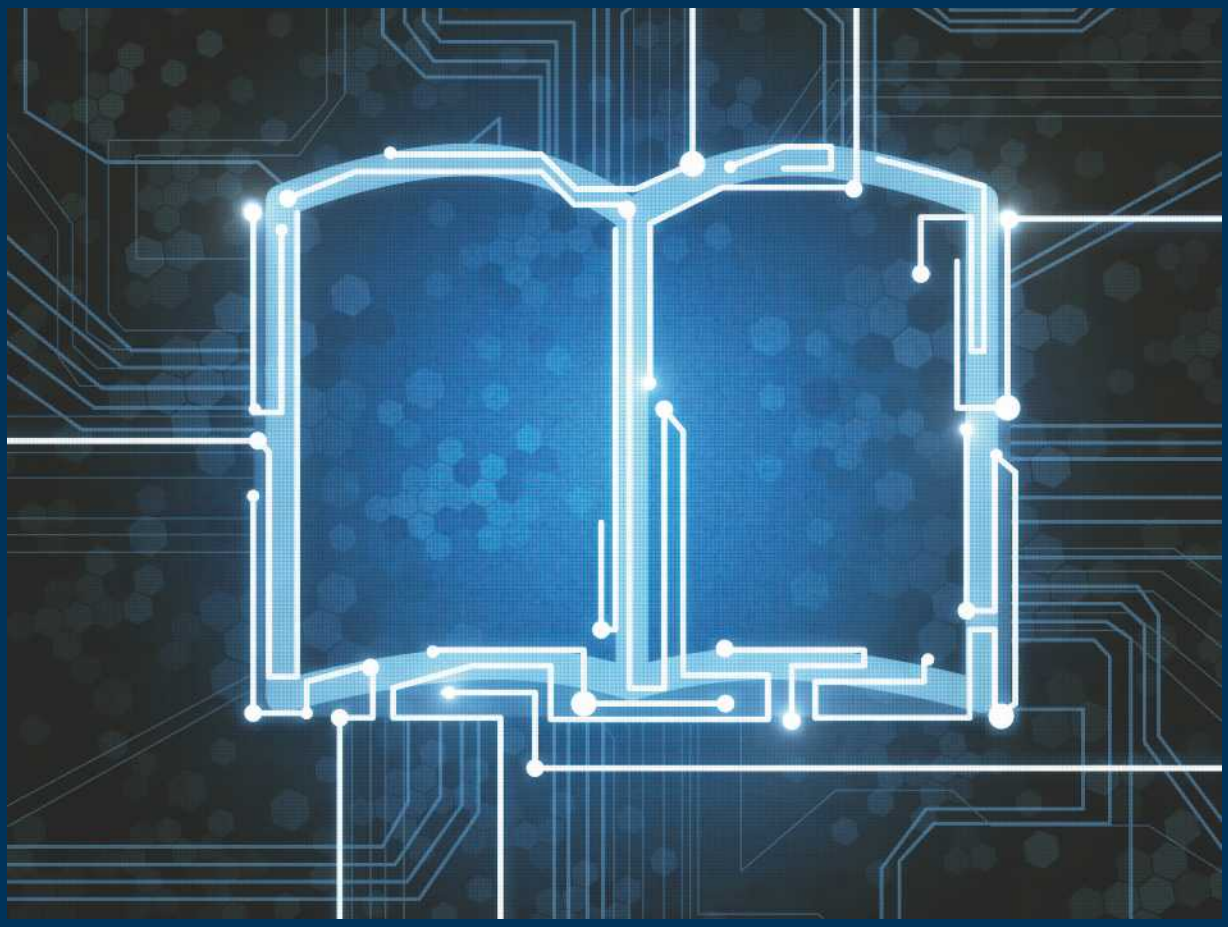
Florian Berding, Julia Pargmann

Iota Reliability Concept of the Second Generation

Measures for Content Analysis Done by Humans or Artificial Intelligences



λογος

Berufs- und Wirtschaftspädagogik 4

herausgegeben von Florian Berding und Tobias Schlömer

Berufs- und Wirtschaftspädagogik

Band 4

Berufs- und Wirtschaftspädagogik

Band 4

Herausgegeben von

Prof. Dr. Florian Berding

Universität Hamburg

Prof. Dr. Tobias Schlömer

Helmut-Schmidt-Universität-Hamburg/Universität der Bundeswehr Hamburg

Florian Berding, Julia Pargmann

Iota Reliability Concept of the Second Generation

Measures for Content Analysis Done by Humans or Artificial Intelligences

Logos Verlag Berlin

Bibliographic information published by Die Deutsche Bibliothek

Die Deutsche Bibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data is available in the Internet at http://dnb.ddb.de.

Cover image: Schulbuch by Greyfebruary via istock

ISBN 978-3-8325-5581-8
ISSN (Print) 2629-3137
DOI https://doi.org/10.30819/5581

Logos Verlag Berlin GmbH
Georg-Knorr-Str. 4, Geb. 10,
12681 Berlin

Tel.: +49 (0)30 / 42 85 10 90
Fax: +49 (0)30 / 42 85 10 92
http://www.logos-verlag.de

Content

1 Introduction

Analyzing textual data via content analysis is a popular research method in the social sciences. Krippendorff (2019, p. 24) defines it as a "research technique for making reliable and valid inference from texts (or other meaningful matter) to the context of their use. [italic in the original]". Yet the range of its applications is not limited to research. Data generated by content analysis can be a valuable source of information in other fields like education in the social sciences (Berding et al., 2022). In educational settings, textual data is omnipresent, manifested in artefacts such as explanations in school books, tasks on worksheets, in written essays or exams, in lesson plans or curricula, or in written communication between teachers and learners in digital learning environments.

Textual data offers deep insights into learning and learning outcomes (e.g., Hußmann et al., 2007; Leuders, 2010; Sjuts, 2007, 2010). For example, if a teacher would like to know if their students have developed the "right" understanding of "prices" in an economic context, an easy way is to ask learners to develop their own explanations. These written explanations do not only give a teacher an idea about the learning outcome but also provide insight into the students' understanding of the topic. This information can then be used for fine-tuning further instruction.

The approach to gather, analyze, and use data to improve learning and learning outcomes of individuals is discussed as learning analytics and is pursued through the creation of educational settings and learning processes that match the learners' individual needs and conditions (e.g., Larusson & White, 2014, pp. 1–2). Utilizing textual data for learning analytics requires a technology that is able to understand text-based sources like humans do. This leads to the application of artificial intelligence (AI), a technology that attempts to simulate human actions (Kleesiek et al., 2020, p. 24). Within educational settings, AI has to understand the textual information, summarize the information in categories of scientific models and theories, and derive the impact of the categories on further learning to provide recommendations for learners and teachers (Berding et al., 2022). For example, if the analysis of a student's text leads to the conclusion that the student has not acquired the "right" understanding of prices, it does not make sense to teach new topics that build on a valid understanding of prices. In this example, the student is expected to have a high risk of failing in a newly introduced topic. Thus, before teaching the new topic, more time and effort has

to be spent to help the student construct a valid understand of prices. In the contrasting case where a student writes an essay that implies a valid understanding of prices, it is more efficient to introduce a new topic as soon as possible.

When employing content analysis in practical educational settings, teachers and other users need to ensure the same quality criteria as within sciences, namely objectivity, reliability and validity (Hesse & Latzko, 2011, p. 70; Ingenkamp & Lissmann, 2008, p. 51). Thus, the issue with quality in the scope of content analysis is the same for practice and research, and for both human and artificial intelligence. For this research method in particular, special criteria focusing on reliability have emerged in the last decades. Reliability is a central characteristic of any assessment instrument, and describes the extent to which the instrument produces error-free data (Schreier, 2012, pp. 166–167). Krippendorff (2019, p. 281, p. 283) suggests replicability as a fundamental reliability concept, which is also referred to as inter-rater reliability. This describes the degree to which "a process can be reproduced by different analysts, working under varying conditions, at different locations, or using different but functionally equivalent measuring instruments" (Krippendorff, 2019, p. 281).

Past decades have seen a large number of reliability measures being suggested. A study by Hove et al. (2018) shows that the 20 reliability measures they investigated differ in their numeric values for the same data, making an evaluation of the quality of content analysis difficult. Krippendorff's Alpha is currently the most frequently recommended reliability measure (Hayes & Krippendorff, 2007), as it can be applied to variables of any kind (nominal, ordinal, and metric), to any number of raters and to data with missing cases and unequal sample sizes; all while comprising chance correction (Krippendorff, 2019, p. 291). Thus, it is not surprising that Krippendoff's Alpha has become one of the most popular measures of reliability for content analysis in research (Lovejoy et al., 2016, p. 1150). Currently, this measure is also evaluated for characterizing the quality of input data for machine learning (Song et al., 2020).

Recent years however have seen the advantages of Krippendorff's Alpha being questioned and controversially discussed (Feng & Zhao, 2016; Krippendorff, 2016; Zhao et al., 2018). Zhao et al. (2013) analyzed different reliability measures, concluding that Krippendorff's Alpha contains problematic assumptions and produces the highest number of paradoxes and abnormalities amongst all

included measures. For example, they argue that Alpha penalizes improved coding. That is, if raters correct errors, the values for Alpha can decrease (Zhao et al., 2013, p. 457). Furthermore, cases exist where rater agreement is nearly 100%, while the Alpha values are about 0, indicating the absence of reliability. Based on their findings, Zhao et al. (2013, p. 475) recommend developing and trying new reliability measures. Feng and Zhao (2016, p. 146) suggest not to base new approaches on classical test theory, but on item response theory.

In classical test theory, reliability is characterized with measures such as Cronbach's Alpha. These measures produce a single numeric value for a complete scale. Item response theory however, is more detailed. With the help of the test information curve, the reliability of a scale can be investigated for different characteristics of that scale (e.g., Ayala, 2009, pp. 27–33; Baker & Kim, 2017, pp. 96–98). Furthermore, some models of item response theory such as Rasch models offer the opportunity to investigate if a scale produces bias for different groups of individuals. That is, they allow to investigate if an instrument functions similarly for different groups of people (subgroup invariance) (e.g., Baker & Kim, 2017, pp. 38–42).

Berding et al. (2022) transferred the idea of item response theory to content analysis by suggesting the *Iota Reliability Concept*. This concept provides several measures for characterizing the reliability of *every single category* of a coding scheme. In addition, the concept is able to produce insights into how errors in one category influence the data representing other categories, and how data may be biased for different groups of individuals. In the first study, the Iota Concept showed promising statistical properties such as high values for recovering the true reliability of a category, independence from the number of raters, the number of categories, the underlying distribution of categories, and the sample size. The Iota Concept showed a comparable ability as Krippendorff's Alpha to predict the deviation between the true and estimated relationship for two ordinal variables.

However, the study by Berding et al. (2022) is limited. The true reliability is modeled with the assumption that the reliability is the same over all categories and that the second ordinal variable is measured with perfect reliability. Both assumptions pose restrictions for applications in science and practice. In particular, rules of thumb for judging about the quality of a content analysis may be inadequate if these rules rely on the assumption that the other variables are

measured with perfect reliability. More realistic rules have to be derived on the assumption that all variables vary in their reliability and that a relationship between the variables is not perfect.

Thus, the work in this researcher's guide aims to address these limitations and to further improve the methodology of content analysis. To reach this aim, the guide develops a new and improved version of the Iota Concept. The scientific and practical value added is composed of

- *providing insights into the reliability of every single category.* Previous measures often used in content analysis such as Krippendorff's Alpha, Percentage Agreement, Scott's Pi, and Cohen's Kappa (Lovejoy et al., 2016) describe the reliability of a scale with one single numeric value assuming that the reliability is constant for the complete scale (Feng and Zhao, 2016). The Iota Concept adapts the basic ideas of modern test theory (Ayala, 2009; Baker and Kim, 2017; Bonifay, 2020; Paek and Cole, 2020) and overcomes this limitation (Berding et al., 2022). These insights can help in the construction of a coding scheme by directly showing where the coding scheme is performing well and where improvements are necessary.
- *providing insights into how a coding scheme may produce bias for different groups of individuals/materials.* These insights can be used to evaluate scientific research and/or to review coding schemes to avoid these contortions. This is particular important if the data from content analysis is used within AI-based learning analytics, since artificial intelligence can reproduce advantages or disadvantages for specific groups of learners, which are an implicit part of data (Luan et al., 2020, p. 5; Seufert et al., 2021, pp. 14–15).
- *providing rules of thumb for evaluating content analysis.* Deriving rules of thumb for evaluating the quality of content analysis under realistic conditions ensures a high certainty for generating reliable data while save costs at the same time.
- *providing possibilities for data replication*. Most literature on content analysis concentrates on the development and evaluation of new coding schemes (Früh, 2017; Krippendorff, 2019; Kuckartz, 2018; Mayring, 2015; Schreier, 2012). Less emphasis is put on the application of an existing coding scheme to new data by new raters. The new Iota Concept aims to provide a framework that allows the assessment of reliability in these situations. The

generated insights can be used for the specific training of new raters or to evaluate the quality of machine learning.

- *providing error-corrected data.* Modern statistical techniques such as latent modeling do not assume that a construct is measured with perfect reliability and provide methods for calculating error-corrected data (Geiser, 2013, p. 40; Wang & Wang, 2020, p. 1). The updated Iota Concept aims to provide such a technique for the method of content analysis.

In chapter 2, we present a short summary of the old Iota Concept as described in Berding et al. (2022) as an introduction into the concept. The third chapter presents a refined version of the Iota Concept and compares the new version to the old one.

Chapter 4 introduces an estimation algorithm based on the techniques of Maximum Likelihood Estimation to estimate the parameters of the new concept. In chapter 5, a simulation study is performed to investigate the probabilities of the new algorithm where 415,291 coding processes based on 10,514 sets of true parameters are simulated. Altogether 7,093,054 parameters are estimated. The analysis of these parameters includes the effects of the sample size, the number of raters, the number of categories, and the shape of the true distribution on the accuracy of the estimates.

After providing insight into the quality of the estimates, a second simulation study investigates the new concept's ability to predict coding quality (chapter 6). This is characterized by the deviation between the true and the estimated sample association/correlation, the risk of Type I and Type II Errors, and by the chance to correctly classify the effect size according to the work of Cohen (1988). In this study, the assessment of a coding scheme's reliability and the corresponding coding process are being simulated. The analysis is based on 6,044,572 coding processes. These coding processes are nested within 201,486 reliability estimations prior to coding. Both processes (reliability estimation and coding) are based on 25,399 different sets of true parameters. In this simulation, the reliability of both variables varies and the strength of association/correlation is modeled through the respective rules for effect sizes developed by Cohen (1988). These rules provide guidelines for deciding if an association/a correlation is of no, small, medium or strong practical importance. The simulation also includes a comparison with the old Iota Concept, Krippendorff's Alpha and Percentage Agreement. In consequence, realistic cut-off values will be derived from

modeling the relationship between estimated reliability values and the measures which characterize the quality of the generated data. Chapter 7 describes a third simulation study which aims to improve the reliability measures on the scale level, based on the insights from the second simulation study.

The illustrations end with a summary and a discussion of the results in chapter 8. In the final chapter, we discuss some practical examples for the application of the new concept using the *R* package *iotarelr*.

2 Summarizing the Iota Concept of the First Generation

The Iota Concept of the first generation (Berding et al., 2022) is based on the following six assumptions:

1. The core of content analysis is a scheme guiding raters to assign a coding unit to a category. Here, reliability is a property of a coding scheme, not of raters.
2. The categories form a nominal or ordinal scale with discrete values.
3. Every coding unit must be assigned completely to one category (thus, it is not possible to assign 50% of the unit to category 1 and the other 50% to category 2).
4. Every coding unit can be assigned to at least one category.
5. Raters judge the category of a coding unit by using a coding scheme or by guessing.
6. Reliability can vary for each category.

Figure 1 illustrates the model behind these assumptions. Raters use a coding scheme as a tool for assigning a category to a coding unit (see the arrows in the figure). The coding scheme defines the range of possible categories and provides definitions, rules, and examples as clearly as possible to guide raters in their work (see the coding scheme in the figure). In the best case, the coding scheme is so clear that every rater assigns the same category to a coding unit if the conditions for the specific category are satisfied (see arrow between the table on right and the coding units on the bottom left).

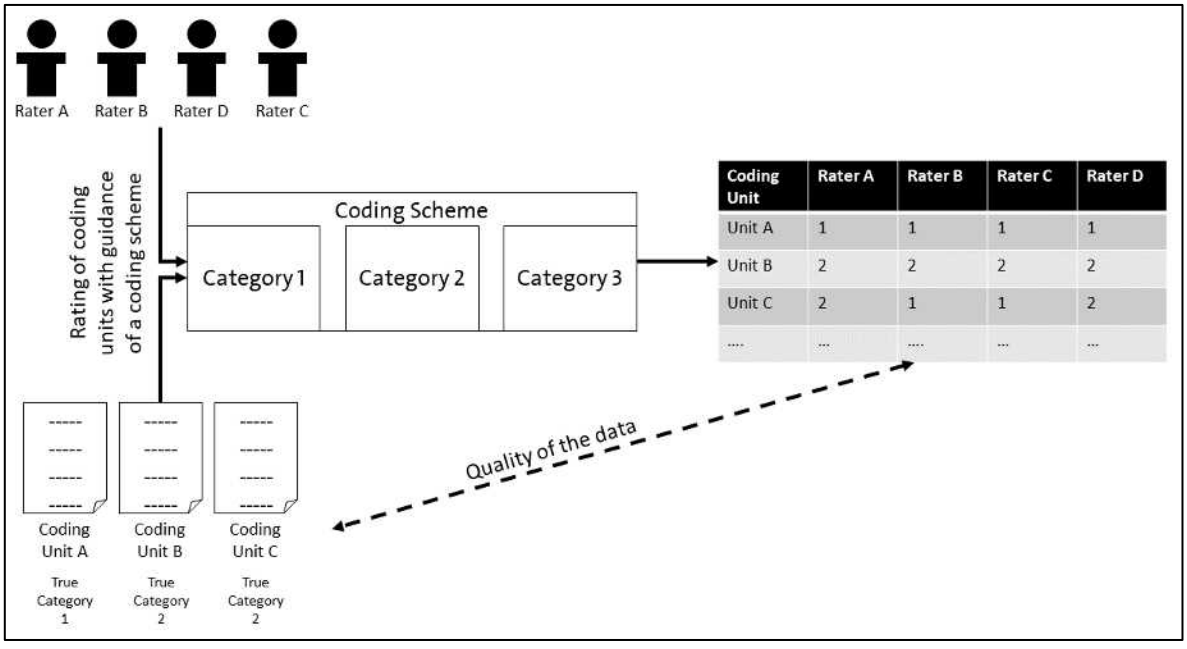


Figure 1. Coding Model of the Iota Concept

In practice, however, a coding scheme is not perfect. Thus, errors may occur. In Figure 1, rater B assigns coding unit A to category 1 and coding unit B to category 2, which is both correct. In the case of coding unit C, a mistake occurs. Rater B assigns this unit to category 1 instead of category 2. This mistake has two implications. First, the data generated by rater B underestimates the frequency of category 2. Second, the mistake leads to an overestimation of category 1. Thus, an error in recovering one true category always has an impact on the data generated for the other categories. To account for these different types of errors, the Iota Concept differentiates between the true category of a coding unit and the assigned category of a coding unit. If a coding scheme is reliable, both categories match each other. Consequently, the concept consists of the following three elements, and it also entails a special matrix for the description of the coding scheme's reliability, as well as two corresponding types of errors:

- **Alpha Elements:** The Alpha Elements account for the probability that a rater using the coding scheme is able to assign the true category to a coding unit. Thus, the A*lpha Reliability* of category i describes the conditional probability to assign category i to a coding unit if the true category of that coding unit is i. The A*lpha Error* of category i is the complementary probability to the Alpha Reliability. That is, the probability to assign category j with $j \neq i$ to the coding unit if the true category of that coding unit is i.
- **Beta Elements:** The Beta Elements are used to describe situations in more detail in which Alpha Errors occur. In these situations, a coding unit with the true category j is assigned to category i with $j \neq i$. Thus, the data of category i is biased since it comprises frequencies which truly belong to any of the other categories. The *Beta Error* of category i describes the probability to assign a coding unit of any other category as i to category i. The *Beta Reliability* of category i is the complementary probability and characterizes the probability that all other coding units are not assigned to category i.
- **Iota Elements:** The Iota Elements combine both Alpha and Beta Elements and correct the values for guessing. Mathematically, Iota of category i is the mean of the chance-corrected Alpha and Beta Reliability. The index ranges from 0 – indicating that the ratings of category i equals guessing – to 1 – indicating perfect reliability.
- **Assignment Error Matrix (AEM):** The Assignment Error Matrix provides the most detailed description of a coding scheme's reliability. The diagonal cells

(top left to bottom right) show the Alpha Error for the specific category. The remaining cells describe the probability that an Alpha Error guides the raters towards assigning a coding unit to another specific category. Thus, these cells describe the Beta Error in more detail. The interpretation of this matrix can best be explained using the example shown in Table 1. The Alpha Error for category 1 is about 47%. That is, in about 47% of cases, a coding unit that truly belongs to category 1 is assigned to another category. When this error occurs, about 71% of cases are assigned to category 2, and about 29% of cases are assigned to category 3. Here, category 2 is more strongly impacted by coding errors of category 1 than category 3.

Table 1. An Example of an Assignment Error Matrix of First-Generation Iota

		Assigned Category		
		1	2	3
True Category	1	.471	.709	.291
	2	.690	.959	.310
	3	.478	.522	.941

- **Minimum/Average Iota:** In many applications, not only the reliability of each category is important but also how the categories work together as a scale. For the description of the reliability on a scale level, the first generation of the concept suggests the average or the minimum of the Iota values as a reliability measure.

In the development study conducted by Berding et al. (2022), the new measure yielded promising results as it is unaffected by sample size, the number of categories and the number of raters.

3 Introducing Iota Concept of the Second Generation

3.1 Refining the Assignment Error Matrix, the Alpha and the Beta Elements

In contrast to the first generation of Iota, where the Assignment Error Matrix is a result of the Alpha and Beta elements, in the second generation, the Assignment Error Matrix represents the central object. This includes a small redefinition of its components, illustrated by Table 2.

Table 2. Example of an Estimated Assignment Error Matrix of the Second Generation

		Assigned Category		
		0	1	2
True Category	0	.508	.392	.100
	1	.000	.823	.177
	2	.237	.000	.763

The Assignment Error Matrix represents the true category in the rows and the assigned category in the columns. The values in the cells describe the conditional probability that a coding unit of category i is assigned to category j. For example, the probability to assign category 0 to a coding unit truly belonging to category 0 is about 51%%. The probability to assign category 1 to a coding unit which truly belongs to category 0 is about 39%% and the probability to assign such a coding unit to category 2 is about 10%%. Thus, the cells on the diagonal represent the probability to assign the right category to a coding unit.

Based on the Assignment Error Matrix, the Alpha Reliability can be defined. The Alpha Reliability of category i is the probability that a coding unit with the true category i is assigned to category i. Thus, the different Alpha Reliabilities equal the diagonal of the Assignment Error Matrix. The Alpha Error of category i is the complementary probability. That is, the conditional probability to assign a coding unit of category i to any of the other categories. The following equations apply:

$$\alpha_i^{rel} = aem(i,i) \quad [1]$$

$$\alpha_i^{err} = 1 - aem(i,i) \quad [2]$$

In Table 2, the Alpha Reliability of category 0 is about 51%%, meaning that in about every second case a coding unit of category 0 is assigned to category 0. For category 1, the Alpha Reliability is about 82%%. That is, a coding unit truly belonging to category 1 is assigned to category 1 in most cases. For category 3, a

similar result occurs. About 76%% of the coding units belonging to category 3 are assigned to the right category.

Related to the data generated by the coding scheme of Table 2, 49%% of the coding units belonging to category 0 are not correctly recognized. Thus, these units are missing in the data characterizing category 0. Furthermore, these 49%% are falsely assigned either to category 1 or 2 and thus skew the data of both of these categories. The Beta Reliability and the Beta Error characterize this phenomenon.

With the help of the Assignment Error Matrix, the Beta Error can be defined. The Beta Error describes the probability that coding units truly belonging to any other category than i are assigned to category i. Equation 3 shows the corresponding expression.

$$\beta_i^{err} = \frac{\sum_{k \neq i} p_k * aem(k,i)}{\sum_{k \neq i} p_k * \alpha_k^{err}} \qquad [3]$$

First, we concentrate on the denominator. The necessary condition for the Beta Error is that an Alpha Error occurs for a coding unit belonging to any other category than i. That is, a rater has to fail on discovering the true category of a coding unit in order to have the chance to assign the coding unit to any other category. These other categories are represented with k in the equation.

Furthermore, the probability of assigning a coding unit to any other category than the correct one increases if the amount of the categories increases in the population. The more coding units of a specific category exist in a population, the greater the chance to assign these coding units to another category. Thus, the Alpha Errors have to be weighted with the frequency with which a category appears in the population, represented with $p_k$in Equation 3.

While the denominator characterizes the general probability for making an error, the numerator concentrates on the category under investigation. The numerator is the probability that coding units of any other category are assigned to category i.

Based on this equation, the Beta Reliability is the complementary probability of the Beta Error, defined as the probability that coding units of any other categories than i are *not* assigned to category i.

$$\beta_i^{rel} = 1 - \beta_i^{err} \quad [4]$$

The meaning of the Beta Elements can be explained with an example from Table 2. All values which Table 2 implies can be found in Table 3. It is assumed that the quantity of coding units within the categories of the population are about $p_0 = .674, p_1 = .182$ and $p_2 = .144$. About 67.4%% of the coding units truly belong to category 0, 18.2%% to category 1 and about 14.4%% of the coding units belong to the true category 2. The Assignment Error Matrix in Table 2 implies a Beta Reliability of about 48%% for category 0, of 28%% for category 1 and of 73%% for category 2. The value of category 0 means that in 48%% of cases, Alpha Errors in the *other* categories are *not* assigned to category 0. In contrast, the Beta Reliability of category 1 means that only in 28%% of cases Alpha Errors in the *other* categories are *not* assigned to category 1. In other words: in 72%% of cases, an Alpha Error on coding units belonging truly to category 0 or 2 leads to an assignment of these coding units to category 1. Thus, category 1 suffers more strongly from errors in other categories than category 0.

Table 3. Example of the Different Elements

		α_i^{rel}	β_i^{rel}	p_i	A_i^{rel}	B_i^{rel}
Category	0	.508	.484	.674	.262	-.032
	1	.823	.279	.182	.734	-.443
	2	.763	.723	.144	.644	.451

With the Beta Elements, Alpha Elements and the Assignment Error Matrix, the Iota Concept of the second generation provides detailed insight into the data generated by a coding scheme. For a correct interpretation, it is helpful to compare these values with values which would occur with complete guessing. The following section introduces these values.

3.2 Introducing the Chance-Correction for Alpha and Beta Elements

To compare the quality of a coding scheme with guesswork, the equations of section 3.1 can be applied. Only the Assignment Error Matrix changes. In this case, the cells of the matrix equal $1/c$ as shown in Table 4. Such a matrix assigns the categories to a coding unit completely randomly, regardless of their true category.

