free to set the segment boundaries, there is no sense in calculating chance-corrected coefficients of agreement. The reason for this is obvious: even for a one-page text of 2000 characters, the probability Pc that two coders will randomly select exactly the same characters and assign the same code to them tends to zero. A random correction is therefore not necessary.

References

Brennan, R. L., & Prediger, D. J. (1981). Coefficient kappa: Some uses, misuses, and alternatives. *Educational and Psychological Measurement, 41*(3), 687–699. https://doi.org/10.1177/001316448104100307.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37–46. https://doi.org/10.1177/001316446002000104.

Feinstein, A. R., & Cicchetti, D. V. (1990). High agreement but low Kappa: I. The problems of two paradoxes. *Journal of Clinical Epidemiology, 43*(6), 543–549. https://doi.org/10.1016/0895-4356(90)90158-L.

Gwet, K. L. (2008). Computing inter-rater reliability and its variance in the presence of high agreement. *The British Journal of Mathematical and Statistical Psychology, 61*(Pt 1), 29–48. https://doi.org/10.1348/000711006X126600.

Krippendorff, K. (1970). Bivariate agreement coefficients for reliability of data. *Sociological Methodology, 2*, 139–150. https://doi.org/10.2307/270787.

Krippendorff, K. (2004). 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.

Kuckartz, U. (2014). *Qualitative text analysis: A guide to methods, practice & using software*. Thousand Oaks, CA: SAGE.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159–174. https://doi.org/10.2307/2529310.

Mayring, P. (2014). *Qualitative content analysis: Theoretical foundation, basic procedures and software solution*. Klagenfurt. Retrieved from http://nbn-resolving.de/urn:nbn:de:0168-ssoar-395173

Schreier, M. (2012). *Qualitative content analysis in practice*. Thousand Oaks, CA: SAGE.

Scott, W. A. (1955). Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quarterly, 19*(3), 321–325. https://doi.org/10.1086/266577.