Table 4. Example of an Assignment Error Matrix in the Case of Complete Guessing

		Assigned Category		
		0	1	2
True Category	0	$1/3$	$1/3$	$1/3$
	1	$1/3$	$1/3$	$1/3$
	2	$1/3$	$1/3$	$1/3$

With Equation 2, the chance-corrected Alpha Reliability is given by Equation 5.

$$A_i^{rel} = \frac{\alpha_i^{rel} - \frac{1}{c}}{1 - \frac{1}{c}} \quad [5]$$

The denominator is used to normalize the values. Thus, the chance-corrected Alpha Reliability equals 0 in the case that the probability to assign the true category to a coding unit equals guessing. The chance-corrected Alpha Reliability equals 1 in the case that the true coding unit is always recovered.

Similarly, the chance-corrected Beta Reliability can be calculated. The Beta Error under the condition of complete guessing without normalization is given with Equation 6 by

$$b_i^{err} = \frac{\sum_{k \neq i} p_k * \frac{1}{c}}{\sum_{k \neq i} p_k * (1 - \frac{1}{c})} \quad [6]$$

Thus, the chance-corrected normalized Beta Reliability is given by

$$B_i^{rel} = \frac{\beta_i^{rel} - (1 - b_i^{err})}{1 - (1 - b_i^{err})} \quad [7]$$

A chance-corrected Beta Reliability of 0 indicates that the reliability equals complete guessing. A value of 1 indicates perfect reliability meaning that errors in the other categories do not influence the category under investigation. A value below zero indicates that the Beta Reliability is even lower than it should be expected by complete guessing.

The chance-corrected values can be illustrated with the example from section 3.1. Table 3 reports the corresponding values. The chance-corrected Alpha Reliabilities for the three categories are greater than zero, indicating that the coding scheme offers some stable recovery of the true categories. In other

words: the coding scheme recovers the true category better than complete guessing. In particular, the Alpha Reliability of category 1 and 2 is high. Concentrating on the chance-corrected Beta Reliabilities, Table 3 shows that only category 2 has a value above zero, indicating that the coding scheme does not heavily bias the data of this category due to errors in the other categories. The reliability is better compared to complete guessing. In contrast, the Beta Reliabilities of category 0 and 1 are below zero. This means that the data of both of these categories suffers from errors made on the other categories more strongly than expected by complete guessing. In other words: Complete guessing produces a smaller bias than working with this coding scheme.

3.3 Refining Iota

In the Iota Concept of the first generation, the mean of the chance-corrected Alpha Reliability and chance-corrected Beta Reliability is used to characterize the complete amount of reliability of a category. In the second generation, this definition is no longer used since both values are not standardized to values above zero. Instead, a new definition that allows an improved interpretation is introduced. Figure 2 illustrates this redefinition.

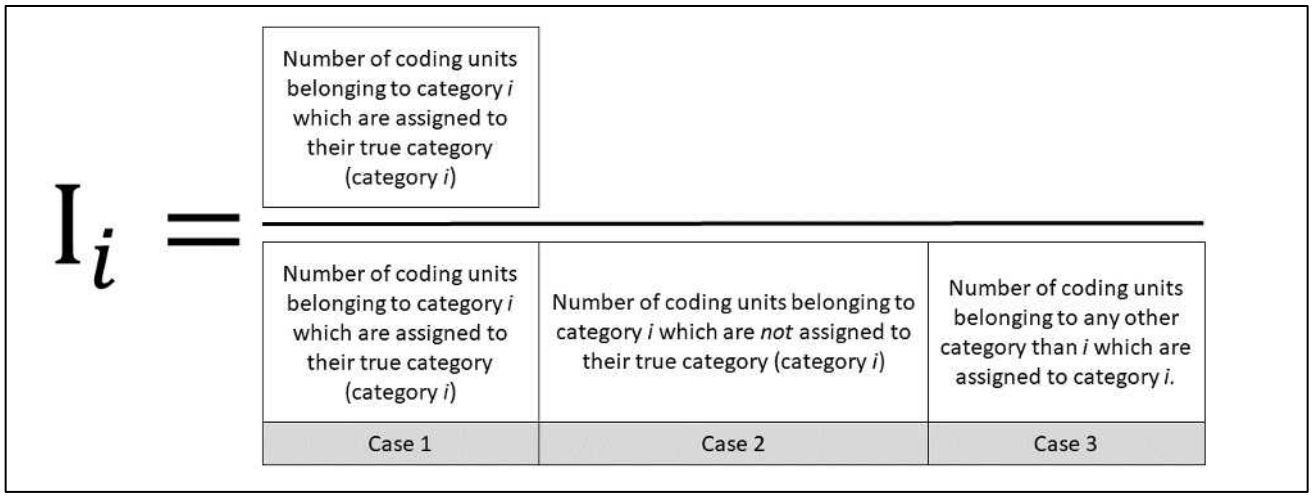


Figure 2. Iota in the Concept of the Second Generation

Three cases can occur when concentrating on the data generated by a coding scheme. First, the true category of a coding unit can be recovered. Thus, the data reflects the true category correctly (Iota). Second, the true category of a coding unit is not recovered. Thus, some observations are missing in the data of that category (Iota Error – Type I). Third, mistakes in the other categories lead raters to assign a coding unit to the category under investigation. Thus, the data of that category comprises too many observations (Iota Error – Type II). For a reliable

reflection of a category, case 1 (Iota) should be maximized and cases 2 (Iota Error – Type I) and 3 (Iota Error – Type II) minimized. With Figure 3, the idea behind Iota can be explained in more detail.

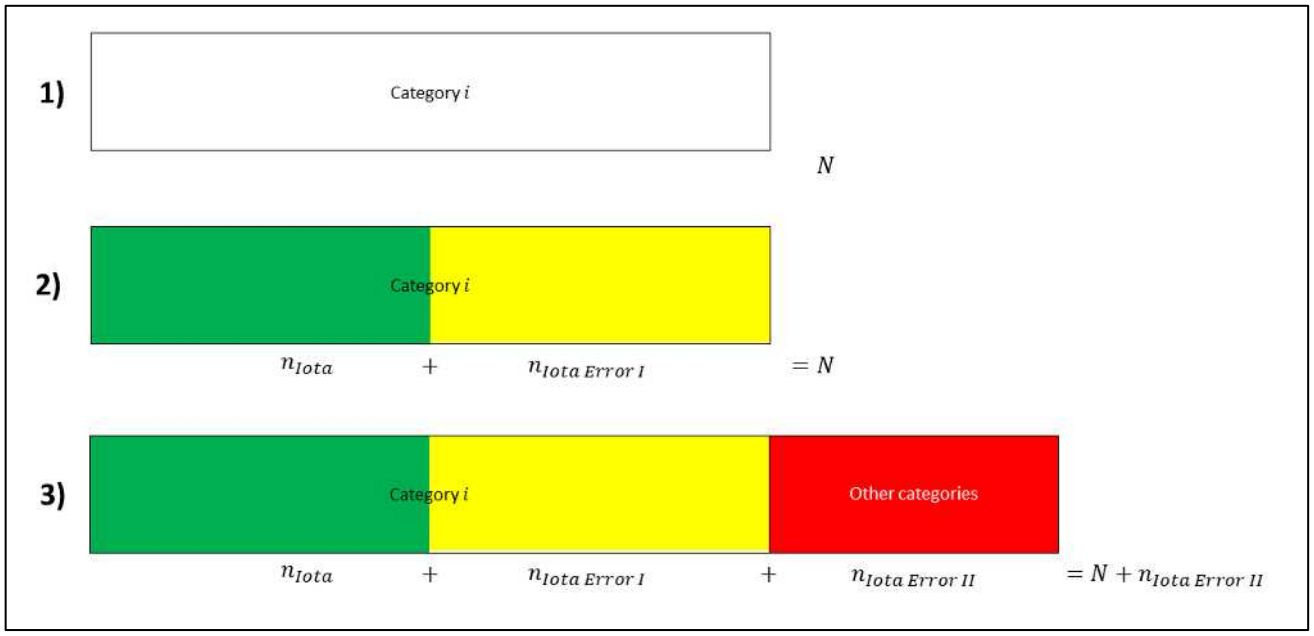


Figure 3. Illustrating Iota and its Errors

The first row in Figure 3 represents all coding units which truly belong to category i. During coding, the true category of these coding units can be either assigned correctly or incorrectly. This is illustrated in the second row of Figure 3. The green part of the bar represents the number of correctly assigned units while the yellow part of the bar represents the coding units which are not assigned to their true category. As a consequence, the data that should represent category i comprises only a part of the truly relevant coding units (in Figure 3 about half of the relevant units). The missing coding units lead to an underestimation of the number of units belonging to category i. Additionally, errors made in the other categories can lead raters to assign coding units to category i which do not belong to the data of category i. This is illustrated by the red part of the bar in the third row of Figure 3. Thus, the red part of the bar represents the number of coding units that belong to other categories and thus are incorrectly assigned to category i. These coding units contribute to an overestimation of the number of coding units of category i. In Figure 3, the green part of the bar of row three refers to Iota, the yellow part of the bar of row three to Iota Error – Type I and the red part of the bar of row three to Iota Error – Type II.

The new Iota takes these suggestions into account and describes the number of correctly assigned coding units on all assignments referring to that category. Mathematically, this can be expressed with Equation 8.

$$\mathrm{I}_i = \frac{p_i * \alpha_i^{rel}}{p_i * \alpha_i^{rel} + p_i * \alpha_i^{err} + b_i^{err} * \sum_{k \neq i}(p_k * \alpha_k^{err})} \qquad [8]$$

Figure 4 shows the results for the example from section 3.1. The green rectangle represents the Iota value while the orange and red rectangles characterize the amounts of cases 2 and 3.

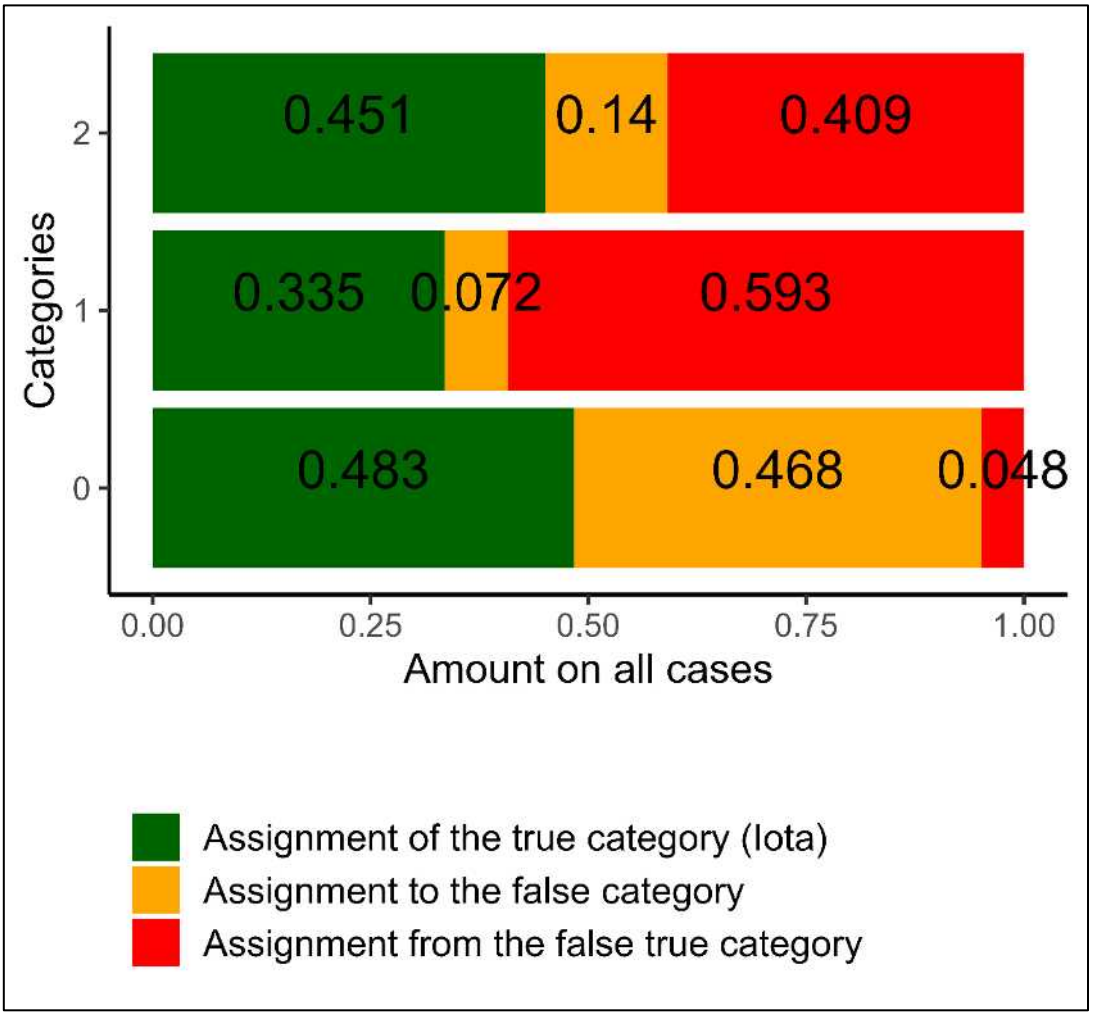


Figure 4. Example for Iota of the Second Generation

In Figure 4 it is visible that Iota is the highest for category 0. The data representing category 0 is composed of 48% of coding units of category 0. About 47% of the data are missing coding units of that category. Thus, only about half of the coding units truly belonging to category 0 are represented in the data ($\frac{.48}{.48+.47} = .505$).

About 5% of the data representing category 0 is made up of coding units from other categories.

Regarding category 1, a Iota value of .335 indicates that about one-third of the data representing this category is made up of the correct coding units. Still, about 7% of the relevant data are missing coding units. Thus, the data representing category 1 comprises about 82% of the coding units truly belonging to category 1 ($\frac{.335}{.335+.072} = .823$). However, Iota Error – Type II is about 60%. This indicates that more than half the coding units representing category 1 are truly coding units belonging to other categories. As a consequence, the total number of coding units in category 1 is overestimated.

For category 2, the situation is similar. About 41% of the coding units forming the data for category 2 truly belong to another category while only about 14% of the data are "forgotten" coding units with the correct true category. In line with these results, Iota is about .451. Thus, the data of this category is heavily biased by errors in the other categories.

Referring to Table 3, Iota is significantly smaller than the average Alpha and Beta Reliabilities for category 2. The reason for this interesting result is that the coding scheme leads raters to assign a part of the coding units truly belonging to categories 0 and 1 to category 2 and the size of category 0 is meaningfully greater than the size of category 2 (see Table 3). Due to the large size of category 0, a great number of coding units is assigned to a category of a very small size. This leads to a high number of errors in the data that should represent category 2.

As this example shows, the Iota concept provides a deep insight into the quality of a coding scheme. The following section describes how the reliability of the different categories is summarized to a reliability measure for the whole scale.

3.4 Reliability on the Scale Level

For describing the reliability of a whole scale, we introduce the new Iota Index. The idea behind this approach is to compare the scale-specific Assignment Error Matrix with the Assignment Error Matrix that implies a completely random assignment of the categories. Figure 5 illustrates this idea for three categories.

Perfect Reliability $aem_1 = \begin{pmatrix} 1.00 & 0.00 & 0.00 \\ 0.00 & 1.00 & 0.00 \\ 0.00 & 0.00 & 1.00 \end{pmatrix}$

Absence of Reliability $aem_2 = \begin{pmatrix} 0.3\overline{3} & 0.3\overline{3} & 0.3\overline{3} \\ 0.3\overline{3} & 0.3\overline{3} & 0.3\overline{3} \\ 0.3\overline{3} & 0.3\overline{3} & 0.3\overline{3} \end{pmatrix}$

Idea for the Degree of Reliability $aem_3 = \begin{pmatrix} a_{1,1} & a_{1,2} & a_{1,3} \\ a_{2,1} & a_{2,2} & a_{2,3} \\ a_{3,1} & a_{3,2} & a_{3,3} \end{pmatrix} - \begin{pmatrix} 0.3\overline{3} & 0.3\overline{3} & 0.3\overline{3} \\ 0.3\overline{3} & 0.3\overline{3} & 0.3\overline{3} \\ 0.3\overline{3} & 0.3\overline{3} & 0.3\overline{3} \end{pmatrix}$

Figure 5. Illustration of the Idea for Measuring the Reliability on a Scale Level

In Figure 5, the first matrix represents the case of perfect reliability where every coding unit is assigned to its true category. The second matrix in Figure 5 shows the case for the absence of the reliability. In this situation, every coding unit is assigned completely randomly to a category. Thus, the degree of reliability can be characterized by measuring the distance between the estimated Assignment Error Matrix from the matrix representing absence of reliability. In that case, the greater the estimated matrix differs from the matrix of perfect absence of reliability, the more reliable the generated data. The following equation shows the corresponding Iota Index which summarizes the distance into one single value.

$$Iota_{Index} = 2\frac{c-1}{c} * \sum_{\forall i} p_i \left(\sum_{\forall j} \left| a_{ij} - \frac{1}{c} \right| \right) \qquad [9]$$

On the right side, Equation 9 contains the term for measuring the distance between single elements of the estimated Assignment Error Matrix from the corresponding single element of the matrix for random assignment ($\left|a_{ij} - \frac{1}{c}\right|$). These values are summarized for each column of a row and weighted with the categorical size of the corresponding category. The reason for weighting is representation: the degree of reliability of large categories should be represented stronger in a value for the complete scale than the degree of reliability for small categories. Finally, the values for each row/category are summarized to achieve a measure of reliability for the whole scale. The term

$2\frac{c-1}{c}$ is used to standardize the range of possible values to 0 and 1. Thus, a value of 0 implies that the Assignment Error Matrix equals the matrix for random assignments while a value of 1 implies that the estimated Assignment Error Matrix has the greatest possible distance from "randomness". That is, the Assignment Error Matrix represents perfect reliability on the scale level.

3.5 Assumption of Weak Superiority

The elements of the Iota Concept have their roots in the Assignment Error Matrix and the categorical sizes. In order to produce meaningful results and to provide a high chance of unique estimates for a given data set, the Assignment Error Matrix must follow a specific structure. Figure 6 visualizes the structure that is created by the introduction of an additional assumption called *weak superiority*.

Weak Superiority fulfilled

Perfect Reliability

$$aem_1 = \begin{pmatrix} 1.00 & 0.00 & 0.00 \\ 0.00 & 1.00 & 0.00 \\ 0.00 & 0.00 & 1.00 \end{pmatrix}$$

Absence of Reliability

$$aem_2 = \begin{pmatrix} 0.3\overline{3} & 0.3\overline{3} & 0.3\overline{3} \\ 0.3\overline{3} & 0.3\overline{3} & 0.3\overline{3} \\ 0.3\overline{3} & 0.3\overline{3} & 0.3\overline{3} \end{pmatrix}$$

$$aem_3 = \begin{pmatrix} 0.5 & 0.5 & 0 \\ 0.25 & 0.5 & 0.25 \\ 0.25 & 0.25 & 0.5 \end{pmatrix}$$

Weak Superiority not fulfilled

$$aem_4 = \begin{pmatrix} 0.25 & 0.5 & 0.25 \\ 0.25 & 0.5 & 0.25 \\ 0.25 & 0.25 & 0.5 \end{pmatrix}$$

Figure 6. Illustrating the Assumption of Weak Superiority

In Figure 6, the first aem shows the situation for three categories with perfect reliability. The true category is recovered and no assignments to the wrong categories occur. The second aem is an example of a situation with absence of reliability. Every category is assigned completely randomly. The third and fourth

matrix describe situations between these two extremes. The third is in line with the assumption of weak superiority, the fourth is not. The assumption of weak superiority demands that a value on the diagonal is at least as high as the other values on the corresponding row. Formula 10 expresses this requirement

$$a_{i,i} \geq a_{i,j} \qquad [10]$$

The reason for the introduction of this assumption is the range of possible categories within a coding scheme, as it is a constructive process of the users of content analysis. It seems clear that the different categories must be coherent and that they reflect *different* characteristics or levels of a phenomenon. In situations like the fourth aem, coding units belonging to category 0 are more often assigned to category 1 than to category 0. Thus, the raters agree that these units represent coding units of category 1. As the existence of a category is based on the agreement of the users of content analysis, this would imply that category 0 and 1 are the same. Thus, there is an incoherence in the differentiation of categories in the coding scheme. In this case it is logical to use only two different categories rather than three.

The entailing problem can be solved with the assumption of weak superiority. It ensures that the different categories do not clash into each other and makes a distinct estimation. Before the following chapter describes how the Assignment Error Matrix and the categorical sizes can be estimated in order to fulfill weak superiority, the next section compares both concepts of Iota in order to make the differences between both concepts clear.

3.6 Comparing the Iota Concepts

The Iota Concepts of the first and the second generation share most of their components. However, some components are redefined and imply a different meaning in the second generation. To prevent confusion, the following Table 5 compares the meaning of the different components.

Table 5. Comparison of the First and Second Iota Generation

Component and Characteristics	First Generation	Second Generation
Assumptions	Assumptions 1 to 6	Assumptions 1 to 6 Weak Superiority
Alpha Reliability	Probability that all raters agree on a category under the condition that that at least one rater assigns the category to a coding unit.	Probability to assign a coding unit of category i to category i.
Alpha Error	Complementary probability of the Alpha Reliability.	Probability to assign a coding unit of category i to another category as i.
Beta Reliability	Complementary probability of the Beta Error.	Probability that coding units truly belong to any other categories as i are *not* assigned to category i.
Beta Error	Probability that at least one rater assigns a coding unit to a category (without the case that all raters agree on that category) under the condition that an Alpha Error occurs in all other categories.	Probability that coding units truly belong to any other categories as i are assigned to category i.
Iota	Mean of the chance-corrected Alpha and Beta Reliability.	Percentage of cases involving category i that correctly represent category i.
Iota Error Type I	-/-	Percentage of cases involving category i that are missing in the data representing category i.
Iota Error Type II	-/-	Percentage of cases involving category i that belong to other categories than i.
Diagonal of the Assignment Error Matrix	Alpha Errors of the specific category.	Alpha Reliability of the specific category.
Other Cells in a Row of the Assignment Error Matrix	Probability to assign a category to another category under the condition that an Alpha Error occurs.	Probability to assign a coding unit of category i to category j.
Average Iota	Mean of Iota. Index for describing the reliability of the complete scale.	-/-
Minimum Iota	Minimum of Iota. Index for describing the reliability of the complete scale.	-/-
Iota Index		Distance between the estimated Assignment Error Matrix and an Assignment Error Matrix representing complete guessing. Index for describing the reliability of the complete scale.

Second generation Iota founds on the same assumptions as the first generation, but in addition introduces weak superiority as a new demand for the Assignment Error Matrix. With the help of the additional assumption, the different components of the concept are redefined with more clarity. For example, in the first generation, the Assignment Error Matrix comprises the Alpha Error and a complex combination of Alpha and Beta Errors. In the new version, the Assignment Error Matrix simply describes the probability to assign a coding unit of category i to category j. Thus, the interpretation of the Assignment Error Matrix is clearer in the second generation.

The clear definition of the Assignment Error Matrix leads to improved definitions of the other components. While in the old concept, Alpha Reliability is defined as the probability that all raters agree on a category under the condition that at least one rater assigns the category to a coding unit, the new concept just defines it as the probability to assign a coding unit of category i to category i. The idea behind both definitions is the same, but in the new concept the definition is much clearer.

The complicated conceptualization of the first generation is due to the attempt to derive the estimates directly from frequencies tables. For the new generation, an estimation algorithm, which uses techniques of Maximum Likelihood Estimation, is developed and applied. The following section introduces this new algorithm.

4 Estimation and Log-Likelihood

In comparison to the first generation of the concept, the estimation of the relevant values in the second generation is more complicated. All measures described in section 3 build on the Assignment Error Matrix aem and the sizes of the true categories in a population p.

The current approach for estimating these values uses an analogy to the estimation within latent class analysis (LCA). This provides a matrix that shows the average latent class probabilities for the most likely latent class membership for each latent class (Geiser, 2013, p. 248) that has a similar interpretation as the Assignment Error Matrix. In LCA, this matrix characterizes the probability to assign individuals to the different classes based on their true class membership. In the suggested reliability concept, the Assignment Error Matrix does the same. It characterizes the probabilities to assign the different categories to a coding unit based on the true category of the coding unit. In LCA, the sizes of the different classes have to be estimated (Andreß et al., 1997, pp. 211–218), similarly to the categorical sizes in the proposed concept.

LCA is based on the assumption that the latent class membership influences the observable patterns on different items (Andreß et al., 1997, pp. 212–214). This idea can be transferred to the reliability concept as shown in Figure 7 by analyzing the coding patterns different raters produce.

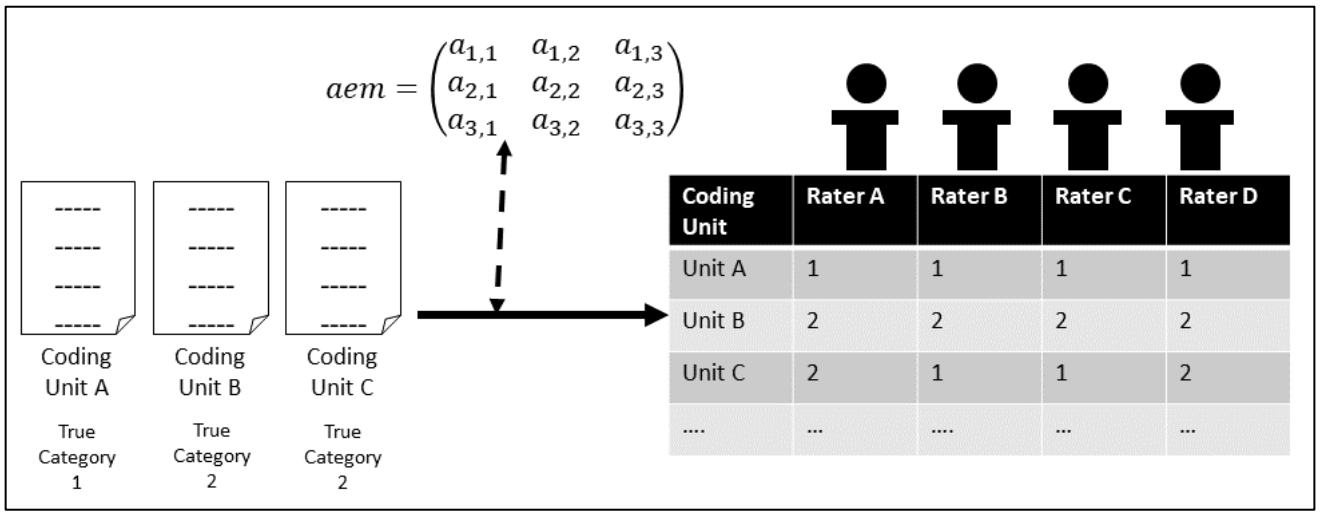


Figure 7. Analogy of the Concept to Latent Class Analysis

In the terminology of LCA, the assignment of different categories to a coding unit made by the raters represents the manifest variables. For example, rater A can assign category 0, 1 and 2 to coding unit A. Raters B, C and D can do the same.

Thus, there are four manifest variables. The pattern of assignments expressed by these four variables/raters is influenced by the true categories of the coding units and the quality of the coding scheme, represented by the Assignment Error Matrix. In consequence, there is a link between the observable assignment patterns on the one hand and the Assignment Error Matrix on the other hand. In terms of LCA, the number of categories represents the number of classes (the number of levels of the latent variable).

To sum up, the estimation of the Assignment Error Matrix can be interpreted as a kind of latent class analysis where the number of classes equals the number of categories and the number of manifest variables equals the number of raters. Furthermore, the number of characteristics/levels of the manifest variables equal the number of categories.

The analogy to LCA allows the application of the Expectation Maximization Algorithm (EM Algorithm) as described in Andreß et al. (1997, pp. 218–226). The EM Algorithm is very useful in practical applications because it is relatively independent from the concrete set of starting values, offers fast computations within each iteration and provides a high chance of convergence. The disadvantages are that it does not provide standard errors and does not allow specifications for each kind of restrictions (Andreß et al., 1997, pp. 220–221). In the following, the EM Algorithm is transferred into the notation of the Iota Concept. Please note that the rules for identification of LCA do not apply to the Iota Concept.

Expectation Stage

The EM Algorithm starts with a randomly chosen set of values for the aem and the categorical sizes p. The first step is to estimate the expected frequencies of the different assignment patterns (E-Step). This requires a list of all unique observed patterns represented with π. For example, in Figure 7, three different patterns can be seen on the right side of the table: $\pi_1 = 1{,}1{,}1{,}1$, $\pi_2 = 2{,}2{,}2{,}2$, $\pi_3 = 2{,}1{,}1{,}2$. How often these patterns can be observed in the data is described with n_π. Referring to the table in Figure 7, $n_{2,1,1,2} = 1$. The number of coding units is represented with N.

First, the *conditional* probabilities of the different assignment patterns are calculated. Equation 11 provides the conditional probability for a concrete pattern π under the condition that the true category is t.

$$P_{\pi t} = p_t \prod_{j=1}^{r} a_{i(t),i_\pi(j)} \quad [11]$$

In Equation 11, $i_\pi(j)$ represents the index of the Assignment Error Matrix for the category that rater j has assigned in a specific pattern. For example, if the pattern is $\pi = 2,1,1,2$, then the third rater assigned category 1. Category 1 refers to the second column in the Assignment Error Matrix since in the example the possible categories are 0,1 and 2. Thus, $i_{\pi=2,1,1,2}(3) = 2$. Something similar applies to $i(t)$ which refers to the corresponding row or column in the Assignment Error Matrix for the true category t.

The unconditioned probability for an assignment pattern is given by the sum over all true categories t (Equation 12).

$$P_\pi = \sum_{\forall t} P_{\pi t} \quad [12]$$

With the help of these probabilities, the expected pattern frequency within each true category can be calculated given by Equation 13. In this equation, n_π represents the observed frequency of pattern π.

$$\hat{n}_{\pi t} = \frac{n_\pi * P_{\pi t}}{P_\pi} \quad [13]$$

Maximization Stage

Based on these results, the estimates are improved (M-Step). The new estimate for the categorical size of category t is given by Equation 14.

$$\hat{p}_t = \frac{1}{N} \sum_{\forall \pi} \hat{n}_{\pi t} \quad [14]$$

The new entries in the Assignment Error Matrix are given by Equation 15. In this equation, t represents the true categories and c the assigned categories. $\phi_{\pi,c}$ describes the amount of category c within the pattern π. For example, referring to Figure 7, $\phi_{\pi=1,1,1,1,c=0} = 0$ and $\phi_{\pi=1,1,1,1,c=1} = 1$.

$$\hat{a}_{t,c} = \frac{\sum_{\forall \pi} (\phi_{\pi,c} * \hat{n}_{\pi t})}{\hat{p}_t * N} \quad [15]$$

The new values are now used as new starting values for the algorithm. The EM Algorithm ensures that the quality of the estimates increases in every iteration.

To characterize the quality of the estimates, the likelihood has to be calculated. The corresponding log-likelihood for a given aem, p and the observed assignment patterns π is characterized by

$$LL = -\sum_{\forall \pi} n_\pi * \log\left(\sum_{\forall t} p_t \left(\prod\nolimits_{j=1}^{r} a_{i(t),i_\pi(j)}\right)\right) \qquad [16]$$

Conditioning Stage

The EM Algorithm described above tries to find the most plausible estimates. In general, the EM Algorithm will not produce estimates of the Assignment Error Matrix that are in line with the assumption of weak superiority. Thus, a new stage is added to the algorithm above for the case that an iteration does produces an incorrectly structured aem.

Finding estimates of the aem that are in line with weak superiority represent an estimation problem with inequality constraints. To solve this problem, the suggested approach uses a local optimization by adapting the rows of the aem to fulfill weak superiority and to describe the patterns belonging to that category as plausible as possible.

During the conditioning stage, a row t of an aem is interpreted as a multinomial distribution with the probabilities m_i. The number of observed categories within a true category is given by the corresponding rows of the aem from the maximization stage. These are derived from Equation 15. Thus, the log-likelihood can be described by Equation 17.

$$LL_t = -\sum_{i=1}^{c} \hat{a}_{t,c} \log(m_i) \qquad [17]$$

During the conditioning stage, the values for m_i are varied to maximize likelihood. For this aim, the gradient of LL_t is used. The entry i of the gradient is given by Equation 18.

$$grad(LL_t)_i = \frac{\hat{a}_{t,i}}{m_i} - \frac{\hat{a}_{t,c-1}}{1 - \sum_{j=1}^{c-1} m_j} \qquad [18]$$

Please note that i ranges from 1 to $c - 1$ since the sum of all probabilities must equal one, reducing the dimensionality of the optimization problem. With the

help of the gradient and the log-likelihood, the following algorithm produces estimates for a row of the aem that are in line with weak superiority.

1) Chose the row t of the aem that is supposed to be estimated.
2) Reorder the columns of the row so that column t is the first column.
3) Reorder the rest of the columns so that the column with the smallest value $\hat{a}_{t,i}$ is the last column. The reordered values form the observations for which the values of m_i should be optimized.
4) Choose a set of start values for m that is in line with weak superiority ($m_1 \geq m_j \; \forall j \neq 1$).
5) Calculate the log-likelihood with Equation 17 for m based on the observations.
6) Calculate the gradient of log-likelihood for m with Equation 18.
7) If the gradient for m_i is greater than zero, reduce m_i by δ_1.
   If the gradient for m_i is smaller than zero, increase m_i by δ_1.
   if the gradient for m_i equals zero, do not change m_i.
8) Check the new value for m_1.
   If $m_1 < \frac{1}{c}$, set $m_1 = \frac{1}{c}$.
   If $m_1 > 1, set\ m_1 = 1$.
9) Now check the other probabilities
   If $m_i > m_1$, set $m_i = m_1$.
   If $m_i < 0$, set $m_i = 0$.
10) Set $m_c = 1 - \sum_{j=1}^{c-1} m_j$
11) Calculate the log-likelihood with Equation 17 for the new m based on the observations.
12) Compare both values for the log-likelihood. If the log-likelihood for the new values of m only slightly increases compared to log-likelihood for the old values, go to step 13. In the other case restart the algorithm with step 5 and the new values for m.
13) Reorder the columns of m. Set the last column to the original column from step 3.
14) Reorder the columns for m. Set column 1 to column t (see step 2).
15) Use m as new values in the aem for row t.

The reorder of the columns is necessary to provide the same algorithm for all rows of an aem. That is, for all true categories. Step 3 is crucial since the

suggested algorithm explicitly investigates only $c - 1$ entries. By assigning the column with the smallest value to the last column, the algorithm indirectly proves that the result of the estimation is in line with weak superiority. The following section describes the design of a simulation study to investigate the quality of the estimation algorithm and to provide insights into the statistical properties of the new concept.

5 Simulation Study I

5.1 Hypotheses and Design of Simulation Study I

To prove the new concept and its estimation algorithm, a simulation study is performed. This study concentrates on the quality of the estimates on the one hand and on the other hand on the ability to characterize the reliability of the complete scale. First, the following basic hypotheses about the estimates will be investigated:

Categorical Level

H1: Increasing the sample size leads to more accurate estimates for the Assignment Error Matrix and the categorical sizes.
H2: Increasing the number of raters leads to more accurate estimates of the Assignment Error Matrix and the categorical sizes.
H3: The number of categories decreases the accuracy of the estimates.
H4: The kind of distribution of categories in the population does not influence the estimates of the Assignment Error Matrix and the categorical sizes.

Scale Level

H5: Increasing the sample size leads to more accurate estimates for the Iota Index.
H6: Increasing the number of raters leads to more accurate estimates for the Iota Index.
H7: The number of categories decreases the accuracy of the Iota Index.
H8: The kind of distribution of categories in the population does not influence the estimates for the Iota Index.

As more raters and a greater sample size provide more information, the accuracy of the estimates should increase. A higher number of categories implies more parameters to be estimated, leading to a decreasing accuracy for a given sample size and a given number of raters. In contrast, the kind of distribution of the categories in a population should ideally not influence the accuracy. In particular the independence of the kind of distribution would be a great benefit for users as it would ensure more valid conclusions in practical applications, since the true distribution is seldomly known. Figure 8 illustrates the design of the study.

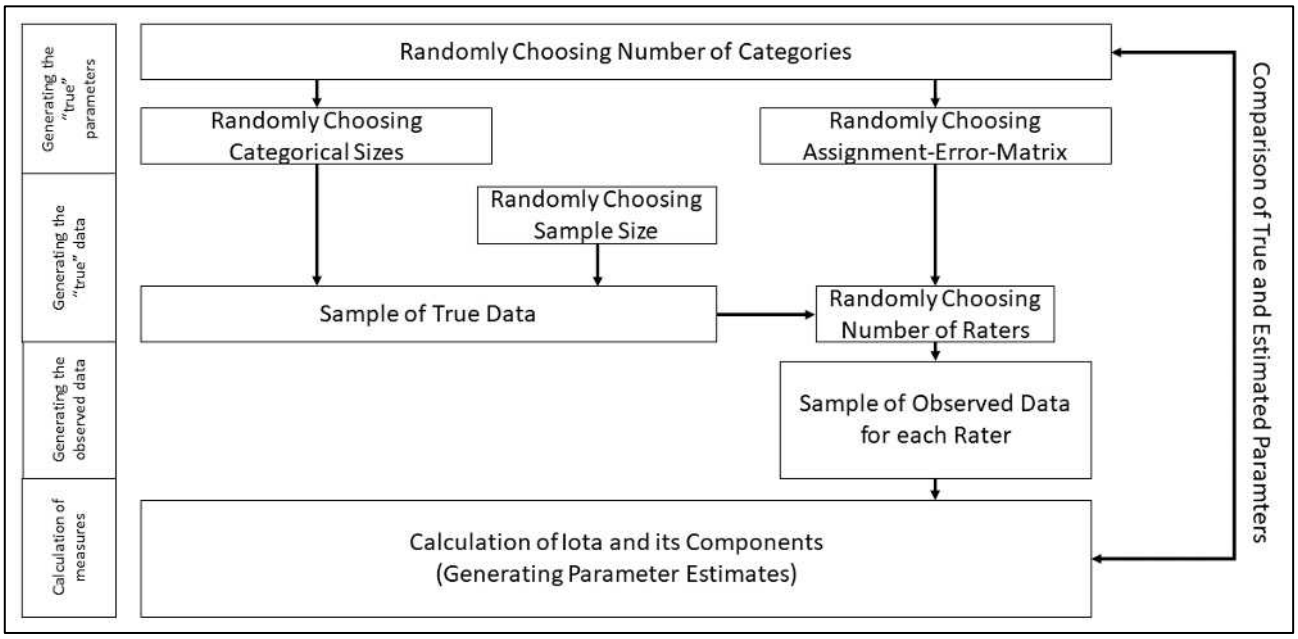


Figure 8. Design of Simulation Study I.

The simulation begins with the generation of the true parameters. In a first step, this implies the selection of the number of categories of a coding scheme. The number of categories in this simulation varies between two and five. The specific number is determined by chance. After this step, the true parameters for the categorial sizes and the Assignment Error Matrix have to be chosen. The categorical sizes represent the probability of a multinomial distribution, which is randomly chosen with each probability greater than zero. Similarly, the true parameters of the Assignment Error Matrix are randomly chosen but it is ensured that the matrix is in line with weak superiority in every case. Ten percent of all parameter sets refer to an Assignment Error Matrix representing the absence or perfect reliability.

After modeling the coding scheme, a sample of true data is generated. The corresponding sample size can vary between 20 and 1,500 coding units. Again, the specific size is selected by chance.

In the next step, the number of raters is determined by chance. The number can vary between two and five. Each rater assigns a category to every coding unit. The probability for assigning the categories to a coding unit is determined by the true Assignment Error Matrix. This process results in a table where every rater assigned a category to every coding unit. This table forms the observed data.

The observed data is used for the calculation of Iota of the second generation and its components, resulting in estimates for the Assignment Error Matrix and

the categorical sizes. Now it is possible to compare the estimated with the true parameters and to analyze the influence of the number of raters, number of categories and sample size on the accuracy of the estimates.

In addition, the true distribution is characterized by the Herfindahl Index. Thus, it can be investigated, what influence the shape of the distribution has on the estimates.

In this simulation study, 11,520 different sets of Assignment Error Matrices and categorical sizes are generated. For every set of true parameters, a sample of four different numbers of raters is drawn. For every number of raters, the simulation chooses ten different sample sizes. Thus, at the end, 460,800 cases are available to investigate the properties of the Iota Concept of the second generation

Since the generated data always refers to one of the 11,520 sets of true parameters, the data has a clustered structure. The sets of true parameters form level 1 and the nested coding processes form level 2. Level 3 is made up of the parameters of the Assignment Error Matrix, the categorical sizes, the Alpha Elements, Beta Elements and Iota within each coding process.

To take this structure into account in the evaluation of accuracy, the following analyses are done in *R* (R Core Team, 2021) with help of the package *estimatr* (Blair et al., 2022), which allows for cluster adjusted regression. Furthermore, the software *Mplus 8.8* (Muthén Linda & Muthén, 2022) is used in combination with *R* (Hallquist & Wiley, 2018), which allows for an additional and even deeper analysis of these data structures by including the framework of structural equation modeling (SEM) and by expending the number of cluster levels up to three. Heck and Thomas (2020, pp. 33–35) summarize the advantages of multilevel modeling: the possibility to model variables on their correct level, allowing for a more complete specification of errors and providing more accurate standard errors. In consequence, ignoring the clustered structure generally leads to smaller standard errors, in turn leading to a higher chance of a wrongful confirmation of the research hypotheses although in truth, they are false (Heck & Thomas, 2020, p. 33).

In the following analysis, no centering is applied since the number of raters, the number of categories, the concentration, the sample size and the true reliability are scales with a meaningful zero point. In these cases, centering is not necessary

(Heck & Thomas, 2020, p. 74). The analysis applies Bayes Estimation since it provides a lot of advantages. These include more realistic predictions, the possibility of modelling more complex structures and avoiding estimation problems (Wang & Wang, 2020, pp. 17–18). To evaluate the global fit, the Posterior Predictive P-value (PPP) can be used. The PPP describes the amount in which the model-generated data are more plausible than the empirical data. A PPP of .500 means that the model-generated data is equally plausible as the observed data (Zyphur & Oswald, 2015, p. 402), pointing to an excellent fit (Wang & Wang, 2020, p. 26). A value of at least .100 (Cain & Zhang, 2019, p. 48) or of at least .05 (Wang & Wang, 2020, p. 26) indicates an acceptable level of global fit. The PPP is advantageous for small samples but tends to over-reject a “good” model in increased sample sizes (N > 1,000), leading to the rejection of a “good” model due to only small misspecifications. Thus, Hoofs et al. (2018) suggest the Bayesian root mean square error of approximation to solve this situation. However, such alternative fit indices are currently not available for the present kind of fitted model in this study. All simulations in this study were performed using the High Performance Computing Cluster "Hummel" at the University of Hamburg. The following section presents the results.

5.2 Results of Simulation Study I

5.2.1 Data Description and Preparation

The simulation generated 11,472 true parameter sets (Cluster Level 1) and simulated about 39.50 coding processes per parameter set, resulting in a total of 453,140 coding processes (Cluster Level 2). In every coding process, the parameters for the Assignment Error Matrix and the categorical sizes are estimated. This resulted in 7,743,512 single parameter estimates (Cluster Level 3).

To provide first insights into the quality of the generated data, Figure 9 shows the scatter plot for the estimated and true Iota Index. The Iota Index is used since it is a reliability measure for the complete scale and depends on both the parameters of the Assignment Error Matrix and the categorical sizes. Thus, abnormalities in this measure have a high chance to reflect problems in the estimation process.

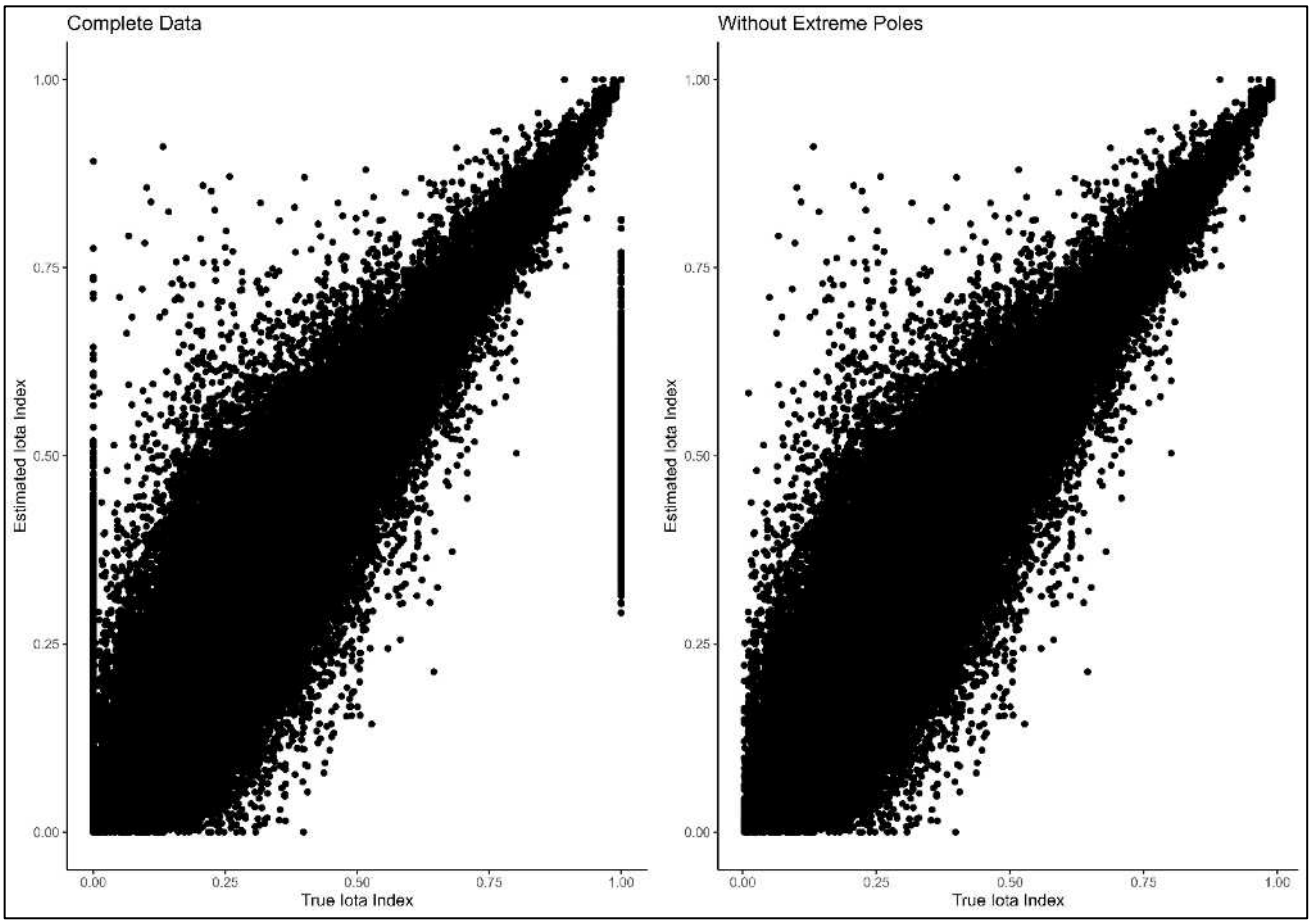


Figure 9. Scatter Plot for the Iota Index

Figure 9 shows that the estimates become more accurate the higher the true Iota Index is. In particular, the points deviate nearly symmetrically from the true values. However, on the extreme poles of the scale, a high number of outliners can be identified. This suggests that the current implementation of parameter estimation is problematic for the extreme cases, or the cases of perfect absence of reliability and for perfect reliability. In order to ensure stable results, these extreme situations are omitted from further analyses (true Iota Index equals 1 or 0). After omitting these cases, the data consist of 10,514 true parameter sets (Cluster Level 1) with about 39.50 coding processes per set, resulting in 415,291 simulated coding processes (Cluster Level 2). Altogether, 7,093,054 parameters are estimated. The following section analyses the accuracy of the primary parameters.

5.2.2 Accuracy of the Estimated Assignment Error Matrix and Categorical Sizes

During the estimation, about 1.1% of the estimated parameter sets showed boundary values. Thus, the risk of generating estimates with categorical sizes is

less than 1%. This is very low and indicates that the estimates are plausible values in terms of content.

Table 6. Descriptive Statistics of Estimation Errors

	Total Error	Deviation between True and Estimated Parameter
$N_{Cluster\ Level\ 1}$	10,514	10,514
$N_{Cluster\ Level\ 2}$	415,291	415,291
$N_{Cluster\ Level\ 3}$	-/-	7,093,054
Minimum	0.001321	.000000
25% Percentile	0.644948	.021500
Median	1.377410	.054250
Mean	1.405852	.082310
75% Percentile	2.065311	.114780
95% Percentile	2.914430	.252408
Maximum	5.885020	.969520
Standard Deviation	0.891520	.086183

To characterize the accuracy of the estimates, the total error is calculated. In this analysis, the total error is given as the sum of the pairwise distance between the true and the corresponding estimated parameter. Table 1 shows basic statistics. On average, the total error of estimation is about $M = 1.405852$ with a standard deviation of $SD = 0.89152$. To provide deeper insight into the total error, Table 7 presents the results of a hierarchical regression analysis.

Table 7. Influences on the Total Error of Estimation

	Model 1		Model 2		Model 3		Model 4		Model 5	
	R^2	.0176	R^2	.0214	R^2	.0252	R^2	.0339	R^2	.0424
	b	β	b	β	b	β	b	β	b	β
True Reliability (True Iota Index)	-0.112	-0.133	-0.117	-0.139	-0.129	-0.153	-0.129	-0.153	-0.129	-0.153
Concentration			0.045	0.061	0.029	0.040	0.030	0.040	0.030	0.041
Number of Categories					-0.006	-0.068	-0.006	-0.068	-0.006	-0.068
Number of Raters							-0.007	-0.093	-0.007	-0.093
Sample Size									0.000	-0.092

Notes: All coefficients are significant at 0.1%.
Cluster variable: true parameter set (Level 1).

According to Table 7, the variance in the total sum of errors can be explained up to about 65%. The strongest impact comes from the number of categories, explaining about 44.89% of the total variance. This implies that the higher the number of categories, the higher the total estimation error. The variable with the second strongest impact is the level of true reliability measured by the true Iota Index, accounting for about 12.70% of the total variance. This suggests that the higher the true reliability, the more accurate the estimates, implying that the estimation algorithm produces more accurate results for situations with a high true reliability. This is plausible as the variance of ratings is higher the more the true reliability converges to situations of perfect absence of reliability. The shape of distribution measured by the Herfindahl Index is less important and explains only about 2.92% of the total variance; the more the values concentrate on a single category, the less accurate the estimates. In addition, the number of raters and the sample size have a positive impact on the accuracy, so the higher the sample size and the more raters are involved in rating the coding units, the more accurate the results. The number of raters explains about 2.38% and the sample size 2.29% of the variance of the total error. Thus, the impact of the number of categories is about 4.7 times higher than the effect of the number of raters and about 4.8 times higher than the effect of the sample size.

Besides the total error, the deviation of single parameters from their true value is important as well, since the Assignment Error Matrix and the categorical sizes can be directly used to evaluate a coding scheme. Table 6 provides the basic statistics. On average, the estimated and the true parameters deviate by 8 percentage points, implying that the estimated value is about 8 percentage points higher or lower than the true counterpart. With a certainty of 75%, the

deviation is not exceeding about 11 percentage points and with a certainty of 95%, the deviation is less than 25 percentage points.

Table 8. Influences on the Deviation Between True and Estimated Parameters

	Model 1		Model 2		Model 3		Model 4		Model 5	
	R^2	.0176	R^2	.0214	R^2	.0252	R^2	.0339	R^2	.04237
	b	β	b	β	b	β	b	β	b	β
True Reliability (True Iota Index)	-0.112	-0.133	-0.117	-0.139	-0.129	-0.153	-0.129	-0.153	-0.129	-0.153
Concentration			0.045	0.061	0.029	0.040	0.03	0.04	0.030	0.041
Number of Categories					-0.006	-0.068	-0.006	-0.068	-0.006	-0.068
Number of Raters							-0.007	-0.093	-0.007	-0.093
Sample Size									0.000	-0.092

Notes: All coefficients are significant at the 0.5% level.
Cluster variable: coding processes (Level 2).

Table 8 provides insight into factors influencing the deviation for the corresponding pairs of true and estimated parameters. In general, the explained variance of all variables is quite low with a R^2 less 5%. Again, the deviation is stronger if the true reliability is low. Except for the concentration of the categories, all other variables contribute to more accurate estimates. The higher the number of categories, the more raters judge the coding units and the greater the sample size, the more accurate the single parameters. The effects of the number of raters and the sample size have an equal relevance.

In order to take the clustered structure into account more clearly, a three level structural equation model is fitted to the data by using *Mplus* 8.8 and Bayes Estimation (Muthén Linda & Muthén, 2022). Figure 10 shows these results.

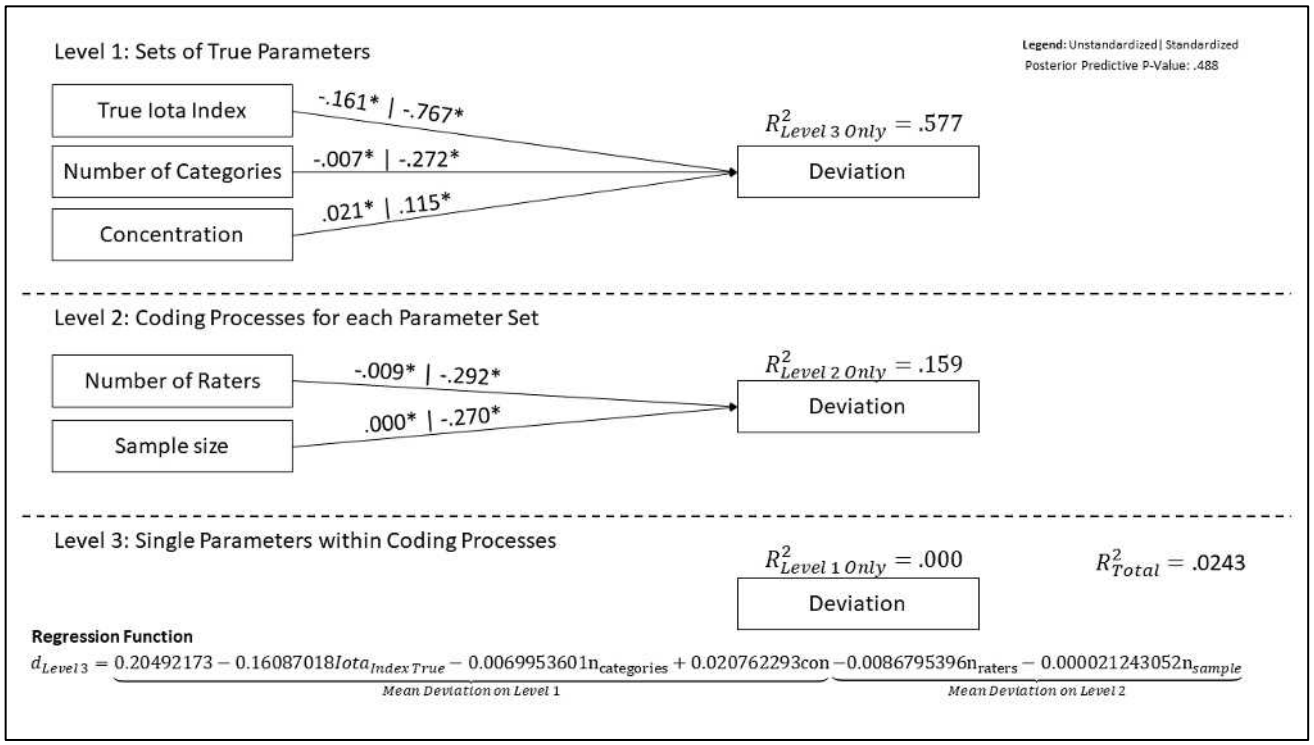


Figure 10. Summary of Multilevel Regression for Deviation of Single Parameters

On the level of true parameters sets, the deviation represents the mean deviation within these clusters and the variance refers to the differences between the different clusters. Here, the true Iota Index, the number of categories and the concentration of the categories in the true distribution explain about 58% of the variance in the average deviation between true and estimated parameters. On this level, the true reliability is the strongest predictor. That is, the higher the true reliability of a coding scheme, the more accurate the parameter's estimates. The number of categories again has a similar influence while a higher concentration on a single category implies less accurate results. This effect, however, is quite small.

On the level of coding processes, deviation refers to the mean deviation within a coding process and the variance refers to differences in the coding processes belonging to a specific set of true parameters. Here, the number of categories and the sample size account for roughly 16% of the corresponding variance in the mean deviation between true and estimated parameters. Both variables have a similar effect on the deviation and a similar effect size. The more raters are involved in the data generation and the greater the sample size, the more accurate the results.

On the last level, deviation refers to the deviation between each pair of true and estimated parameter of the Assignment Error Matrix and categorical sizes. The variance refers to all parameters within a coding process. Since on this level, there are no variables to predict the deviation, R^2 is zero. In this kind of model, the deviation between a concrete pair of true and estimated parameters is estimated with the help of the corresponding cluster means (Geiser, 2013, p. 215), leading to the regression function shown in Figure 10. The total R^2 of .0243 indicates that about 2.5% of the variance's deviation between the true and estimated parameters can be explained by the fitted model and the derived regression function, which is quite low.

5.2.3 Accuracy of the Derived Reliability Measures

The Assignment Error Matrix and the categorical sizes are the foundation of the Alpha Elements, Beta Elements, Iota and the Iota Index. Thus, the accuracy of the estimated Assignment Error Matrix and the categorical sizes have an impact on the accuracy of the derived measures. Table 9 provides a first overview.

Table 9. Descriptive Statistics of Estimation Errors for Alpha, Beta, Iota and the Iota Index

	Deviation Alpha Reliability	Beta Reliability	Deviation Iota	Deviation Iota Index
$N_{Cluster\ Level\ 1}$	10,514	10,514	10,514	10,514
$N_{Cluster\ Level\ 2}$	415,291	415,291	415,291	415,291
$N_{Cluster\ Level\ 3}$	1,458,426	1,458,426	1,458,426	-/-
Minimum	.000000	.000000	.000000	.00000
25% Percentile	.019000	.008578	.021070	.01136
Median	.045450	.030686	.052030	.02621
Mean	.068370	.049921	.072630	.03853
75% Percentile	.091350	.067043	.103710	.05189
95% Percentile	.210172	.160928	.210326	.11669
Maximum	.767210	1.000000*	.914460	.77883
SD	.074152	.070058	.068962	.04027

Note. * The maximum deviation of 1 is a result of the estimation algorithm in extreme situations. If the Alpha Reliability of a category is perfect (the cell in the Assignment Error Matrix equals 1) the condition for the Beta Error equals zero and is not defined. In these cases, Beta Error is set to zero, implying a Beta Reliability of one. If the estimation results in values for the Alpha Reliability unequal 1, the Beta Error is estimated conventionally. Thus, these cases are outliners.

With a certainty of 75% the estimated values for Alpha Reliability do not differ more than 9 percentage point from their true values. The maximum deviation is about 21 percentage points with a certainty of 95%. Focusing on Beta Reliability,

the estimations are more accurate since the values do not differ more than 16 percentage points. For Iota, similar results occur as for Alpha Reliability. Switching to the scale level, the estimated values for the Iota Index do not differ more than .12 from their true values. In the following, regression analyses are presented to provide insight into the influencing variables of the reliability measures. Table 10 presents these results for Alpha Reliability.

Table 10. Influences on the Deviation Between True and Estimated Alpha Reliability

	Model 1		Model 2		Model 3		Model 4	
	R^2 .0019		R^2 .0025		R^2 .0203		R^2 .0376	
	b	β	b	β	b	β	b	β
Concentration	0.024	0.044	0.029	0.052	0.030	0.053	0.030	0.053
Number of Categories			0.002	0.024	0.002	0.024	0.002	0.024
Number of Raters					-0.009	-0.134	-0.009	-0.134
Sample Size							0.000	-0.132
Notes: All coefficients are significant at the 0.1% level.								

All included variables account for about 3.8% of the total variance, which is quite a low number. The most influencing variables are the number of raters and the sample size. The more raters judge coding units and the greater the sample size, the more accurate the estimates. The number of categories and the concentration of the true distribution practically do not influence the estimation results. Similar results occur for the Beta Reliability as Table 11 shows.

Table 11. Influences on the Deviation Between True and Estimated Beta Reliability

	Model 1		Model 2		Model 3		Model 4	
	R^2 .0008		R^2 .0018		R^2 .0076		R^2 .0157	
	b	β	b	β	b	β	b	β
Concentration	-0.015	-0.028	-0.009	-0.017	-0.009	-0.016	-0.009	-0.016
Number of Categories			0.002	0.034	0.002	0.033	0.002	0.033
Number of Raters					-0.005	-0.076	-0.005	-0.076
Sample Size							0.000	-0.090
Notes: All coefficients are significant at the 0.1% level.								

Concentrating on Iota, Table 12 implies that the considered variables explain only about 3.6% of the total variance. Again, the number of raters and the sample size have a positive impact on the accuracy. In contrast to Alpha and Beta, the number of categories shows no significant influence on the deviation between the estimates and true values of Iota. The concentration on a single category in the true distribution does practically not influence the estimation.

Table 12. Influences on the Deviation between True and Estimated Iota

	Model 1		Model 2		Model 3		Model 4	
	R^2	.0024	R^2	.0024	R^2	.0195	R^2	.0357
	b	β	b	β	b	β	b	β
Concentration	0.025	0.049	0.025	0.049	0.026	0.050	0.026	0.050
Number of Categories			0.00	0.00	0.000	0.001	0.000	0.000
Number of Raters					-0.008	-0.131	-0.008	-0.131
Sample Size							0.000	-0.127

Notes: All coefficients are significant at the 0.1% level except Number of Categories

Switching to the scale level, Table 13 reports the results for the Iota Index. Now, all variables explain about 11.5% of the total variance. The concentration on a single category in the true distribution of the categories explains less than 0.1% of the variance. Thus, it is practically not relevant. Similarly, the number of categories explains about 0.39% of the variance, which is very low. In contrast, the number of raters explains about 5.94% and the sample size about 5.14% of the total variance. The more raters are involved in the data generation and the greater the sample size, the more accurate the estimates.

Table 13. Influences on the Deviation Between True and Estimated Iota Index

	Model 1		Model 2		Model 3		Model 4	
	R^2	.0005	R^2	.0039	R^2	.0633	R^2	.1147
	b	β	b	β	b	β	b	β
Concentration	-0.006	-0.023	-0.011	-0.043	-0.011	-0.041	-0.010	-0.041
Number of Categories			-0.002	-0.061	-0.002	-0.062	-0.002	-0.062
Number of Raters					-0.009	-0.244	-0.009	-0.244
Sample Size							0.000	-0.227

Notes: All coefficients are significant at the 0.1% level.

Since R^2 for the Iota Index is quite high, a new function was fitted to the data to provide information for planning studies. Figure 11 presents the results for a different number of raters together with the 95% prediction interval.

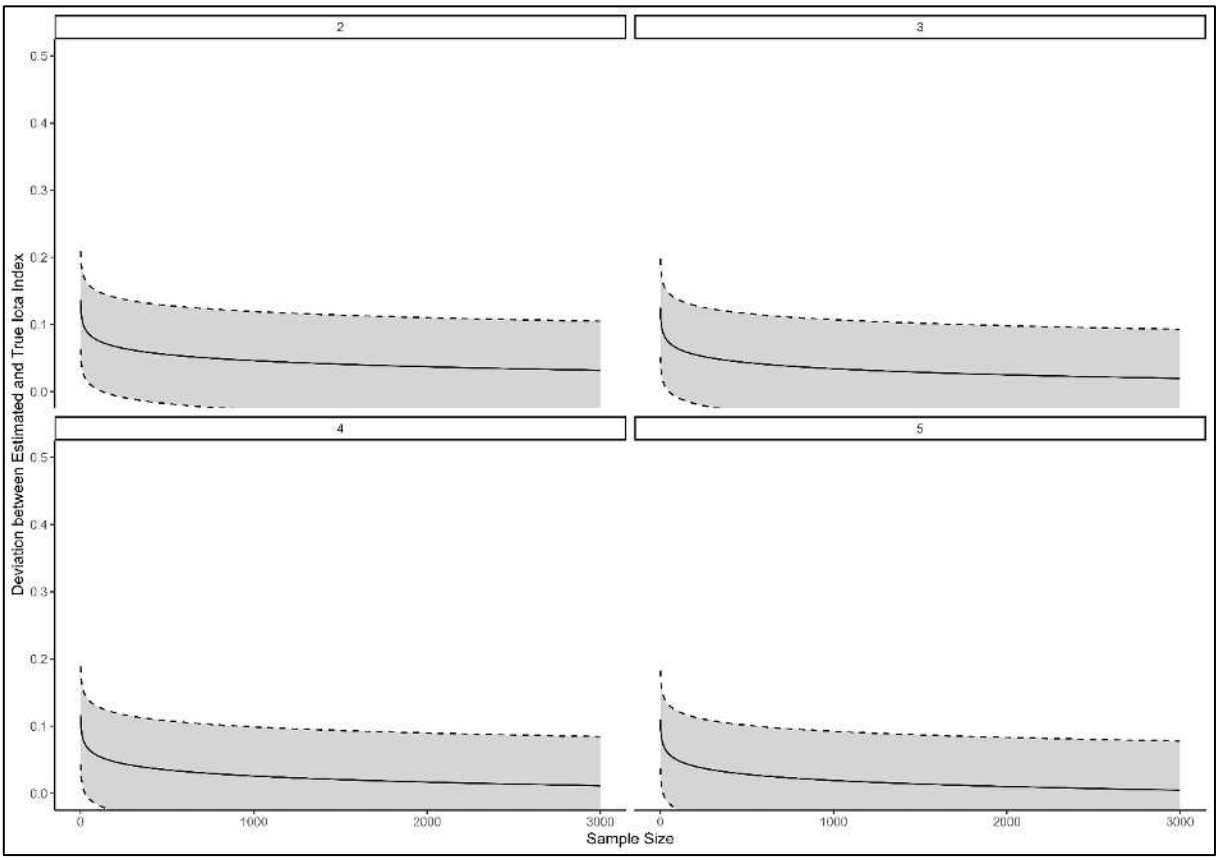


Figure 11. Relationship Between the Deviation of the Iota Index and the Number of Raters and Sample Size.

The function plotted in Figure 11 is given by Equation 19. An increase in the sample size leads to more accurate values, however the effect of the sample size is decreasing with increase n.

$$d_{Iota\ Index}(r,n) = -\ln(0.02965969)\,r - \ln(0.01309752)\,n + 0.15730986, R^2 = .1423 \quad [19]$$

Table 14 shows the necessary sample size to ensure that the Iota Index does not deviate more than 0.10 from its true value with a certainty of 95%.

Table 14. Necessary Sample Size to Ensure a Maximum Deviation of Iota Index l of a Maximum of 0.1

Number of Raters	Maximum Deviation of 0.1 with a Certainty of 95% for the Iota Index
2	4,388
3	1,752
4	913
5	551

In order to take the clustered structure of the data into account in more detail, a multilevel analysis is performed with *Mplus* 8.8 and Bayes Estimation. This analysis includes all measures simultaneously. Figure 12 presents the results.

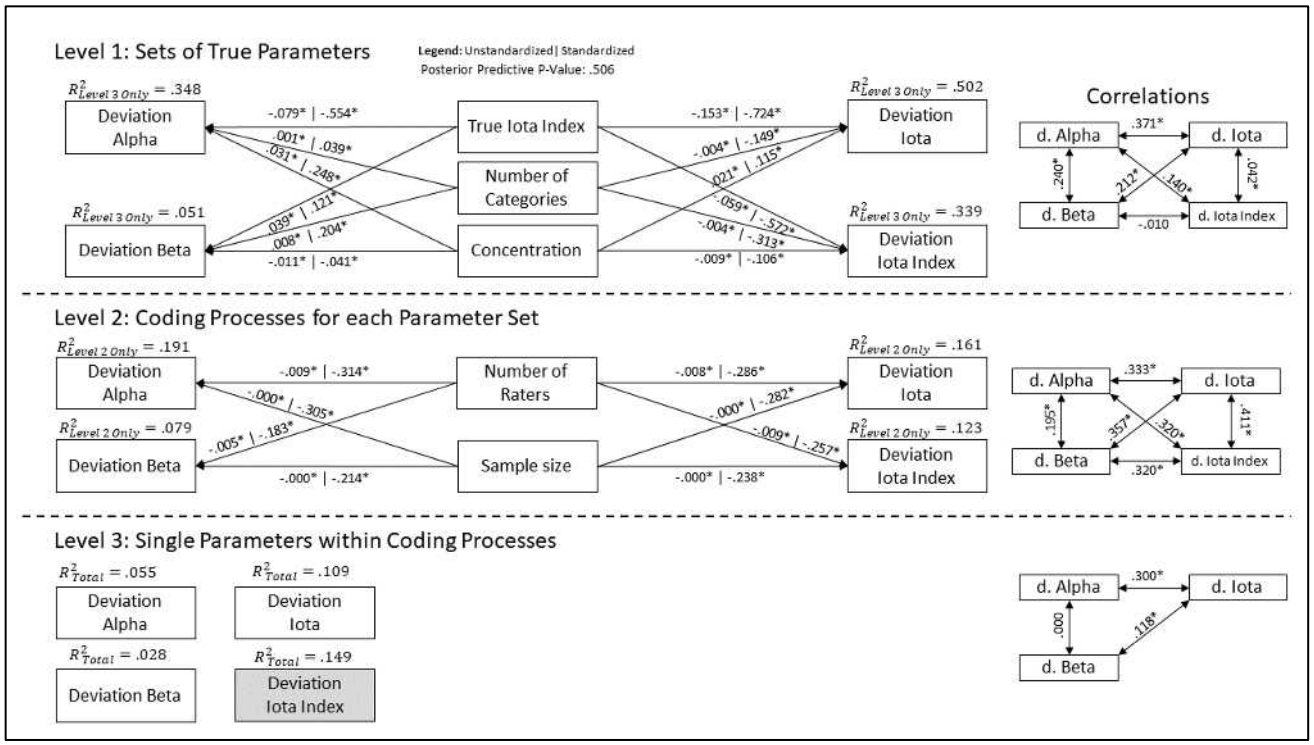


Figure 12. Summary of Multilevel Regression for the Deviation of Reliability Measures

On the first level, the deviation represents the mean deviation within the true set of parameters. The variance on this level can be interpreted as a variance in the mean deviation between the different sets of true parameters. Here, a higher true reliability implies more accurate estimates for all measures except for the Beta Reliability. A higher number of categories leads to more accurate results for Iota and the Iota Index but implies less accurate results for the Alpha and Beta Reliability. A higher concentration of single categories in the true distribution implies more accurate results for Beta and the Iota Index but not for Alpha and Iota. The mean deviation of the true parameter sets can best be explained for Iota with $R^2 = .502$. Differences in the mean deviation of Beta Reliability can be explained on this level only with $R^2 = .051$.

On the top level, the mean deviation between Alpha Reliability, Beta Reliability and Iota shows weak to moderate correlations (Cohen, 1988, pp. 79–80). That is, higher deviations on one measure are associated with higher deviations on the other measures. In contrast, all three measures are not or weakly correlated (Cohen, 1988, pp. 79–80) with the mean deviation of the Iota Index.

On the level of a single coding process, the deviation represents the mean deviation for the coding processes belonging to specific true parameters. The

variance on this level characterizes the variance between the corresponding coding processes. Here, a higher number of raters and a greater sample size lead to more accurate results for every measure. Both variables explain about 8% of the variance for the Beta Reliability, compared to 19% for the Alpha Reliability. On this level, all measures except Alpha and Beta reliability show a moderate to strong correlation (Cohen, 1988, pp. 79–80). That is, a greater deviation on one measure is associated with a greater deviation on the other measures. Only for the correlation between Alpha and Beta this correlation is weak to moderate (Cohen, 1988, pp. 79–80).

Focusing on the individual parameters within a coding process, the R^2 is zero in every case since there are no explanatory variables included. On this level, the deviation represents the mean deviation between the parameters belonging to a coding process. The variance can be interpreted as the variance of these mean values. On this level, higher deviations for Alpha and Beta lead to higher deviations for Iota, which is plausible since Iota can be described as a summary of Alpha and Beta Reliability. In contrast, the deviations of Alpha and Beta do not correlate on this level as they are independent.

Taking all levels together, the model is able to explain about 3% of the variance in the deviation of Beta Reliability, compared to 15% for the Iota Index. Equations 20 to 23 describe this relationship.

$$d_{Alpha} = 0{,}13686617 - 0{,}07898855 I_T + 0{,}03052350 H + 0{,}00067391 n_c - 0{,}00915591 n_r - 0{,}00002355 n_s \quad [20]$$

$$d_{Beta} = 0{,}03596601 + 0{,}03908057 I_T - 0{,}01145610 H + 0{,}00807707 n_c - 0{,}00485815 n_r - 0{,}00001507 n_s \quad [21]$$

$$d_{Iota} = 0{,}17880818 - 0{,}15278928 I_T + 0{,}02100430 H - 0{,}00386700 n_c - 0{,}00831589 n_r - 0{,}00002165 n_s \quad [22]$$

$$d_{Iota\ Index} = 0{,}12037919 - 0{,}05926049 I_T - 0{,}00949960 H - 0{,}00397315 n_c - 0{,}00885201 n_r - 0{,}00002169 n_s \quad [23]$$

Since the deviation depends substantially on the true level of reliability, the following Figure 13 shows the accuracy for different levels of certainty. Specific values can be found in **Appendix A**.

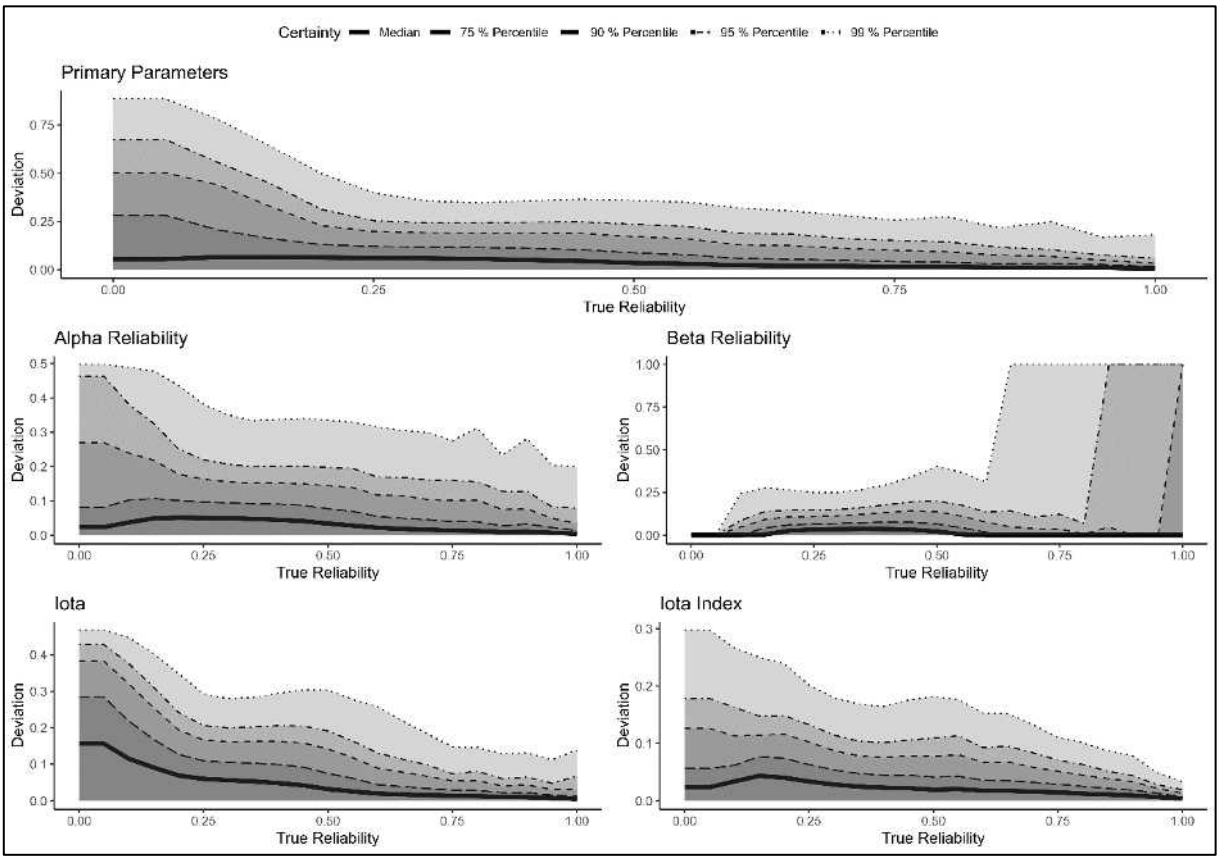


Figure 13. Certainty Levels of Accuracy for Different Levels of True Reliability

For all estimates except Iota, the deviation of true and estimated values does not exceed 10 percentage points with a certainty of 50% (Median). For Iota, this applies from a true reliability of about 0.15. Focusing on the 90% interval, the primary parameters do not exceed 10 percentage points from a true reliability of 0.75, Alpha from 0.8, Iota from 0.6 and the Iota Index from 0.3, respectively. Beta is below 0.1422 for all levels of true reliability with a certainty of 90%. From a level of true reliability of about 0.65, high deviations occur for Beta in 99% of intervals. These extreme values are due to estimation problems for rare, extreme values in which the corresponding true Alpha Reliability is perfect and the denominator of Beta not defined (see note in Table 9 for an explanation).

The reason for the primary parameters' broad confidence intervals in the range of low true reliability (0-0.25) can be traced back to the categorical sizes. In cases where the Assignment Error Matrix is close to perfect absence of reliability, raters assign categories to coding units randomly regardless of their true category. This implies that the data of assigned categories shows an equal distribution of categorical sizes. The true categorical sizes do not matter as a result of random assignments. Thus, high deviations between the estimated and

true categorical sizes have to be expected. The following section summarizes the results of the first simulation study.

5.3 Summary of Simulation Study I

Categorical Level

The first simulation study aims to provide insights into basic properties of the implemented estimation algorithm. Referring to the four hypotheses from section 5.1, two hypotheses can be confirmed. The greater the sample size (H1) and the more raters are involved in the data generation (H2), the more accurate the parameter estimates are.

The influence of the number of categories is more complicated. On the one hand, more categories are associated with a higher total estimation error. On the other hand, the single parameters are not negatively affected by a higher number of categories. Thus, hypothesis (H3) cannot be confirmed completely. Since the application of the Iota Concept relies on the individual parameters of the Assignment Error Matrix and the categorical sizes, it is more plausible to assume that the number of categories does not negatively affect the estimates. That is, a higher number of categories does not lead to less accurately estimated parameters.

According to hypothesis (H4), an influence of the individual categories' concentration on the true distribution could be identified. A higher concentration is associated with a higher deviation from the true parameters. This implies that the estimation algorithm performs best if the categories of a scale are equally distributed. However, the influence of the concentration is small. From a very strict point of view, hypothesis (H4) cannot be confirmed. From a more practical point of view, the influence of the concentration is less important due to low values for R^2.

These results continue for the reliability measures derived from the estimated parameters. The higher the sample size and the more raters are involved, the more accurate are the Alpha Reliability, Beta Reliability and Iota. A higher number of categories contributes to more accurate results for Iota and the Iota Index but decreases the accuracy for Alpha and Beta Reliability. In both cases, the negative impact of a higher number of categories is quite low because of low standardized regression coefficients (Alpha Reliability) or a small R^2 (Beta Reliability). The degree of concentration in the true distribution affects most

notably the Alpha Reliability with a higher concentration, leading to more inaccurate estimates.

Scale Level

Focusing on the scale level, (H5) and (H6) can be confirmed. The greater the sample size and the more raters are involved in rating coding units, the more accurate the estimate for the Iota Index. A higher number of categories is associated with a smaller deviation between true and estimated values, contradicting (H7). Finally, the estimates are more precise if the true distribution of the categories concentrates on single categories. From a strict point of view, hypothesis (H8) cannot be confirmed. The effect is small compared to the other influencing variables on the first level. Again, from a more practical point of view, the influence of the concentration is less important.

Absolute Values of Accuracy

Besides the factors themselves that influence accuracy, the absolute value of accuracy itself is equally important when judging the quality of an estimation algorithm. Referring to the central tendency (median) in Figure 13, the estimates do not deviate more than 5 percentage points for both the Alpha Reliability and Beta Reliability and 0.042 for the Iota Index. For Iota and the primary parameters, higher deviations have to be expected if the true reliability is quite low (lower than 0.25). Taking the uncertainty of the estimation into account, the 95% confidence interval is rather broad for all parameters except the Beta Reliability and the Iota index. Thus, in the central tendency, the estimation algorithm is quite accurate. However, further research should aim to reduce the uncertainty in parameter estimation. In practice, it matters if a coding unit is assigned to the right category in 10% or in 35% of cases. On the scale level, the situation is less problematic. With a certainty of 95%, the values for the Iota Index differ not more than .1 if the true reliability is at least .1. Thus, the values on the scale level are quite accurate and robust. After analyzing the basic properties of the new estimation algorithm, the following section investigates its predictive power.

6 Simulation Study II

6.1 Research Questions and Design of Simulation Study II

The second simulation study investigates how well the new concept is able to predict consequences arising from different degrees of reliability on subsequent analyses. The study addresses the following research questions:

RQ1 How strong does the Iota Index predict the deviation between true and estimated sample association/correlation for nominal/ordinal data to be?

RQ2 How strong does the Iota Index predict Type I and Type II Errors in accepting and rejecting hypothesis of association/correlation to be?

RQ3 How strong does the Iota Index predict the correct effect size of an association/correlation to be?

RQ4 How does the Iota Index perform in predicting consequences in comparison to the old Iota Concept and other measures of inter-rater reliability?

RQ5 What are meaningful cut-off values for reliability in practical situations?

To generate answers to these questions, a simulation study is performed. All simulations in this study were performed using the High-Performance Computing Cluster "Hummel" at the University of Hamburg. Figure 14 shows the design of the study.

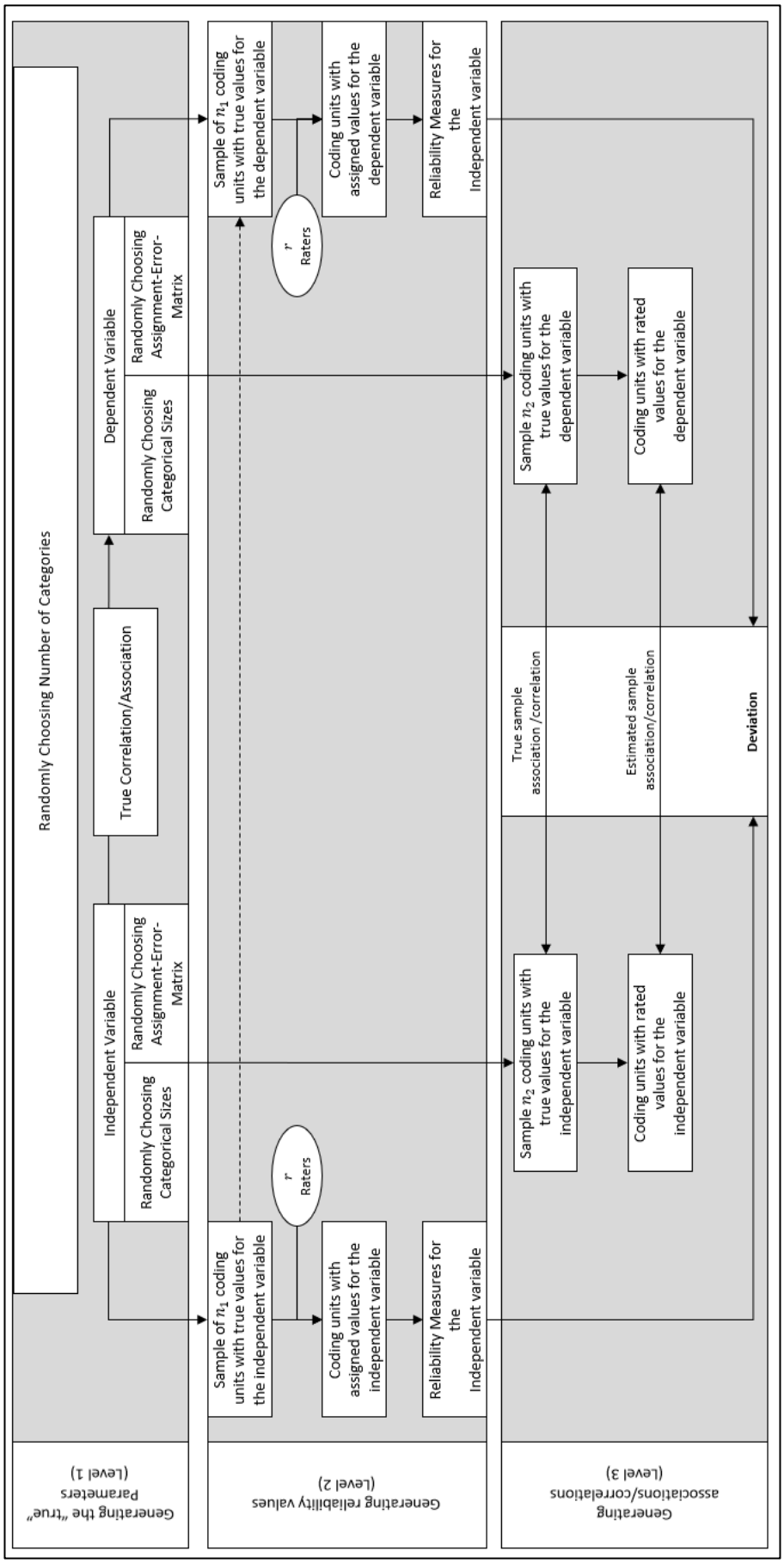


Figure 14. Design of Simulation Study II

Step 1: Generating Sets of True Parameters

In the first step the true values are generated for both the independent and the dependent variable. That is, for both variables an Assignment Error Matrix and categorical sizes are chosen randomly. This assumes that the number of categories is identical. The simulation ensures that the Assignment Error Matrix is in line with weak superiority. The number of categories can vary between two and five.

As the aim of this simulation study is to understand how different degrees of reliability are connected with the quality of subsequent analyses, the sample of true parameters for the Assignment Error Matrix has to ensure that the complete range of reliability is captured. For achieving this, a stratified random sampling is applied. Here, the range between $1/c$ and 1 is divided into 20 equidistant parts. The first part represents a reliability around zero and the 20th part represents a nearly perfect reliability. The part is chosen randomly in the moment where the simulation generates the true parameter set. Within the selected part, a concrete value is chosen randomly. The concrete value is used as the mean of normal distribution. The corresponding standard deviation is set to half the minimum distance of this value to 1 or $1/c$. With help of this normal distribution, c probabilities are generated which form the diagonal elements of the Assignment Error Matrix. The other probabilities are chosen randomly. In every case, the assumption of weak superiority is fulfilled.

Parts associated with a higher degree of reliability have a higher chance than parts associated with a lower reliability (the highest has a chance twice as high as the lowest). In order to allow the derivation of cut-off values, the analysis made must be as accurate as possible at the higher end of the reliability scale. The probabilities for the categorical sizes are drawn with simple random sampling.

Additionally, in this first step, the kind and strength of the association/correlation between both variables are modeled. In the case of nominal data, the function $f(x)$ assigns every category of the independent variable to a category of the dependent variable. This allocation is done randomly to ensure that a category of the independent variable corresponds with a category of the dependent variable at all times. Also, every category can only correspond with one other in both directions. For ordinal data, $f(x)$ models a simple linear relationship that

ensures that the minimal and maximal categories of both variables correspond with each other.

The strength of association/correlation is modeled based on the work of Cohen (1988) and concentrates on the range relevant for practice: “practically no association/correlation”, “practically weak association/correlation”, “practically medium association/correlation” and lastly “practically strong association/correlation”. Every class is assigned a range of probabilities ranging from 0.0 to lower 0.1 for no practical relationship, from 0.1 to lower 0.3 for a weak practical relationship, from 0.3 to 0.5 for a practical medium relationship and from 0.5 to 0.7 for a strong relationship. These probabilities are inspired by the work of Cohen (1988) who suggests that relationships expressed with a value of at least 0.1 should be interpreted as a weak but practical relevant relationship, with at least 0.3 as a medium and with at least 0.5 and above as a strong relationship in the social sciences. Although Cohen (1988, 79-81,224-225) refers to Pearson’s Correlation and w and not to probabilities, these values provide a guide for modeling realistic strength.

As the first step, one class of strength is chosen randomly. In the next step, a probability belonging to that class is selected by chance. For example, if the category medium was chosen, a probability of 0.4 means that in about 40% of cases, a category of the independent variable is assigned to the corresponding category of the dependent variable as determined by $f(x)$. In all other cases, the category of the independent variable is assigned randomly to any of the categories of the dependent variable.

The true effect size of the population is calculated additionally with a sample of 10,000 of true coding units. This effect size is used to assign the simulated process to one of the four classes of association’s/correlation’s strength since the probabilities do not exactly match the classification of Cohen (1988). Furthermore, the effect size for Cramer’s V depends on the number of categories, which is considered with this method.

Step 2: Simulating Reliability Estimation of Content Analysis

In the next step, the estimation of the reliability is modeled. This step assumes that several practitioners of content analysis draw a sample of coding units, rate the coding units and estimate the reliability of the codings. It further assumes that in practice, this sample is quite small (e.g., Früh, 2017, p. 180; Krippendorff,

2019, p. 394). The reason for the small sample size is cost-related, as several raters have to rate the coding units. Furthermore, the information obtained is often used to judge the quality of a coding scheme and, if the quality is not sufficient, to refine it. After each refinement, the raters have to draw a new sample of coding units and have to rate them in order to prove that the changes in the coding scheme indeed advanced the coding process. Thus, this phase poses high efforts for practitioners as it can be realized only with a limited sample size. For example, Früh (2017, p. 180) suggests about 30 to 50 codings for a scale as a minimum and 200 to 300 as a preferred size. Schreier (2012, p. 151) suggests about 10% to 20% of the final sample size for qualitative content analysis.

Such a cycle of improvement can be found at multiple points in content analysis literature (Früh, 2017, p. 185; Krippendorff, 2019, p. 394; Kuckartz, 2018, p. 95; Mayring, 2015, pp. 10–109; Schreier, 2012, pp. 152–165). It implies that reliability is estimated *before* the central study is performed and that reliability estimations are done with small sample sizes.

The simulation study implements these assumptions by splitting the reliability estimation and core study into separate processes. For the reliability estimation, the simulation generates a sample of coding units based on the true parameters' values of step 1 with sample size n_1. The size varies between 20 and 300 units. The true values of the independent variable of the coding units are generated based on the corresponding categorical sizes. The true values of the dependent variable are generated based on the corresponding categorical sizes and the true strength of the correlation/association. For the case of absence of a relationship, this implies that the distribution of the true categories for the dependent variable follow only the multinomial distribution characterized by the categorical sizes of the dependent variable. In the case of a perfect relationship, the distribution is determined by the categorical sizes of the independent variable. For the cases between a perfect and no relationship, the distribution of the dependent variable is a mixture of the multinomial distribution of the independent variable and the categorical sizes for the dependent variable, weighted by the strength of the relationship.

The generated sample is the foundation for the coding process. The number of raters r varies between two and five. Each rater judges the coding units of the sample according to the true Assignment Error Matrix for both variables.

The coding of every rater provides the basis for the computation of different reliability measures. The new Iota Index, Average Iota and Minimum Iota from the first generation, Krippendorff's Alpha and the Percentage Agreement are applied. Krippendorff's Alpha and Percentage Agreement provide comparison standards for the new measure. Percentage Agreement represents a more liberal measure and Krippendorff's Alpha a more conservative one (Zhao et al., 2013, p. 473). Computation of both measures are done with the package *irr* (Gamer et al., 2019). The Iota Concept of the first generation is used to investigate if the second generation indeed provides a progression.

Step 3: Simulating the Core Study of Content Analysis

In the literature on content analysis, the core study is performed *after* the reliability estimation (Früh, 2017, p. 185; Krippendorff, 2019, p. 394; Kuckartz, 2018, p. 95; Mayring, 2015, pp. 10–109; Schreier, 2012, pp. 152–165). This step assumes that not all coding units of the core study are rated by all raters because of the high sample size and the cost and time limitations in practice. Thus, the simulation assumes that the coding units are judged by only one rater.

Similar to step 2, a sample of coding units is generated with sample size n_2, varying between 100 and 3,000. For this sample the true association/correlation is calculated. In the next step, the simulated rater judges the coding units according to the true Assignment Error Matrix for both variables. At the end, all coding units are assigned to a category of both the independent and the dependent variable. On the basis of the coded data, the estimated sample association/correlation is calculated. This allows a comparison of the true and the estimated sample association/correlation.

Analysis

To measure the relationships, Cramer's V is applied for nominal data and Kendall's Tau for ordinal data. Kendall's Tau is part of the basic R program, while the calculation of Cramer's V makes use of the package *sjstats* (Lüdecke, 2018).

While the comparison of estimated and true sample association/correlation refers to RQ1, the simulation provides additional statistics for the other research questions. Concerning RQ2, the simulation compares the p-values associated with Cramers' V and Kendall's Tau and compares if the correct decision about the corresponding hypothesis is implied by the rated data. In this study, a Type I Error refers to the error that the null hypothesis is accepted based on the true sample

data while the estimated data implies a rejection. A Type II Error refers to the situation where the null hypothesis is rejected based on the true sample data while the estimated data implies an acceptance. The significance-level is set to 0.05 which is a broadly accepted convention in the social sciences (Agresti, 2022, p. 181; Hayes, 2005, p. 166).

Besides the level of significance, the effect size is another very important part of statistical analysis as even an effect size irrelevant for practice can be significant if the sample size is large enough (Rasch et al., 2010, p. 83). RQ3 addresses this issue by checking if the estimated data implies the same effect size as the data of the true sample. For ordinal data and Kendall's Tau, the simulation makes use of Cohen (1988, pp. 79–80), implying that values below 0.1 indicate a practically not relevant relationship, between 0.1 to lower 0.3 a small effect, between 0.3 to lower 0.5 a medium and about 0.5 a strong effect. For Cramer's V, the concrete cut-off values depend on the number of columns of the corresponding frequencies table. These can be found in Cohen (1988, p. 222). Thus, the cut-off values depend on of the number of categories in the coding scheme. The table Cohen (1988, p. 222) provides is used exactly.

The Iota Index is compared with other measures (RQ4) and the generated relationships between the degree of reliability on the one hand and the levels of deviation and errors on the other hand can be used to derive cut-off values for practice (RQ5).

The relevant relationships are developed with a multilevel regression analysis performed with *MPlus* 8.8 (see section 5.1 for the reason of this analytical approach), which considers the hierarchal data structure generated by the simulation. During the simulation, 25,399 sets of true parameters are generated. For every set of true parameters, 8 reliability estimations are simulated (two for every possible number of raters). For every reliability estimation, 30 core study processes are generated. The sets of true parameters form level 1. These values are the same for all corresponding coding processes in step 2 and 3. In addition, the reliability estimates are the same for all corresponding coding processes in step 3.

To consider the hierarchical structure of the data, a two-level model is fitted to the data where the reliability estimates form level 1 (between) and the simulated coding processes form level 2 (within). In contrast to the first simulation study,

the level of true parameters is not included in the model for two main reasons. First, the model should reflect practice and in practice the true values are not know. Second, including the level of true parameter sets would cluster the data according to their true reliability, leaving only a small amount of variance for the second and third level.

The analysis assumes that the independent variable is as important as the dependent variable. In statistics, this is modelled by constraining the regression coefficients of both variables to be equal. Figure 15 shows the underlying model. The following section presents the respective results.

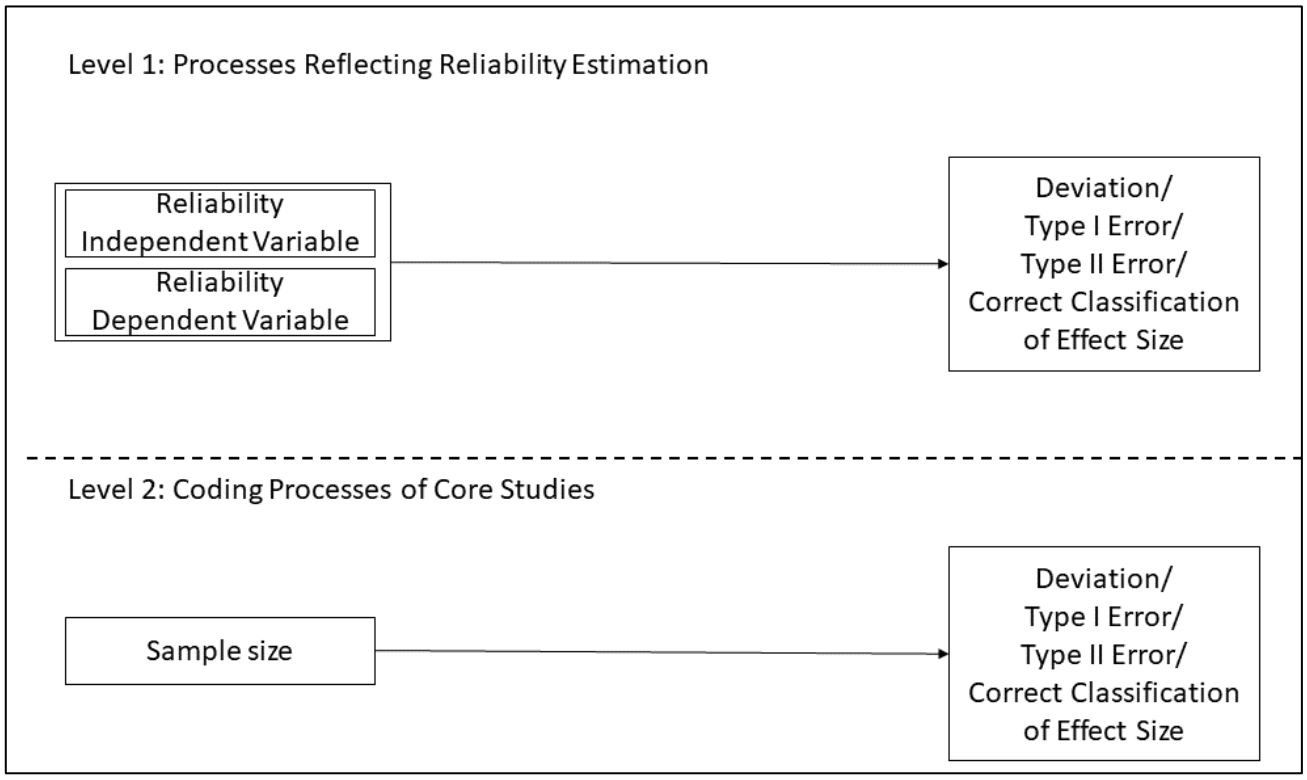


Figure 15. Multi-Level-Model of Simulation Study II

For the deviation between true and estimated sample association/correlation, a linear regression is applied. For Type I and II errors and for the correct classification of the effect size, probit regression is used since these variables represent a binary outcome.

6.2 Results of Simulation Study II

6.2.1 Overview

The simulation generated about 6,044,572 coding processes. These coding processes are nested within 201,486 reliability estimations before coding. Both

processes are based on 25,399 different sets of true parameters. Table 15 shows the distribution of cases to different levels of data (nominal and ordinal) and different strengths of true associations/correlations between the independent and dependent variable.

Table 15. Sample Size of Simulation Study II

	Nominal Data				
	No Relationship	Weak Relationship	Medium Relationship	Strong Relationship	Perfect Relationship
Set of True Parameters	1,703	2,297	1,840	4,382	2,555
$N_{Between}$ (Reliability Estimation)	13,468	18,157	14,511	34,855	20307
N_{Within} (Coding Processes)	404,039	544,710	435,330	1,045,649	609,204
	Ordinal Data				
	No Relationship	Weak Relationship	Medium Relationship	Strong Relationship	Perfect Relationship
Set of True Parameters	2,771	2,902	2,543	1,823	2,583
$N_{Between}$ (Reliability Estimation)	22,041	23,011	20,168	14,532	20,436
N_{Within} (Coding Processes)	661,230	690,330	605,040	435,960	613,080

An initial inspection of the relationship between the different degrees of reliability and the deviation from the true sample association for the case of nominal data and a strong relationship is provided in Figure 16. Figures for ordinal data and other configurations of the effect size can be found in **Appendix B**.

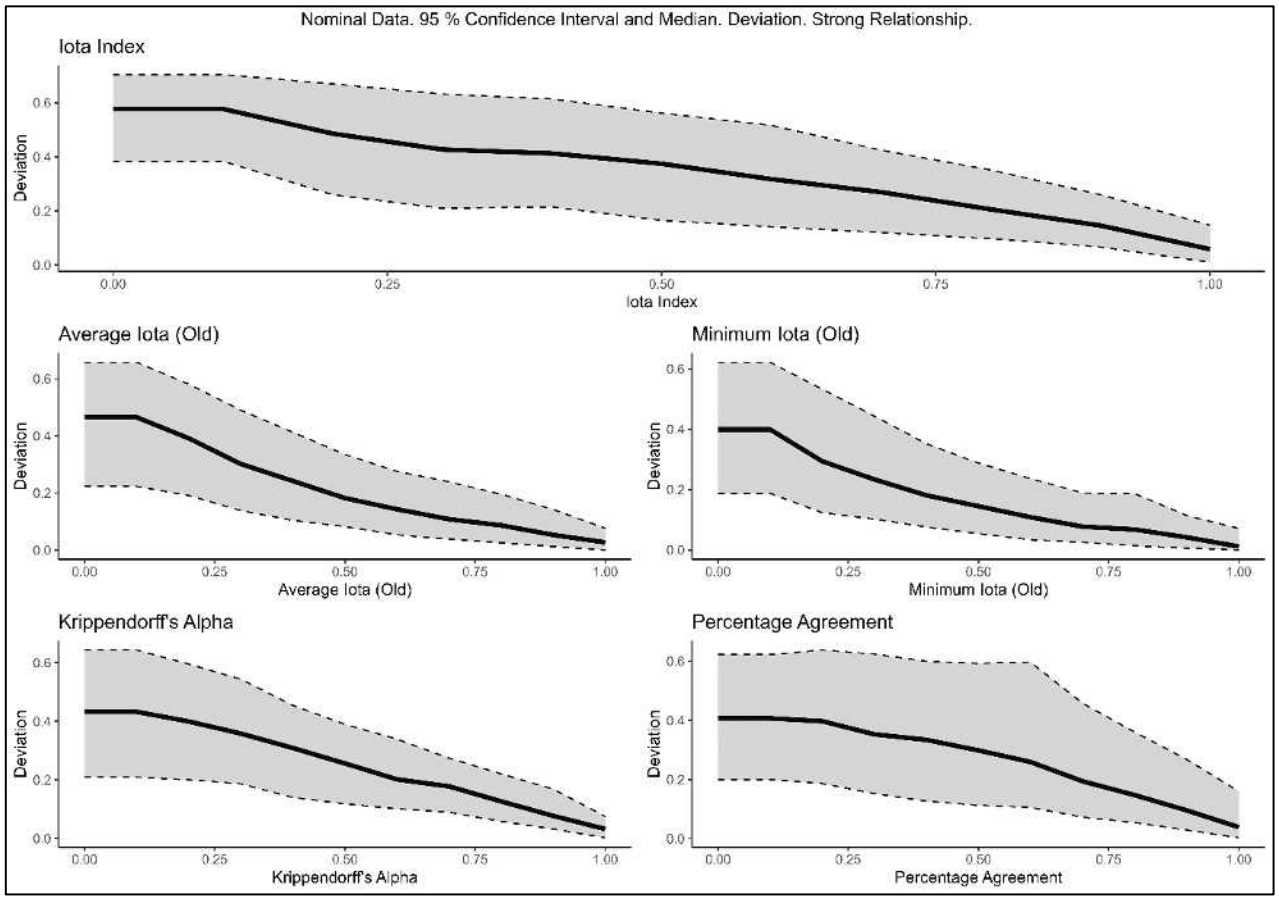


Figure 16. Relationship Between Estimated Reliability and Deviation from the True Association.

For every measure, a higher degree of reliability implies less deviation between estimated and true association. The relationship can be characterized as a linear relationship for all measures. This characterization holds even for other strengths of association and for all configurations of ordinal data (see **Appendix B**). Only for a perfect association and nominal data neither the Iota Index nor the Percentage Agreement follow a linear, but rather a quadratic relationship. A similar result occurs for a strong and a perfect ordinal relationship (see **Appendix B**). For all measures the uncertainty decreases with an increased strength of association/correlations as the width of the 95% interval decreases. Since the linear relationship occurs for most configurations, the following analyses apply linear models. **Appendix C** provides an overview for the global model fit.

According to the simulation study of Hu and Bentler (1999, pp. 27–28), an RMSEA below .06 *and* an SRMR below .09 indicate a global model fit. Additionally, a CFI of at least .950 points to a global model fit. As the tables in **Appendix C** indicate, all estimated models are in line with these rules. The only exceptions are the models for the absence of association in nominal data for the Iota Index, Average

Iota and Krippendorff's Alpha, as well as the absence of a correlation in ordinal data for Minimum Iota in terms of a correct classification of the effect size. Here, the CFI is below .950. Since RMSEA and SRMR are in line with the rules of thumb developed by Hu and Bentler (1999), they remain part of the following analysis.

Furthermore, in terms of the correct classification of the effect size, most models could not be estimated under the condition of a perfect association/correlation in both nominal and ordinal data. For nominal data, the corresponding model could not be estimated for Minimum Iota and Percentage Agreement under the condition of a strong association. Thus, these models cannot be considered in the following analysis. Table 16 reports the values for R^2 on the first level showing how well the different reliability measures are able to predict the average deviation and the average error rates.

Table 16. $R^2_{Levl\ 1\ only}$ for the Prediction of the Average Deviation and Error Rates

Nominal Data					
Deviation					
	No Relationship	Weak Relationship	Medium Relationship	Strong Relationship	Perfect Relationship
Iota Index	.101	.200	.420	.541	.775
Average Iota	.081	.148	.323	.451	.693
Minimum Iota	.030	.061	.170	.332	.583
Krippendorff's Alpha	.110	.199	.382	.527	.821
Percentage Agreement	.006	.019	.085	.299	.488
Type I Error					
	No Relationship	Weak Relationship	Medium Relationship	Strong Relationship	Perfect Relationship
Iota Index	.044	.706	.746	.725	.669
Average Iota	.045	.599	.586	.556	.509
Minimum Iota	.029	.541	.522	.423	.578
Krippendorff's Alpha	.052	.681	.694	.654	.632
Percentage Agreement	.013	.456	.450	.355	.443
Type II Error					
	No Relationship	Weak Relationship	Medium Relationship	Strong Relationship	Perfect Relationship
Iota Index	.008	.053			
Average Iota	.001	.035			
Minimum Iota	.004	.003			
Krippendorff's Alpha	.007	.055			
Percentage Agreement	.008	0			
Effect Size					
	No Relationship	Weak Relationship	Medium Relationship	Strong Relationship	Perfect Relationship
Iota Index	.002	.484	.538	.757	
Average Iota	.012	.408	.495	.658	
Minimum Iota	.048	.264	.397		
Krippendorff's Alpha	.006	.492	.553	.766	
Percentage Agreement	.058	.174	.320		

Table 16. $R^2_{Levl\ 1\ only}$ for the Prediction of the Average Deviation and Error Rates (Continued)

Ordinal Data					
Deviation					
	No Relationship	Weak Relationship	Medium Relationship	Strong Relationship	Perfect Relationship
Iota Index	.288	.385	.605	.680	.708
Average Iota	.244	.345	.556	.607	.642
Minimum Iota	.191	.297	.502	.541	.579
Krippendorff's Alpha	.293	.412	.658	.723	.779
Percentage Agreement	.169	.261	.446	.482	.512
Type I Error					
	No Relationship	Weak Relationship	Medium Relationship	Strong Relationship	Perfect Relationship
Iota Index	.094	.644	.681	.673	.637
Average Iota	.093	.522	.563	.535	.502
Minimum Iota	.082	.429	.480	.458	.415
Krippendorff's Alpha	.100	.665	.678	.666	.627
Percentage Agreement	.074	.391	.413	.375	.326
Type II Error					
	No Relationship	Weak Relationship	Medium Relationship	Strong Relationship	Perfect Relationship
Iota Index	.005	.037			
Average Iota	.001	.041			
Minimum Iota	.000	.047			
Krippendorff's Alpha	.001	.040			
Percentage Agreement	.002	.046			
Effect Size					
	No Relationship	Weak Relationship	Medium Relationship	Strong Relationship	Perfect Relationship
Iota Index	.006	.480	.636	.831	.523
Average Iota	.004	.436	.548	.684	
Minimum Iota	.001	.386	.487	.593	
Krippendorff's Alpha	.005	.538	.676	.864	.943
Percentage Agreement	.000	.353	.492	.701	

Focusing on the deviation between true and estimated association for nominal data, the values for R^2 increase for stronger true relationships. This is plausible since the relevance of reliability increases the more a real structure exists behind the relationship of two variables. Independently from the concrete level of the true relationship, percentage agreement performs the worst. Average Iota and Minimum Iota perform significantly better. For example, R^2 is about 45.1% for Average Iota in contrast to 29.9% for Percentage Agreement for a strong relationship. Krippendorff's Alpha and the new Iota Index perform best. Both

perform similarly well with the Iota Index performing slightly better in the practically relevant range.

Concentrating on the prediction of Type I Errors, the ranking of the reliability measure is similar to the ranking for deviation. The Iota Index and Krippendorff's Alpha perform best, followed by Average Iota, Minimum Iota and lastly Percentage Agreement. Here, the Iota Index has generally higher values for R^2 than Krippendorff's Alpha. The values for R^2 are stable across different strengths of true associations. Only in the case where the true association is practically absent, the explained variance is meaningfully lower for all measures.

A similar result occurs for the proper characterization of the effect size. Percentage Agreement performs the worst in predicting error in the characterization of the association's effect size. The other measures perform better. The Iota Index and Krippendorff's Alpha perform best.

For Type II Errors, all measures perform badly. The reason could be in the low frequency of Type II Errors in the simulated data and the comparatively high sample sizes for the coding processes on level 2. A high sample size increases the chance that a significance test becomes significant even for very low effect sizes.

For ordinal data, the results are similar. The Iota Index and Krippendorff's Alpha perform best while Krippendorff's Alpha shows slightly higher values for R^2 in terms of deviation. Average Iota and Minimum Iota are in the middle and Percentage Agreement performs the worst. The following section analyzes the relationships in detail, starting with the deviation between true and estimated sample associations/correlations.

6.2.2 Analyses of the Deviation Between True and Estimated Sample Association/Correlation

Table 17 provides detailed insights into the estimated models by showing the standardized regression coefficients and the corresponding values for R^2.

Table 17. Standardized Regressions Coefficient for the Deviation Between True and Estimated Sample Association/Correlation.

		Nominal Data				Ordinal Data			
Measure		Measure	$R^2_{Level\ 1}$	Sample Size	$R^2_{Level\ 2}$	Measure	$R^2_{Level\ 1}$	Sample Size	$R^2_{Level\ 2}$
Index	0	-.218*	.101	-.190*	.036	-.372*	.288	-.245*	.06
	1	-.316*	.200	.158*	.025	-.441*	.385	-.018*	0
	2	-.460*	.420	.187*	.035	-.532*	.605	.001	0
	3	-.514*	.541	.206*	.043	-.583*	.680	.001	0
	4	-.614*	.775	.145*	.021	-.587*	.708	.001	0
Average	0	-.194*	.081	-.190*	.036	-.338*	.244	-.245*	.06
	1	-.267*	.148	.158*	.025	-.414*	.345	-.018*	0
	2	-.395*	.323	.187*	.035	-.506*	.556	.001	0
	3	-.462*	.451	.206*	.043	-.543*	.607	.001	0
	4	-.576*	.693	.145*	.021	-.552*	.642	.001	0
Minimum	0	-.117*	.030	-.190*	.036	-.296*	.191	-.245*	.06
	1	-.168*	.061	.158*	.025	-.381*	.297	-.018*	0
	2	-.284*	.170	.187*	.035	-.480*	.502	.001	0
	3	-.394*	.332	.206*	.043	-.514*	.541	.001	0
	4	-.526*	.583	.145*	.021	-.519*	.579	.001	0
Alpha	0	-.228*	.110	-.190*	.036	-.375*	.293	-.245*	.06
	1	-.313*	.199	.158*	.025	-.462*	.412	-.018*	0
	2	-.438*	.382	.187*	.035	-.559*	.658	.001	0
	3	-.508*	.527	.206*	.043	-.603*	.723	.001	0
	4	-.627*	.821	.145*	.021	-.608*	.779	.001	0
Percent	0	-.052*	.006	-.190*	.036	-.265*	.169	-.245*	.06
	1	-.091*	.019	.158*	.025	-.340*	.261	-.018*	0
	2	-.192*	.085	.187*	.035	-.431*	.446	.001	0
	3	-.357*	.299	.206*	.043	-.458*	.482	.001	0
	4	-.455*	.488	.145*	.021	-.463*	.512	.001	0

Note.
* significant at the 5%-Level.
Coefficient of reliability measures for independent and dependent variable constrained to be equal.
0 = no true relationship, 1 = weak relationship, 2 = medium relationship, 3 = strong relationship,
4 = perfect relationship.

Table 17 shows that the relevance of the reliability measures increases with an increasing strength of true association/correlation. In contrast, the relevance of the sample size decreases when the strength of the true relationship increases. This is shown by the decreasing values for R^2 on the level of the core studies.

In the case of ordinal data, a higher sample size in a core study is associated with less deviation. In contrast, a higher sample size leads to a stronger deviation between the estimated and true sample association for nominal data, except for the cases where there is practically no true relationship. This result is surprising since a higher sample size should normally decrease the deviation. A potential

reason for the strange behavior of the sample size for nominal data could be that low levels of reliability imply a false structure of measured data that becomes more prominent with a higher sample size.

Figure 17 shows which deviation between the true and estimated sample association for nominal data has to be expected for different degrees of reliability and measures. Due to limited space, Figure 17 stands as a representative for all investigated configurations of ordinal data.

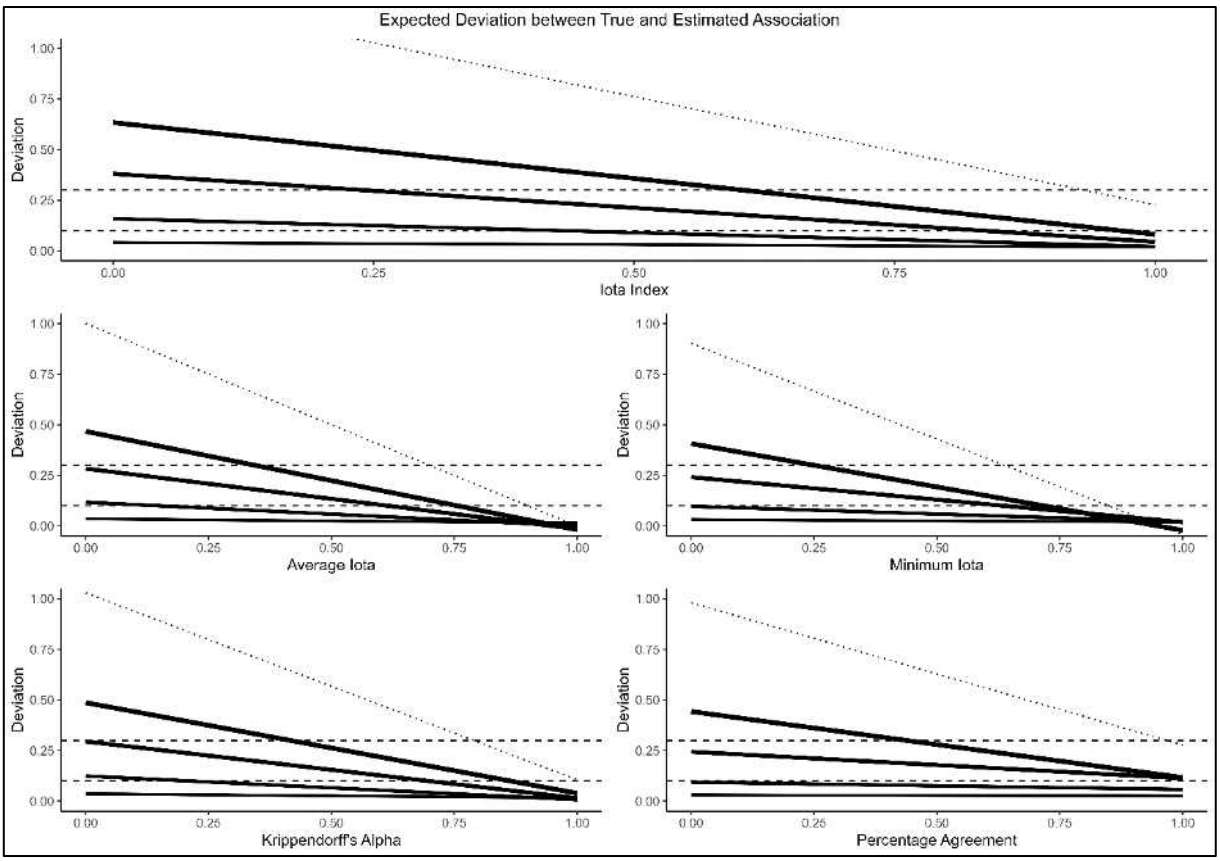


Figure 17. Expected Deviation Between True and Estimated Sample Association (Nominal Data).

The dotted lines in Figure 17 represent the relationships for the case of a perfect true relationship, which is not realistic for practice. The bold solid lines represent the more realistic cases ranging from practically no true association up to a practically strong true association. The density of the lines indicates the strength of association. The thicker the line, the stronger is the true association. The dashed lines on the horizontal emphasize a deviation of 0.1 and 0.3 which can be interpreted as the borders for a practically irrelevant deviation (0.1) and only a small practical deviation (0.3). These cut-off values are inspired by the taxonomy

of Cohen (1988, pp. 79–80) and are intended to classify the practical relevance of product-moment-correlations.

In Figure 17, the lines for the weaker associations are below the lines for the stronger associations. This implies that cases with a stronger true association have higher demands on the reliability than situations with a lesser true association. The intersections between a) the horizontal lines and b) the lines representing the expected deviation can be used to derive cut-off values. They characterize the necessary level of reliability to meet the assumption that the estimated association does not deviate more than 0.1 or 0.3 from the true value, respectively. Table 18 reports the specific intersections.

Table 18. Potential Cut-Off Values for Deviation

		Nominal				Ordinal			
		Expected Deviation		95% Interval		Expected Deviation		95% Interval	
		less 0.1	less 0.3	less 0.1	less 0.3	less 0.1	less 0.3	less 0.1	less 0.3
Iota Index	0	0	0	0	0	0	0	.645	0
	1	.427	0	1	0	.791	0	1	.347
	2	.834	.240	1	.646	.996	.553	1	.841
	3	.966	.604	1	.890	1	.762	1	.992
	4	1	.932	1	1	1	.976	1	1
Average Iota	0	0	0	0	0	0	0	.403	0
	1	.146	0	.997	0	.546	0	1	.087
	2	.614	0	1	.433	.766	.298	1	.616
	3	.755	.344	1	.696	.847	.521	1	.791
	4	.899	.700	1	.913	.931	.740	1	.974
Minimum Iota	0	0	0	0	0	0	0	.335	0
	1	0	0	1	0	.471	0	1	.005
	2	.634	0	1	.453	.694	.215	1	.557
	3	.717	.250	1	.687	.785	.450	1	.747
	4	.849	.637	1	.899	.859	.664	1	.923
Krippendorff's Alpha	0	0	0	0	0	0	0	.461	0
	1	.206	0	1	0	.627	0	1	.104
	2	.695	0	1	.482	.871	.356	1	.673
	3	.863	.416	1	.774	.963	.606	1	.861
	4	1	.789	1	.969	1	.842	1	1
Percentage Agreement	0	0	0	0	0	0	0	.564	0
	1	0	0	1	0	.740	0	1	.135
	2	1	0	1	.844	1	.399	1	.857
	3	1	.437	1	1	1	.697	1	1
	4	1	.967	1	1	1	.995	1	1

0 = no true relationship, 1 = weak relationship, 2 = medium relationship, 3 = strong relationship, 4 = perfect relationship.

Table 18 indicates that regarding the Iota Index, a value of .604 for both variables (independent and dependent) is associated with a deviation between true and estimated sample association (nominal data) of less than 0.3 for cases with a strong true association. For Krippendorff's Alpha, the corresponding value for both variables is about .416,.344 for Average Iota,.250 for Minimum Iota and .437 for Percentage Agreement.

Table 18 also illustrates the values for ordinal data. For example, if the expected deviation is not supposed to exceed a small value relevant for practice, the Iota Index must be at least .762, Average Iota .521, Minimum Iota .450, Krippendorff's

Alpha .606 and Percentage Agreement .697 for situations with a strong correlation.

The described cut-off values are the values for the expected deviation. They represent the best estimate for the deviation. However, they do not account for uncertainty. The prediction's precision can be characterized by the *standard error of the estimate*. This error also allows the calculation of *prediction intervals,* which characterizes the probability that the true value is within a specific range around the prediction (Afifi et al., 2020, p. 119). Table 19 reports the standard error of the estimate for the different measures. The standard errors show that the Iota Index and Krippendorff's Alpha perform similarly, with Percentage Agreement showing the highest error. Something similar is true for ordinal data.

Table 19. Standard Error of Estimate for the Deviation

Nominal Data					
	No Relationship	Weak Relationship	Medium Relationship	Strong Relationship	Perfect Relationship
Iota Index	.02302185	.05912678	.08318308	.09618919	.11236856
Average Iota	.02310696	.06059429	.08881300	.10417783	.12999450
Minimum Iota	.02332581	.06299857	.09699768	.11392911	.15039592
Krippendorff's Alpha	.02298175	.05917576	.08542422	.09747899	.10128250
Percentage Agreement	.02342492	.06409401	.10124673	.11654122	.16610964
Ordinal Data					
	No Relationship	Weak Relationship	Medium Relationship	Strong Relationship	Perfect Relationship
Iota Index	.03240891	.0623095	.07901177	.09202779	.13563398
Average Iota	.03266069	.06364934	.08275053	.10054227	.14939469
Minimum Iota	.03295476	.06519173	.08672944	.10765697	.16150227
Krippendorff's Alpha	.03238294	.06142252	.07479100	.08671020	.11896303
Percentage Agreement	.03307854	.06633126	.09058814	.11363413	.17349891

With the help of the standard error of estimates, the 95% interval is estimated and the intersection with the horizontal lines is calculated. These values ensure that with a certainty of 95% that the deviation between true and estimated association/correlation is less than 0.1 and 0.3, respectively.

As can be seen in Table 18, all measures need to be on their highest values to ensure a deviation of less than 0.1 in nearly all situations. This points to a weakness of all investigated reliability measures and points to the need of better estimation methods for the reliability to enable more precise predictions.

In addition, cut-off values can be derived for situations in which the deviation should not exceed a small practical effect. For example, assuming a strong true association in practice, a value of .890 for the Iota Index, of .696 for Average Iota, of .687 for Minimum Iota and of.774 for Krippendorff's Alpha ensures with a certainty of 95% that the deviation between the estimated and true association does not exceed more than 0.3 for nominal data. The next section characterizes the measures' predictive power for Type I Errors.

6.2.3 Analyses of Type I Errors

Table 20 reports the standardized regression coefficients along with the values for R^2. It shows that in the core study, a higher sample size implies a reduction of Type I Errors, except for a weak true association/correlation. The sample size explains between 0% and 35.3% of variations between the core studies.

Table 20. Standardized Regression Coefficient for Type I Errors.

		Nominal Data				Ordinal Data			
Measure		Measure	$R^2_{Level\ 1}$	Sample Size	$R^2_{Level\ 2}$	Measure	$R^2_{Level\ 1}$	Sample Size	$R^2_{Level\ 2}$
Index	0	-.144*	.044	.348*	.121	-.213*	.094	.309*	.096
	1	-.594*	.706	-.035*	.001	-.57*	.644	-.296*	.087
	2	-.613*	.746	-.528*	.279	-.565*	.681	-.485*	.236
	3	-.595*	.725	-.594*	.353	-.579*	.673	-.498*	.248
	4	-.571*	.669	-.564*	.318	-.556*	.637	-.502*	.252
Average	0	-.146*	.045	.348*	.121	-.208*	.093	.309*	.096
	1	-.536*	.599	-.036*	.001	-.509*	.522	-.297*	.088
	2	-.533*	.586	-.528*	.279	-.510*	.563	-.486*	.236
	3	-.513*	.556	-.594*	.353	-.510*	.535	-.498*	.248
	4	-.493*	.509	-.564*	.318	-.488*	.502	-.502*	.252
Minimum	0	-.116*	.029	.348*	.121	-.194*	.082	.309*	.096
	1	-.502*	.541	-.037*	.001	-.459*	.429	-.297*	.088
	2	-.497*	.522	-.528*	.279	-.469*	.480	-.486*	.236
	3	-.445*	.423	-.594*	.353	-.473*	.458	-.498*	.248
	4	-.523*	.578	-.529*	.280	-.440*	.415	-.502*	.252
Alpha	0	-.157*	.052	.348*	.121	-.219*	.1	.309*	.096
	1	-.581*	.681	-.036*	.001	-.587*	.665	-.296*	.088
	2	-.591*	.694	-.528*	.279	-.568*	.678	-.486*	.236
	3	-.566*	.654	-.594*	.353	-.579*	.666	-.498*	.248
	4	-.551*	.632	-.564*	.318	-.545*	.627	-.502*	.252
Percent	0	-.073*	.013	.348*	.121	-.176*	.074	.309*	.096
	1	-.442*	.456	-.036*	.001	-.416*	.391	-.297*	.088
	2	-.442*	.45	-.528*	.279	-.415*	.413	-.486*	.236
	3	-.389*	.355	-.594*	.353	-.403*	.375	-.498*	.248
	4	-.434*	.443	-.533*	.284	-.370*	.326	-.502*	.252

Note.
* significant at the 5%-Level.
Coefficient of reliability measures for independent and dependent variable constrained to be equal.
0 = no true relationship, 1 = weak relationship, 2 = medium relationship, 3 = strong relationship,
4 = perfect relationship.

Figure 18 shows the expected probability for a Type I Error to occur for different configurations of true association in nominal data. Likewise, as in Figure 17, the horizontal dashed lines represent the probability of 5% and 10% for a Type I Error to occur, which can be used to derive cut-off values for the necessary reliability. The figure illustrates that with increasing reliability, the chance for a Type I Error decreases. That is, the higher the reliability, the smaller the risk that the estimated association implies the rejection of an association although the true sample association would imply acceptance. Figure 18 also illustrates that the curves for stronger true associations are below the curves for the weaker true associations. Thus, situations with a smaller true association put higher demands

on the reliability in terms of Type I Errors than situations with a stronger true association.

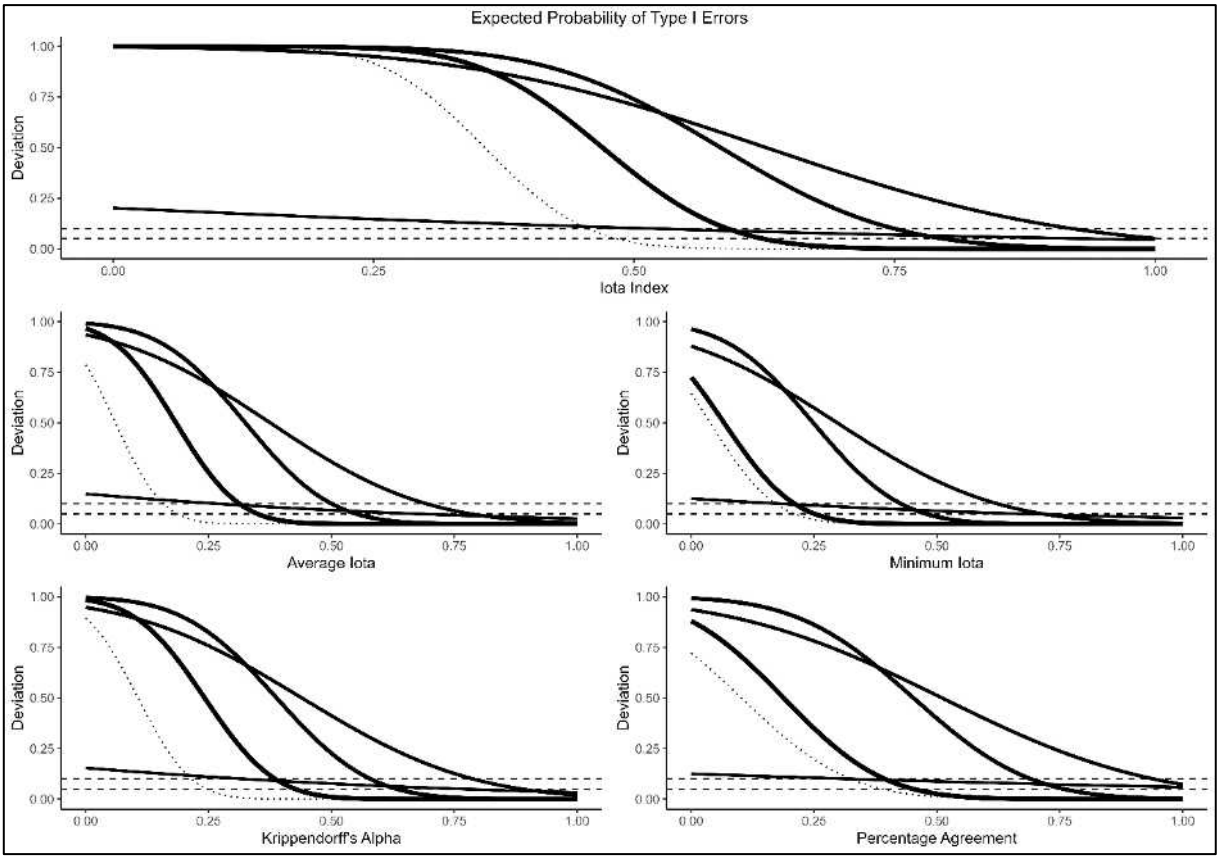


Figure 18. Expected Risk of Type I Error (Nominal Data).

Table 21 reports the intersections of the curves with the horizontal dashed lines. If the risk for a Type I Error to occur needs to be less than 5%, values of .797for the Iota Index for a medium weak true association, of .624 for a strong true association and of .483 for a perfect true association are necessary. Thus, the stronger the true association, the lower the demands for the reliability to lead researchers to make the “right” decisions.

If the risk for a Type I Error is supposed to be less than 10%, a value of .797 for the Iota Index, of .500 for Average Iota, of .427 for Minimum Iota, of .579 for Krippendorff’s Alpha and of .691 for Percentage Agreement is necessary for a medium true association.

Table 21. Potential Cut-Off Values for Type I Errors

		Nominal				Ordinal			
		Expected Deviation		95% Interval		Expected Deviation		95% Interval	
		less 5%	less 10%	less 5%	less 10%	less 5%	less 10%	less 5%	less 10%
Iota Index	0	.937	.517	1	1	1	.792	1	1
	1	1	.920	1	1	.919	.858	1	.971
	2	.797	.749	.874	.827	.732	.691	.800	.759
	3	.624	.590	.670	.636	.642	.607	.697	.663
	4	.483	.455	.517	.489	.563	.532	.610	.580
Average Iota	0	.674	.268	1	1	.802	.540	1	1
	1	.783	.693	.965	.875	.687	.621	.817	.751
	2	.550	.500	.640	.590	.474	.433	.551	.509
	3	.349	.313	.405	.369	.375	.338	.438	.401
	4	.186	.158	.225	.197	.280	.248	.334	.302
Minimum Iota	0	.672	.182	1	1	.731	.461	1	1
	1	.715	.623	.904	.812	.624	.554	.767	.697
	2	.477	.427	.572	.521	.397	.353	.480	.437
	3	.251	.210	.318	.277	.293	.255	.362	.324
	4	.210	.172	.263	.225	.181	.148	.241	.207
Krippendorff's Alpha	0	.764	.320	1	1	.938	.634	1	1
	1	.886	.788	1	.982	.774	.703	.906	.835
	2	.633	.579	.726	.672	.555	.508	.637	.589
	3	.426	.385	.486	.445	.450	.411	.515	.476
	4	.246	.215	.285	.254	.342	.307	.397	.362
Percentage Agreement	0	1	.321	1	1	1	.720	1	1
	1	1	.940	1	1	.943	.850	1	1
	2	.758	.691	.889	.821	.649	.589	.767	.707
	3	.452	.394	.550	.491	.510	.456	.611	.557
	4	.387	.324	.479	.416	.365	.317	.451	.403

0 = no true relationship, 1 = weak relationship, 2 = medium relationship, 3 = strong relationship, 4 = perfect relationship.

In order to take the uncertainty of estimations into account, the standard error of estimates (Table 22) is used to calculate prediction intervals that ensure with a probability of 95% that the occurrence of Type I Errors is less than 5% or 10%, respectively. According to Table 21, a value of .827 for the Iota Index, of .590 for Average Iota, of .521 for Minimum Iota, of .672 for Krippendorff's Alpha and of .821 for Percentage Agreement is necessary in the case of a medium true association to ensure with a certainty of 95% that a Type I Error occurs in less than 10%. The following section presents the results for Type II Errors.

Table 22. Residual Standard Error of Prediction for Type I Errors

Nominal Data					
	No Relationship	Weak Relationship	Medium Relationship	Strong Relationship	Perfect Relationship
Iota Index	.42856035	.42958436	.36225307	.30446269	.26792103
Average Iota	.42855644	.44557885	.39834627	.33746289	.30799488
Minimum Iota	.42975522	.45208895	.41157883	.36516304	.30803597
Krippendorff's Alpha	.42805091	.43490002	.37833164	.32354269	.27818507
Percentage Agreement	.43092232	.45790905	.4264159	.37081686	.31950746
Ordinal Data					
	No Relationship	Weak Relationship	Medium Relationship	Strong Relationship	Perfect Relationship
Iota Index	.45801823	.40695612	.37214645	.35501956	.34025083
Average Iota	.45829256	.43491035	.40602906	.38261695	.36704729
Minimum Iota	.45918752	.44925069	.42215707	.39961533	.39505928
Krippendorff's Alpha	.4576995	.40989475	.38421015	.36218248	.34144588
Percentage Agreement	.45979262	.45411463	.44047686	.41527418	.39605687

6.2.4 Analyses of Type II Errors

Table 23 reports the details for Type II Errors. Compared to the absence of a true association/correlation, the degree of reliability is more important when there is *at least* a true weak association/correlation present. Furthermore, the sample size becomes the relevant influencing factor showing high values for R^2. Surprisingly, if there is a weak true association/correlation, higher reliability values imply an increased risk for Type II Errors while a larger sample size implies a lower risk.

Table 23. Standardized Regression Coefficient for Type II Errors.

		Nominal Data				Ordinal Data			
Measure		Measure	$R^2_{Level\ 1}$	Sample Size	$R^2_{Level\ 2}$	Measure	$R^2_{Level\ 1}$	Sample Size	$R^2_{Level\ 2}$
Index	0	.062*	.008	-.068*	.005	.047*	.005	-.096*	.009
	1	.163*	.053	-.515*	.266	.137*	.037	-.769*	.591
Average	0	.024	.001	-.068*	.005	.020	.001	-.096*	.009
	1	.130*	.035	-.515*	.266	.143*	.041	-.769*	.591
Minimum	0	-.043*	.004	-.068*	.005	.009	0	-.096*	.009
	1	.040*	.003	-.515*	.265	.152*	.047	-.769*	.591
Alpha	0	.058*	.007	-.068*	.005	.024*	.001	-.096*	.009
	1	.164*	.055	-.515*	.266	.145*	.040	-.769*	.591
Percent	0	-.057*	.008	-.068*	.005	.026*	.002	-.096*	.009
	1	.013	0	-.515*	.265	.143*	.046	-.769*	.591

Note.
* significant at the 5%-Level.
Coefficient of reliability measures for independent and dependent variable constrained to be equal.
0 = no true relationship, 1 = weak relationship, 2 = medium relationship, 3 = strong relationship, 4 = perfect relationship.

Figure 19 illustrates the expected risk for a Type II Error. The curves are below the dashed lines for 10% in every case and even below the line for 5% of occurrence for a broad range of reliability. This illustrates that the sample size rather than the reliability is a crucial factor of influence.

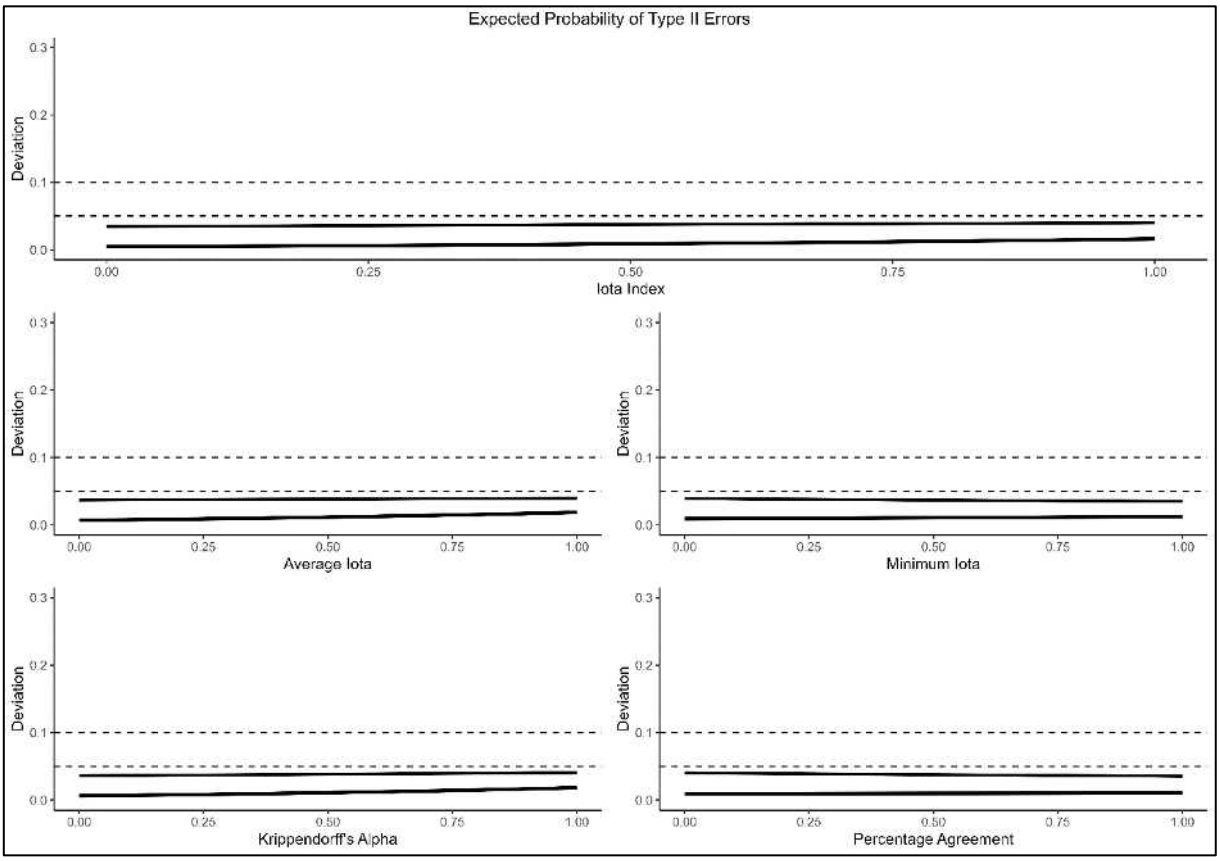


Figure 19. Expected Risk of Type II Errors (Nominal Data).

It is not possible to derive cut-off values for Type II Errors, since higher values of reliability are more error-prone. Furthermore, the values of R^2 are low for the reliability measures.

6.2.5 Analysis of Correct Classification of Effect Sizes

Table 24 presents the standardized coefficients and the values for the explained variance on the different chances for the correct classification of effect sizes. In general, a higher degree of reliability leads to an increased chance that the estimated sample associations/correlations lead to the same classification of effect size as the true sample association/correlation. Surprisingly, a higher sample size in the core study is associated with a decreasing chance to classify the effect size correctly. However, this influence is only weak. In cases with no true association/correlation, a higher sample size increases the chance for a correct classification.

Table 24. Standardized Regression Coefficient for the Correct Classification of Effect Sizes.

		Nominal Data				Ordinal Data			
Measure		Measure	$R^2_{Level\ 1}$	Sample Size	$R^2_{Level\ 2}$	Measure	$R^2_{Level\ 1}$	Sample Size	$R^2_{Level\ 2}$
Index	0	.031*	.002	.205*	.042	.052*	.006	.369*	.136
	1	.492*	.484	-.342*	.117	.492*	.48	-.162*	.026
	2	.521*	.538	-.25*	.062	.546*	.636	-.156*	.024
	3	.608*	.757	-.231*	.053	.644*	.831	-.206*	.043
	4					.504*	.523	-.044*	.002
Average	0	.076*	.012	.205*	.042	.044*	.004	.369*	.136
	1	.442*	.408	-.342*	.117	.465*	.436	-.162*	.026
	2	.49*	.495	-.249*	.062	.502*	.548	-.155*	.024
	3	.558*	.658	-.231*	.054	.577*	.684	-.206*	.043
	4								
Minimum	0	.148*	.048	.205*	.042	.019*	.001	.369*	.136
	1	.350*	.264	-.342*	.117	.435*	.386	-.162*	.026
	2	.434*	.397	-.249*	.062	.473*	.487	-.155*	.024
	3					.538*	.593	-.206*	.043
	4								
Alpha	0	.051*	.006	.205*	.042	.047*	.005	.369*	.136
	1	.494*	.492	-.342*	.117	.528*	.538	-.162*	.026
	2	.528*	.553	-.25*	.062	.567*	.676	-.156*	.024
	3	.613*	.766	-.235*	.055	.659*	.864	-.207*	.043
	4					.669*	.943	-.039*	.002
Percent	0	.155*	.058	.205*	.042	.007	0	.369*	.136
	1	.273*	.174	-.342*	.117	.395*	.353	-.162*	.026
	2	.373*	.32	-.249*	.062	.453*	.492	-.155*	.024
	3					.551*	.701	-.206*	.043
	4								

Note.
* significant at the 5%-Level.
Coefficient of reliability measures for independent and dependent variable constrained to be equal.
0 = no true relationship, 1 = weak relationship, 2 = medium relationship, 3 = strong relationship,
4 = perfect relationship.

Figure 20 presents the expected probability for a correct classification of the effect size for the different measures and different configurations of true association. The dashed horizontal lines indicate a chance of 90% or 95% to correctly classify the effect size. Table 25 shows the concrete intersections of the curves with these horizontal lines in order to derive cut-off values.

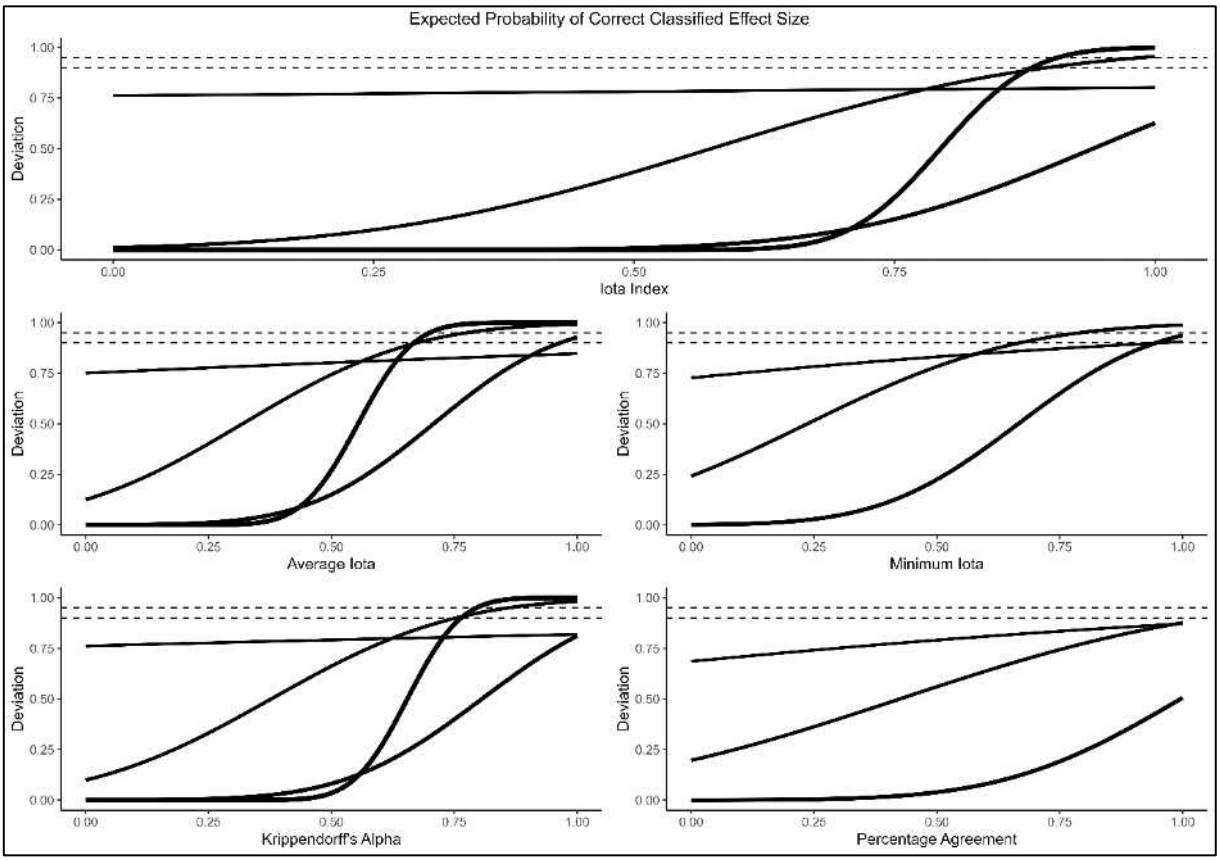


Figure 20. Expected Chance for a Correct Classification of Effect Sizes (Nominal Data)

Table 25. Potential Cut-Off Values for Effect Size Classification

		Nominal				Ordinal			
		Expected Deviation		95% Interval		Expected Deviation		95% Interval	
		more 95%	more 90%	more 95%	more 90%	more 95%	more 90%	more 95%	more 90%
Iota Index	0	1	1	1	1	1	.532	1	1
	1	.987	.896	1	1	1	1	1	1
	2	1	1	1	1	1	1	1	1
	3	.907	.882	.941	.916	1	1	1	1
	4					.817	.809	.827	.819
Average Iota	0	1	1	1	1	1	.269	1	1
	1	.775	.674	.979	.878	.859	.783	.992	.917
	2	1	.965	1	1	.971	.918	1	.975
	3	.7	.668	.747	.715	.915	.884	.937	.906
	4								
Minimum Iota	0	1	.957	1	1	1	.095	1	1
	1	.791	.669	1	.924	.788	.71	.927	.85
	2	1	.946	1	1	.915	.859	.975	.92
	3					.877	.843	.902	.868
	4								
Krippendorff's Alpha	0	1	1	1	1	1	.317	1	1
	1	.860	.753	1	.963	.964	.882	1	1
	2	1	1	1	1	1	1	1	1
	3	.795	.764	.837	.806	.977	.949	.995	.968
	4					.649	.634	.662	.648
Percentage Agreement	0	1	1	1	1	1	0	1	1
	1	1	1	1	1	1	1	1	1
	2	1	1	1	1	1	1	1	1
	3					1	1	1	1
	4								

0 = no true relationship, 1 = weak relationship, 2 = medium relationship, 3 = strong relationship, 4 = perfect relationship.

The interpretation is similar to that of the other sections. If the chance to correctly classify the effect size should be at least 90% under the condition of a strong true association, the Iota Index must be at least .882, Average Iota at least .668 and Krippendorff's Alpha at least .764. For Minimum Iota and percentage agreement, no model could be estimated.

To consider the uncertainty in estimations, Table 26 reports the standard error of estimates. This is employed when calculating the curves to ensure that the chances of a correct classification hold true in 95% of cases. According to Table 25, the Iota Index must be at least .916, Average Iota at least .715 and

Krippendorff's Alpha at least .806 to ensure that the chance for a correct classification is at least 90% (for a strong true association) with a certainty of 95%. The following section summarizes the results.

Table 26. Residual Standard Error of Prediction for Classification of Effect Sizes

Nominal Data					
	No Relationship	Weak Relationship	Medium Relationship	Strong Relationship	Perfect Relationship
Iota Index	.39153958	.43651241	.26855155	.30533175	
Average Iota	.39123590	.44689353	.26566314	.32420636	
Minimum Iota	.39047224	.46099874	.27235053		
Krippendorff's Alpha	.39139074	.43610799	.25940090	.29713650	
Percentage Agreement	.39018655	.46794704	.28554146		
Ordinal Data					
	No Relationship	Weak Relationship	Medium Relationship	Strong Relationship	Perfect Relationship
Iota Index	.28583459	.37786995	.23781684	.15467564	.26651092
Average Iota	.28586547	.39029558	.23713670	.15774974	
Minimum Iota	.28597555	.39910332	.24090050	.16234238	
Krippendorff's Alpha	.28586989	.36912086	.22394821	.14539892	.20545627
Percentage Agreement	.28601844	.39940086	.25039935	.16569055	

6.3 Summary of Simulation Study II

Simulation study II aims to provide first insights into the predictive power of the new Iota Index, its performance compared to other measures and potential cut-off values for practice. To provide a realistic picture, only the data for the range of true associations/correlations relevant for practice is used for both this summary and further analysis since a perfect relationship cannot be assumed to be plausible in practical situations. The range of strength of association/correlation relevant for practice is based on the classification provided by Cohen (1988). With the help of simulation study II, the following answers are currently possible:

RQ1 How strong does the Iota Index predict the deviation between true and estimated sample association/correlation for nominal/ordinal data to be?

Two-level regression analysis indicates that the Iota Index explains between 10.1% (nominal data, no true association) to 68.0% (ordinal data, strong correlation) of the variation in the deviation between true and estimated association.

RQ2 How strong does the Iota Index predict Type I and Type II Errors in accepting and rejecting hypothesis of association/correlation to be?

In the current study, a Type I Error refers to situations where the true sample association/correlation implies the acceptance of an association (the null hypothesis is rejected) while the estimated sample association/correlation implies a rejection (the null hypothesis is accepted). The two-level probit regression of simulation study II shows that for nominal data, the corresponding R^2 is between 70.6% and 74.6% for nominal data and between 64.4% and 68.1% for ordinal data. In the case of absence of a true association/correlation, R^2 ranges between 4.4% (nominal data) and 9.4% (ordinal data). Thus, for a broad range of practical applications the Iota Index shows a relatively high predictive power.

A Type II Error occurs in situations where the existence of an association/correlation in the true sample is rejected (acceptance of null hypothesis) while the estimated sample implies the acceptance of an association/correlation (rejection of null hypothesis). In this case, the Iota Index explains about 5.3% of the variance in the best case.

RQ3 How strong does the Iota Index predict the correct effect size of an association/correlation to be?

The two-level probit regression analyzes the relationship between the numerical value of the Iota Index and the chance to correctly classify an effect size. That is, the chance to correctly classify an association or correlation as practically irrelevant, as a small, medium or strong effect based on the work of Cohen (1988). The analysis reveals that R^2 increases for stronger true effects ranging from 48.4% (weak association) to about 75.7% for a strong association. In the case of ordinal data, R^2 ranges from 48.0% to 83.1%. In case of absence of a true association (nominal data), R^2 is only 0.2% and in case of absence of a true correlation (ordinal data), R^2 is only 0.6%.

RQ4 How does the Iota Index perform in predicting consequences in comparison to the old Iota Concept and other measures of inter-rater reliability?

Comparing the new Iota Index with the old measures from the Iota Concept of the first generation (Average Iota, Minimum Iota), it is revealed that the new index produces higher values for nearly all prediction tasks and both kinds of data

(nominal, ordinal). The differences are significant. For example, referring to nominal data and a strong true association, the Iota Index explains about 72.5% of variance in Type I Errors while Average Iota explains only 55.6% and Minimum Iota about 42.3%. Thus, the second generation of the Iota Concept is a real improvement compared to the old concept.

Comparing the Iota Index with other established measures, only Krippendorffs' Alpha reveals a predictive power comparable to the Iota Index. For all estimated models, Krippendorff's Alpha and the Iota Index perform in a similar manner. In some situations, Krippendorff's Alpha is slightly better, in some the Iota Index. Percentage Agreement performs the worst in nearly all models.

RQ5 What are meaningful cut-off values for reliability in practical situations?

The derivation of meaningful cut-off values based merely on the analysis is a difficult task. The results indicate that even on their highest level, not all measures can ensure error-free measurements since the reliability values themselves are estimates for reliability, thus suffering from estimation errors. Table 27 summarizes the derived cut-off values. The table shows the most demanding conditions and the highest values necessary for the different target variables for both nominal and ordinal data. For Type I Errors, the situations with the most demanding conditions are ambiguous. Thus, the row with the highest value for the weak practical deviations is chosen to ensure the highest possible minimal level of reliability.

Table 27. Summary of Cut-Off Values

		Nominal					Ordinal			
		Expected Effect		95% Interval			Expected Deviation		95% Interval	
		No practical	Weak practical	No practical	Weak practical		No practical	Weak practical	No practical	Weak practical
Iota Index										
Deviation	3	.966	.604	1	.890	3	1	.762	1	.992
Type I Error	1	1	.920	1	1	1	.919	.858	1	.971
Type II Error	-/-	-/-	-/-	-/-	-/-	-/-	-/-	-/-	-/-	-/-
Effect Size	1	.987	.896	1	1	1	1	1	1	1
Average Iota										
Deviation	3	.755	.344	1	.696	3	.847	.521	1	.791
Type I Error	1	.783	.693	.965	.875	1	.687	.621	.817	.751
Type II Error	-/-	-/-	-/-	-/-	-/-	-/-	-/-	-/-	-/-	-/-
Effect Size	2	1	.965	1	1	2	.971	.918	1	.975
Minimum Iota										
Deviation	3	.717	.250	1	.687	3	.785	.450	1	.747
Type I Error	1	.715	.623	.904	.812	1	.624	.554	.767	.697
Type II Error	-/-	-/-	-/-	-/-	-/-	-/-	-/-	-/-	-/-	-/-
Effect Size	2	1	.946	1	1	2	.915	.859	.975	.92
Krippendorff's Alpha										
Deviation	3	.863	.416	1	.774	3	.963	.606	1	.861
Type I Error	1	.886	.788	1	.982	1	.774	.703	.906	.835
Type II Error	-/-	-/-	-/-	-/-	-/-	-/-	-/-	-/-	-/-	-/-
Effect Size	2	1	1	1	1	2	1	1	1	1
Percentage Agreement										
Deviation	3	1	.437	1	1	3	1	.697	1	1
Type I Error	1	1	.940	1	1	1	.943	.850	1	1
Type II Error	-/-	-/-	-/-	-/-	-/-	-/-	-/-	-/-	-/-	-/-
Effect Size	2	1	1	1	1	1	1	1	1	1

0 = no true relationship, 1 = weak relationship, 2 = medium relationship, 3 = strong relationship, 4 = perfect relationship.

The cells in Table 27 with the highest possible value of 1 indicate that the measure cannot ensure that the deviation or error rate has no or only weak impact on practice. This applies to nearly all target variables and all measures for the 95% prediction interval for no practically relevant deviation and error rate. This means that even if the measure reaches its highest value, it cannot ensure with a certainty of 95% that the deviation and error rates have no practical impact on the generated data. For only weak practical impacts, the situation is less problematic regarding the 95% interval. In most cases regarding the expected values, the measures do not need their maximum value to ensure no or only a weak practical impact.

Table 28 provides a first suggestion for cut-off values in practice by choosing the highest values for the measures. The correct classification of effect sizes is not considered in this table as it is a stricter application of the deviation.

Table 28. Recommendations for Cut-Off Values

	Minimal	Satisfactory	Good	Excellent
Evaluation Category	Expectation of weak practical effect	Expectation of no practical effect	Certainly, only a weak practical effect	Certainly, no practical effect
Iota Index	.920	1.00*	-/-	-/-
Average Iota	.693	.847	.875	-/-
Minimum Iota	.623	.785	.812	-/-
Krippendorff's Alpha	.788	.963	.982	-/-
Percentage Agreement	.940	1.00*	-/-	-/-

Note:
weak practical effect = Deviation less 0.3 and Type I Error less 10%.
no practical effect = Deviation less 0.1 and Type I Error less 5%.
* Limit of the Scale reached

Table 28 shows the minimum values necessary for both variables to ensure no or only weak practical impact on the data. For example, a value of .693 for Average Iota justifies the expectation that the deviation and the Type I Error rates only have a weak practical impact on the data (deviation less 0.3, Type I Error less 10%). A value of .847 justifies the expectation that deviation and the Type I Error rates have no practical effect (deviation less 0.1, Type I Error less 5%) and a value of .875 ensures with a certainty of 95% that the deviation and the error rate only have a weak practical effect (deviation less 0.3, Type I Error less 10%). Ensuring no practical effects with a certainty of 95% is not possible with Average Iota as Average Iota must be greater 1, which is not possible.

The structure for the other measures is similar. No measure provides values which can ensure with a certainty of 95% that generated data is practically not biased. For the Iota Index and Percentage Agreement, even for a small practical effect, there are no values derivable. Figure 21 illustrates the problem.

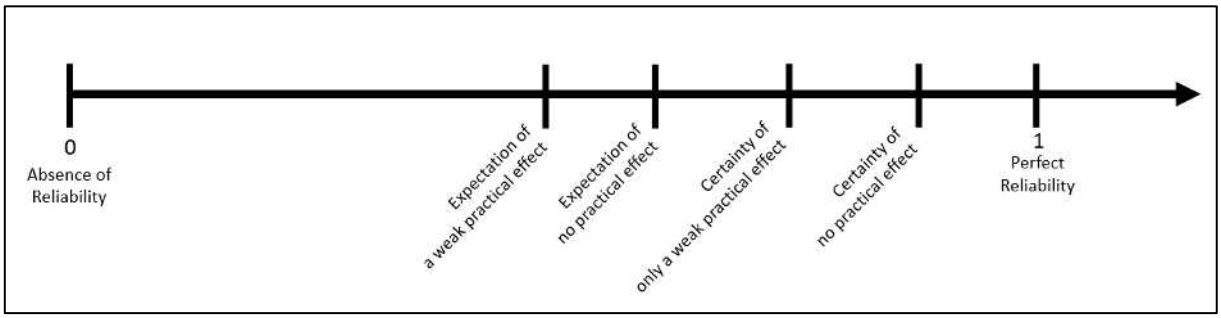


Figure 21. Demands for an Ideal Reliability Measure

Ideally, a reliability measure would achieve its highest value only if it is certain that the generated data is not biased by errors. As Table 28 shows, the necessary values for certainty of no practical effect are outside the possible range. Computationally, the values must be greater 1. The consequence of the results in Table 28 is that the metrics used in the investigated measures of reliability do not account for the uncertainty of the reliability estimation itself. High values are reached too fast. To achieve an ideal measure as it is shown in Figure 21, "breaks" should be integrated to ensure that the values do not increase too quickly. For the Iota Index, this aim is realizable by adapting Equation 9, as shown in Equation 24.

$$Iota_{Index}^{d} = \frac{1}{\left(1-\frac{1}{c}\right)^{d} + (c-1)\left(0-\frac{1}{c}\right)^{d}} * \sum_{\forall i} p_i \left(\sum_{\forall j} \left(\left| a_{ij} - \frac{1}{c} \right| \right)^{d} \right), d \geq 1 \qquad [24]$$

For values of d greater 1, the assigned values are lower compared to the normal Iota Index. At the same time, the new index is still normalized in the range between 0, indicating the absence of reliability, and 1, indicating perfect reliability. This frees capacities at the end of the scale for mapping the degree of certainty on the scale. The point "certainty of no practical effect" is shifted into the possible range of values (see Figure 21). The next simulation study tries to prove this idea and to find a suitable value for d.

7 Simulation Study III

7.1 Design of the Study

Simulation Study II revealed that the different reliability measures do not account for the uncertainty of estimates and predictions. Simulation study III takes up this result and tries to transform the Iota Index so that high values imply a high certainty that the estimated sample statistics do not deviate from error-free sample statistics in a range relevant for practice.

This requires a transformation of the Iota Index that has meaningful end points. That is, a value of zero should indicate the absence of reliability while a value of one indicates perfect reliability with certainty. In terms of the Iota Concept, this implies that zero still must correspond to an Assignment Error Matrix for random assignments while the value of one must still correspond to an Assignment Error Matrix with maximal distance from random assignments.

To provide capacities in the range between zero and one for the case of uncertainty, the values of the Iota Index should be transformed in a way that is more difficult to achieve high values compared to low values. In other words, the scale needs a "brake" at the top level of the scale. In the very best case, the transformed scale has the same predictive power as the original scale of the Iota Index.

To achieve this aim, two different types of transformations of the Iota Index are investigated. Equation 25 shows a static transformation. It represents a generalization of the original equation of the Iota Index (Equation 9). Values of d greater one map the generated values to smaller values. At the same time, they ensure that the order of the values is the same as for the Iota Index and the scale still ranges from zero to one.

$$Iota_{Index}^{d} = \frac{1}{\left(1-\frac{1}{c}\right)^{d} + (c-1)\left(0-\frac{1}{c}\right)^{d}} * \sum_{\forall i} p_i \left(\sum_{\forall j} \left(\left| a_{ij} - \frac{1}{c} \right| \right)^{d} \right), d \geq 1 \qquad [25]$$

Equation 26 shows a dynamic transformation. Here, the values of d_{dyn} depend on the level of the Iota Index. Higher levels of the Iota Index lead to a higher exponent which in turn leads to smaller values. The parameter d_{dyn} controls where the "brake" should affect the scale. For example,

$d_{dyn} = 0.5$ decreases the value from the very beginning of the scale while $d_{dyn} = 2$ compresses the value later.

$$Iota_{Index}^{d_{dyn}} = (Iota\ Index)^{1+(Iota\ Index)^{d_{dyn}}}, d_{dyn} > 0 \qquad [26]$$

Since the static transformation makes a recalculation necessary, the simulation from study II is repeated with the same configuration. The static transformation of the Iota Index is done for $d = 1.5, 2, 3, 4$. For the dynamic transformation, d is either 0.5 or 2. The following section reports the results.

7.2 Results of Simulation Study III

7.2.1 Overview

In the third simulation study, 5,925,743 coding processes are simulated which are nested in 197,525 processes for reliability estimation. Both processes are based on 24,910 sets of true parameters. Table 29 shows the distribution of the processes to nominal and ordinal data as well as to different strengths of true association/correlation.

Table 29. Sample Size of Simulation Study III

	Nominal Data				
	No Relationship	Weak Relationship	Medium Relationship	Strong Relationship	Perfect Relationship
Set of True Parameters	1,741	2,191	1,828	4,197	2,518
$N_{Between}$ (Reliability Estimation)	13,807	17,340	14,496	33,368	19,971
N_{Within} (Coding Processes)	414,210	520,199	434,880	1,001,040	599,124
	Ordinal Data				
	No Relationship	Weak Relationship	Medium Relationship	Strong Relationship	Perfect Relationship
Set of True Parameters	2,741	2,851	2,592	1,821	2,430
$N_{Between}$ (Reliability Estimation)	21,697	22,595	20,538	14,453	19,260
N_{Within} (Coding Processes)	650,910	677,850	616,140	433,590	577,800

Appendix D presents values for assessing the global model fit. All models could be successfully estimated, except some models for the correct classification of effect sizes for a perfect and strong relationship. Thus, these models cannot be considered in the following analysis. The reaming models show a global model fit according to the criteria developed by Hu and Bentler (1999). There are only two exceptions concerning the CFI. The models for the Iota Index and its dynamic transformation with $d_{dyn} = 2$ show values below .950 for the correct

classification of effect sizes in ordinal data for the condition of practically no relationship. Since RMSEA and SRMR are in line with the combination rule of Hu and Bentler (1999), both models remain part of the analysis. Table 30 reports the values for R^2 for the Iota Index and its transformations.

Table 30. $R^2_{Levl\ 1\ only}$ for Prediction of the Average Deviation and Error Rates

Nominal Data					
Deviation					
	No Relationship	Weak Relationship	Medium Relationship	Strong Relationship	Perfect Relationship
Iota Index	.114	.189	.415	.552	.778
$d = 1.5$	.097	.162	.377	.528	.770
$d = 2.0$	.082	.137	.338	.496	.740
$d = 3.0$	.062	.108	.280	.440	.673
$d = 4.0$	.054	.095	.248	.402	.620
$d_{dyn} = 0.5$	.091	.155	.360	.514	.748
$d_{dyn} = 2.0$	.106	.178	.395	.535	.764
Type I Error					
	No Relationship	Weak Relationship	Medium Relationship	Strong Relationship	Perfect Relationship
Iota Index	.064	.699	.747	.722	.654
$d = 1.5$	.058	.685	.722	.687	.606
$d = 2.0$	.051	.657	.683	.638	.530
$d = 3.0$	.042	.595	.605	.550	.412
$d = 4.0$	.037	.544	.545	.487	.325
$d_{dyn} = 0.5$	.056	.659	.684	.656	.570
$d_{dyn} = 2.0$	.061	.680	.719	.699	.634
Effect Size					
	No Relationship	Weak Relationship	Medium Relationship	Strong Relationship	Perfect Relationship
Iota Index	.003	.476	.556	.721	
$d = 1.5$	.007	.439	.54	.676	
$d = 2.0$	.011	.399	.519	.639	
$d = 3.0$	.018	.343	.481		
$d = 4.0$	.020	.311	.453		
$d_{dyn} = 0.5$	.008	.430	.534	.655	
$d_{dyn} = 2.0$	.004	.458	.551	.684	

Table 16. $R^2_{Levl\ 1\ only}$ for Prediction Average Deviation and Error Rates (Continue)

Ordinal Data					
Deviation					
	No Relationship	Weak Relationship	Medium Relationship	Strong Relationship	Perfect Relationship
Iota Index	.271	.398	.607	.679	.704
$d = 1.5$	.277	.400	.615	.686	.714
$d = 2.0$	.273	.309	.602	.607	.699
$d = 3.0$	.256	.359	.558	.619	.649
$d = 4.0$	.239	.331	.516	.571	.601
$d_{dyn} = 0.5$	.271	.388	.600	.666	.696
$d_{dyn} = 2.0$	.271	.392	.603	.672	.699
Type I Error					
	No Relationship	Weak Relationship	Medium Relationship	Strong Relationship	Perfect Relationship
Iota Index	.084	.660	.683	.657	.609
$d = 1.5$	.096	.648	.670	.648	.601
$d = 2.0$	.102	.618	.638	.619	.566
$d = 3.0$	.107	.554	.567	.551	.474
$d = 4.0$	.106	.503	.511	.492	.402
$d_{dyn} = 0.5$	.098	.618	.641	.618	.561
$d_{dyn} = 2.0$	.092	.635	.662	.637	.600
Effect Size					
	No Relationship	Weak Relationship	Medium Relationship	Strong Relationship	Perfect Relationship
Iota Index	.006	.479	.627	.82	
$d = 1.5$	.006	.488	.612	.781	.591
$d = 2.0$	.006	.482	.591	.751	
$d = 3.0$	.006	.455	.549	.701	.385
$d = 4.0$	.006	.424	.508	.656	
$d_{dyn} = 0.5$	.007	.477	.593	.752	.612
$d_{dyn} = 2.0$	.007	.476	.608	.771	

Table 30 emphasizes that the static transformations lead to lower values for R^2 in all cases. R^2 decreases for stronger brakes. That is, for higher values of d. Similar applies for the dynamic transformations but the values for R^2 are very close to the ones from the original Iota Index. Thus, the dynamic transformations are closely connected to the quality of the data. The following sections derive the corresponding cut-off values and prediction intervals.

7.2.2 Potential Cut-off Values and Certainty of Reliability Effects for Deviation

Table 31 reports the necessary values for achieving a specific deviation between true and estimated sample association/correlation for different degrees of certainty.

First, for the static and the dynamic transformations, the necessary values are lower compared to the original Iota Index. For example, achieving an expected deviation of less than 0.1 for nominal data with a strong true association requires a Iota Index value of .829, while for Iota Index ($d = 4$) a value of .627 is sufficient.

Second, the different versions of transformation of the Iota Index decrease the values but they do not provide clear cut-off values for a certainty of 95% under the condition of perfect relationships. The only exception is the static transformation with $d = 4$, which provides a clear cut-off value for nominal data.

In order to provide insights into how certainly the maximal value of the scales are associated with specific effects, Table 32 reports the corresponding probabilities for no and only a weak practical deviation.

Table 31. Potential Cut-Off Values for Deviation

		Nominal				Ordinal			
		Expected Deviation		95% Interval		Expected Deviation		95% Interval	
		less 0.1	less 0.3	less 0.1	less 0.3	less 0.1	less 0.3	less 0.1	less 0.3
Iota Index	0	0	0	0	0	0	0	.664	0
	1	.421	0	1	0	.771	0	1	.343
	2	.829	.241	1	.650	.997	.557	1	.839
	3	.954	.605	1	.882	1	.757	1	.981
	4	1	.932	1	1	1	.967	1	1
$d = 1.5$	0	0	0	0	0	0	0	.575	0
	1	.301	0	1	0	.694	0	1	.229
	2	.769	.107	1	.580	.937	.462	1	.763
	3	.904	.510	1	.829	1	.678	1	.918
	4	1	.870	1	1	1	.905	1	1
$d = 2.0$	0	0	0	0	0	0	0	.51	0
	1	.211	0	1	0	.637	0	1	.150
	2	.728	.008	1	.536	.889	.392	1	.712
	3	.865	.440	1	.795	.957	.62	1	.876
	4	1	.823	1	1	1	.858	1	1
$d = 3.0$	0	0	0	0	0	0	0	.427	0
	1	.093	0	1	0	.557	0	1	.053
	2	.670	0	1	.48	.821	.3	1	.649
	3	.807	.346	1	.75	.897	.539	1	.828
	4	.983	.755	1	1	1	.792	1	1
$d = 4.0$	0	0	0	0	0	0	0	.376	0
	1	.028	0	1	0	.506	0	1	0
	2	.627	0	1	.442	.773	.243	1	.612
	3	.762	.289	1	.716	.853	.488	1	.799
	4	.937	.706	1	.976	.964	.746	1	1
$d_{dyn} = 0.5$	0	0	0	0	0	0	0	.529	0
	1	.258	0	1	0	.647	0	1	.190
	2	.724	.063	1	.541	.884	.418	1	.718
	3	.853	.464	1	.784	.951	.631	1	.875
	4	1	.819	1	1	1	.856	1	1
$d_{dyn} = 2.0$	0	0	0	0	0	0	0	.59	0
	1	.360	0	1	0	.693	0	1	.291
	2	.752	.189	1	.586	.904	.491	1	.756
	3	.868	.534	1	.804	.961	.680	1	.893
	4	1	.843	1	1	1	.878	1	1

0 = no true relationship, 1 = weak relationship, 2 = medium relationship, 3 = strong relationship, 4 = perfect relationship.

Table 32. Certainty at the End of the Scale for Deviation

		Nominal		Ordinal	
		less 0.1	less 0.3	less 0.1	less 0.3
Iota Index	0	1.000	1.000	.986	1.000
	1	.906	1.000	.794	1.000
	2	.755	.999	.507	.995
	3	.608	.991	.343	.963
	4	.126	.747	.119	.608
$d = 1.5$	0	1.000	1.000	.989	1.000
	1	.910	1.000	.844	1.000
	2	.789	.999	.635	.999
	3	.691	.994	.499	.987
	4	.260	.872	.227	.766
$d = 2.0$	0	1.000	1.000	.991	1.000
	1	.909	1.000	.872	1.000
	2	.802	.999	.716	.999
	3	.734	.995	.608	.993
	4	.383	.917	.333	.843
$d = 3.0$	0	1.000	1.000	.993	1.000
	1	.908	1.000	.901	1.000
	2	.817	.999	.801	1.000
	3	.784	.996	.722	.996
	4	.546	.948	.486	.904
$d = 4.0$	0	1.000	1.000	.994	1.000
	1	.913	1.000	.918	1.000
	2	.833	.999	.845	1.000
	3	.821	.997	.781	.997
	4	.649	.964	.581	.928
$d_{dyn} = 0.5$	0	1.000	1.000	.991	1.000
	1	.923	1.000	.880	1.000
	2	.829	1.000	.737	.999
	3	.775	.997	.630	.994
	4	.433	.938	.363	.861
$d_{dyn} = 2.0$	0	1.000	1.000	.991	1.000
	1	.933	1.000	.877	1.000
	2	.849	1.000	.725	.999
	3	.789	.998	.617	.993
	4	.426	.943	.346	.851

Table 32 shows that on their highest values the certainty to ensure no or only a weak practical deviation is very different for the different kinds of transformations. For example, in the case of nominal data and a strong association, a value of “1” for the original Iota Index implies that in about 60.8% of cases, the deviation has no practical relevant size. For the static transformation of the Iota Index with $d = 4$, “1” implies that the estimated and

true sample associations deviate with no practically relevant size with a certainty of 82.1%. For the dynamic transformation with $d_{dyn} = 2$, the certainty is about 78.9%. Thus, the transformations increase the certainty at the end of the scale meaningfully compared to the original Iota Index. The dynamic brakes perform similarly well as the static brake with $d = 4$. Only in case of perfect true relationships the static transformation with $d = 3$ or $d = 4$ performs better.

7.2.3 Potential Cut-off Values and Certainty of Reliability Effects for Type I Errors

Table 33 reports the potential cut-off values for the risk of Type I Errors. Again, the transformations lead to lower necessary cut-off values. Compared to the original Iota Index, all static transformations with at least $d = 2$ and both dynamic transformations provide a clear cut-off value for the 95% prediction interval for only weak practical deviations under the condition of weak relationships. For no practical effects, a clear cut-off value on the 95% prediction interval does not exist for the different transformations of the Iota Index.

Table 33. Potential Cut-Off Values for Type I Errors

		Nominal				Ordinal			
		Expected Deviation		95% Interval		Expected Deviation		95% Interval	
		less 5%	less 10%	less 5%	less 10%	less 5%	less 10%	less 5%	less 10%
Iota Index	0	.837	.514	1	1	1	.786	1	1
	1	.998	.915	1	1	.912	.851	1	.96
	2	.796	.749	.872	.825	.729	.69	.796	.756
	3	.625	.591	.672	.638	.643	.608	.698	.663
	4	.484	.455	.519	.49	.557	.527	.605	.575
$d = 1.5$	0	.779	.417	1	1	.967	.697	1	1
	1	.943	.852	1	1	.848	.782	.969	.903
	2	.719	.669	.803	.753	.643	.6	.716	.673
	3	.525	.488	.578	.54	.548	.511	.607	.571
	4	.371	.34	.409	.379	.455	.423	.506	.474
$d = 2.0$	0	.740	.346	1	1	.898	.629	1	1
	1	.899	.803	1	.995	.798	.729	.929	.859
	2	.661	.609	.751	.699	.578	.534	.656	.612
	3	.451	.412	.509	.469	.477	.439	.54	.502
	4	.282	.248	.327	.293	.379	.345	.433	.399
$d = 3.0$	0	.686	.252	1	1	.802	.535	1	1
	1	.835	.734	1	.939	.728	.654	.871	.797
	2	.582	.528	.681	.626	.491	.444	.576	.53
	3	.355	.312	.419	.377	.383	.344	.451	.412
	4	.162	.124	.215	.178	.274	.237	.335	.298
$d = 4.0$	0	.645	.195	1	1	.738	.476	1	1
	1	.790	.686	1	.898	.679	.604	.828	.753
	2	.531	.476	.635	.579	.436	.389	.524	.477
	3	.296	.252	.365	.321	.325	.284	.397	.356
	4	.081	.040	.142	.101	.203	.165	.271	.232
$d_{dyn} = 0.5$	0	.731	.374	1	1	.903	.644	1	1
	1	.893	.803	1	.983	.798	.733	.921	.855
	2	.670	.620	.755	.706	.592	.55	.664	.623
	3	.476	.439	.528	.491	.497	.462	.556	.521
	4	.322	.292	.361	.331	.406	.373	.458	.425
$d_{dyn} = 2.0$	0	.759	.452	1	1	.935	.698	1	1
	1	.908	.829	1	.985	.825	.768	.931	.874
	2	.713	.670	.786	.743	.648	.611	.710	.673
	3	.548	.517	.592	.561	.566	.534	.618	.586
	4	.418	.392	.450	.424	.486	.458	.530	.502

0 = no true relationship, 1 = weak relationship, 2 = medium relationship, 3 = strong relationship, 4 = perfect relationship.

To provide insights into the certainty at the end of the scales, Table 34 reports the corresponding probabilities for ensuring that the risk of Type I Errors does not exceed a specific value (5% or 10%).

Table 34. Certainty at the End of the Scale for Type I Errors

		Nominal		Ordinal	
		less 5%	less 10%	less 5%	less 10%
Iota Index	0	.666	.899	.440	.739
	1	.507	.804	.909	.988
	2	1.000	1.000	1.000	1.000
	3	1.000	1.000	1.000	1.000
	4	1.000	1.000	1.000	1.000
$d = 1.5$	0	.697	.914	.539	.814
	1	.697	.912	.981	.999
	2	1.000	1.000	1.000	1.000
	3	1.000	1.000	1.000	1.000
	4	1.000	1.000	1.000	1.000
$d = 2.0$	0	.712	.920	.618	.863
	1	.806	.955	.995	1.000
	2	1.000	1.000	1.000	1.000
	3	1.000	1.000	1.000	1.000
	4	1.000	1.000	1.000	1.000
$d = 3.0$	0	.729	.927	.723	.917
	1	.907	.984	.999	1.000
	2	1.000	1.000	1.000	1.000
	3	1.000	1.000	1.000	1.000
	4	1.000	1.000	1.000	1.000
$d = 4.0$	0	.747	.935	.787	.944
	1	.949	.993	1.000	1.000
	2	1.000	1.000	1.000	1.000
	3	1.000	1.000	1.000	1.000
	4	1.000	1.000	1.000	1.000
$d_{dyn} = 0.5$	0	.739	.931	.617	.862
	1	.835	.964	.997	1.000
	2	1.000	1.000	1.000	1.000
	3	1.000	1.000	1.000	1.000
	4	1.000	1.000	1.000	1.000
$d_{dyn} = 2.0$	0	.747	.935	.586	.843
	1	.834	.964	.997	1.000
	2	1.000	1.000	1.000	1.000
	3	1.000	1.000	1.000	1.000
	4	1.000	1.000	1.000	1.000

Again, the transformed values increase the certainty at the end of the scale meaningfully. For example, under the condition of practically no correlation and ordinal data, “1” implies with a certainty of 40% that the risk for Type I Errors is below 5% for the original Iota Index. For the static transformation with $d = 4$, “1” implies with a certainty of 72.3% and for the dynamic transformation with $d_{dyn} = 2$ of 58.6% that the risk for Type I Errors is below 5%.

7.2.4 Potential Cut-off Values and Certainty of Reliability Effects for Classifying Effect Sizes

Table 35 reports the cut-off values for the correct classification of the effect sizes. Similar to simulation study II, the results imply the need for very high values to correctly classify the effect size. Here, the static brakes for $d = 3$ and $d = 4$ do not provide models for predicting the chance for a correct classification under the condition of strong practical effects.

New clear cut-off values compared to the original Iota Index do appear only for the dynamic brake with $d_{dyn} = 2$ in the case of nominal data and for ordinal data. However, according to Table 36, the certainty to correctly classify the effect sizes increases for all transformations compared to the original Iota Index. For example, for nominal data and the absence of a practical relevant association, the highest value of the original Iota Index guarantees with a certainty of 13.6% that the chance to correctly classify the effect size is at least 90%. For the static brake with $d = 4$, this certainty increases to 30.4%. The following section summarizes the results.

Table 35. Potential Cut-Off Values for a Correct Classification of Effect Sizes

		Nominal				Ordinal			
		Expected Deviation		95% Interval		Expected Deviation		95% Interval	
		less 0.1	less 0.3	less 0.1	less 0.3	less 0.1	less 0.3	less 0.1	less 0.3
Iota Index	0	1	1	1	1	1	.577	1	1
	1	.988	.895	1	1	1	1	1	1
	2	1	1	1	1	1	1	1	1
	3	.934	.908	.972	.945	1	1	1	1
	4								
$d = 1.5$	0	1	1	1	1	1	.484	1	1
	1	.948	.843	1	1	1	.952	1	1
	2	1	1	1	1	1	1	1	1
	3	.896	.863	.941	.909	1	1	1	1
	4					.733	.724	.744	.735
$d = 2.0$	0	1	1	1	1	1	.416	1	1
	1	.919	.805	1	1	.986	.905	1	1
	2	1	1	1	1	1	1	1	1
	3	.859	.823	.91	.874	1	.991	1	1
	4								
$d = 3.0$	0	1	1	1	1	1	.325	1	1
	1	.873	.749	1	1	.920	.835	1	.987
	2	1	1	1	1	1	1	1	1
	3					.988	.953	1	.977
	4					.612	.6	.628	.616
$d = 4.0$	0	1	1	1	1	1	.27	1	1
	1	.832	.705	1	.968	.871	.785	1	.941
	2	1	1	1	1	1	.961	1	1
	3					.957	.919	.982	.944
	4								
$d_{dyn} = 0.5$	0	1	1	1	1	1	.440	1	1
	1	.896	.792	1	1	.977	.901	1	1
	2	1	1	1	1	1	1	1	1
	3	.852	.819	.9	.866	1	.981	1	1
	4					.684	.676	.696	.687
$d_{dyn} = 2.0$	0	1	1	1	1	1	.511	1	1
	1	.902	.813	1	.991	.985	.917	1	1
	2	1	1	1	1	1	1	1	1
	3	.856	.828	.894	.867	.996	.972	1	.988
	4								

0 = no true relationship, 1 = weak relationship, 2 = medium relationship, 3 = strong relationship, 4 = perfect relationship.

Table 36. Certainty at the End of the Scale for a Correct Classification of Effect Sizes

		Nominal		Ordinal	
		less 5%	less 10%	less 5%	less 10%
Iota Index	0	.022	.136	.214	.686
	1	.543	.825	.144	.451
	2	.000	.000	.000	.000
	3	.998	1.000	.000	.006
	4				
$d = 1.5$	0	.031	.169	.234	.709
	1	.658	.890	.361	.720
	2	.000	.005	.000	.008
	3	1.000	1.000	.000	.231
	4			1.000	1.000
$d = 2.0$	0	.041	.205	.250	.727
	1	.719	.918	.564	.863
	2	.000	.027	.001	.072
	3	1.000	1.000	.043	.774
	4				
$d = 3.0$	0	.061	.265	.278	.754
	1	.795	.948	.808	.963
	2	.010	.163	.052	.488
	3			.796	1.000
	4			1.000	1.000
$d = 4.0$	0	.077	.304	.304	.777
	1	.854	.968	.913	.988
	2	.065	.425	.282	.842
	3			.998	1.000
	4				
$d_{dyn} = 0.5$	0	.036	.188	.257	.734
	1	.795	.950	.610	.887
	2	.001	.054	.004	.145
	3	1.000	1.000	.208	.952
	4			1.000	1.000
$d_{dyn} = 2.0$	0	.028	.161	.255	.732
	1	.817	.958	.584	.874
	2	.002	.065	.007	.197
	3	1.000	1.000	.645	.998
	4				

7.3 Summary of Simulation Study III

Introducing different transformations of the Iota Index leads to a more effective use of the range between zero and one. In every case, the necessary cut-of values decreased compared to the original Iota Index. In some cases, even a clear cut-

off value occurred that ensures with a certainty of 95% that the deviation does not exceed a small practical effect. Although the transformations are not able to provide a clear cut-off value ensuring no practical effect with a certainty of 95%, the chance for no practical effect increased meaningfully. Thus, the transformations are partly successful.

Choosing the best transformation is a difficult task. The static transformations provide higher certainty at the end of the scale compared to the dynamic transformations. In contrast, the dynamic transformations show higher values for R^2 compared to the static transformations. Thus, the dynamic transformations are more connected to the quality of the data. To decide about the "best" transformation, Table 37 summarizes the cut-off values by showing the highest necessary values for deviation and Type I Errors for the practical relevant range of true associations/correlations. The values for the correct classification of effect sizes are not considered since they do not provide hints for good cut-off values. Values in brackets show the probability that an observation is in line with no or only a weak practical effect.

Table 37. Summary of Cut-Off Values

		Nominal					Ordinal			
		Expected Effect		95% Interval			Expected Deviation		95% Interval	
		No practical	Weak practical	No practical	Weak practical		No practical	Weak practical	No practical	Weak practical
$d = 1.5$										
Deviation	3	0.904	0.510	1 (.691)	0.829	3	1	0.678	1 (.499)	0.918
Type I Error	1	0.943	0.852	1 (.697)	1 (.912)	1	0.848	0.782	0.969	0.903
$d = 2.0$										
Deviation	3	0.865	0.440	1 (.734)	0.795	3	0.957	0.62	1 (.608)	0.876
Type I Error	1	0.899	0.803	1 (.806)	0.995	1	0.798	0.729	0.929	0.859
$d = 3.0$										
Deviation	3	0.807	0.346	1 (.784)	0.750	3	0.897	0.539	1 (.722)	0.828
Type I Error	1	0.835	0.734	1 (.907)	0.939	1	0.728	0.654	0.871	0.797
$d = 4.0$										
Deviation	3	0.762	0.289	1 (.821)	0.716	3	0.853	0.488	1 (.781)	0.799
Type I Error	1	0.790	0.686	1 (.949)	0.898	1	0.679	0.604	0.828	0.753
$d_{dyn} = 0.5$										
Deviation	3	0.853	0.464	1 (.775)	0.784	3	0.951	0.631	1 (.630)	0.875
Type I Error	1	0.893	0.803	1 (.835)	0.983	1	0.798	0.733	0.921	0.855
$d_{dyn} = 2.0$										
Deviation	3	0.868	0.534	1 (.789)	0.804	3	0.961	0.680	1 (.617)	0.893
Type I Error	1	0.908	0.829	1 (.834)	0.985	1	0.825	0.768	0.931	0.874

0 = no true relationship, 1 = weak relationship, 2 = medium relationship, 3 = strong relationship, 4 = perfect relationship.

Weighting the values for R^2 on the one hand and the certainty at the end of the scale on the other hand, two transformations are plausible. The static transformation with $d = 4$ provides a very high certainty at the end of the scale. For example, the chance that the risk for Type I Errors to be less than 5% is 94.9% for nominal data and thus a weak association. The dynamic transformation with $d_{dyn} = 2$ shows a weaker certainty but provides a good compromise between certainty and values of R^2. Table 38 summarizes the cut-off values.

Table 38. Recommendation for Cut-Off Values

	Minimal	Satisfactory	Good	Excellent
Evaluation Category	Expectation of weak practical effect	Expectation of no practical effect	Certainly, only a weak practical effect	Certainly, no practical effect
Iota Index ($d = 4$)	.686	.853	.898	1*
Iota Index ($d_{dyn} = 2$)	.829	.961	.985	1*

Note:
weak practical effect = Deviation less 0.3 and Type I Error less 10%.
no practical effect = Deviation less 0.1 and Type I Error less 5%.
* Limit of the Scale reached

The cut-off values in Table 38 ensure that even under the most demanding conditions, the expected deviation between the true and estimated sample statistics have no or only a small effect on the subsequent analysis. Furthermore, higher values ensure with a high certainty that the quality of the data is in line with small deviations.

Even the correct classification of effect sizes is partly covered by these cut-off values. For example, the value of .961 for the Iota Index ($d_{dyn} = 2$) leads to the expectation that the chance for correctly classifying the effect size is about 74% for nominal data and a medium true association. A value of .985 leads to the expectation of 78%. The following section discusses the results.

8 Discussion

8.1 Conclusions

Content analysis is a popular method in research and a valuable tool in practical settings. The quality of a content analysis is crucial regardless whether it may be in research or in practice as the generated data forms the basis to deduct conclusions and make data-driven decisions. The same quality criteria apply to content analysis, no matter if it is employed in research or in practice (Hesse & Latzko, 2011, p. 70; Ingenkamp & Lissmann, 2008, p. 51). Generally, these are objectivity, reliability and validity.

This book focuses on the criterion of reliability that describes the extent to which an instrument produces error-free data (Schreier, 2012, pp. 166–167). That is the degree to which "a process can be reproduced by different analysts, working under varying conditions, at different locations, or using different but functionally equivalent measuring instruments" (Krippendorff, 2019, p. 281).

In summary, this book provides an updated version of the Iota Reliability Concept which was first described by Berding et al. (2022). The updated concept contributes to the methodology and field of content analysis on the one hand by improving the Iota Concept and on the other hand by adding new opportunities of analyses to the field. To be more specific, the updated concept provides the following progressions.

Applying Maximum Likelihood Estimation

While the first version of the Iota Concept relied on a direct computation from raw data, the new version uses techniques of Maximum Likelihood Estimation, which is inspired by latent class analysis (Andreß et al., 1997). The estimation algorithm uses an EM algorithm and adds a third stage that transforms the Assignment Error Matrix into a structure that is in line with the assumption of weak superiority. The first simulation study showed that this algorithm is able to produce suitable estimates for further applications in practice or research. The estimates in particular are more accurate for coding schemes with a high true reliability than for coding schemes with a low true reliability.

The first simulation study also provides evidence that the new algorithm produces more accurate estimates for an increasing number of raters and an increasing sample size. This is an important difference to Krippendorff's Alpha,

Average Iota and Minimum Iota, which do not perform as well in these cases, since the values are independent from sample size and the number of raters (Berding et al., 2022). The new concept uses the additional information from both an increased number of raters and a larger sample size to produce more accurate results as a consequence.

Providing Insight into the Reliability of Every Single Category.

Previous measures often used in content analysis such as Krippendorff's Alpha, Percentage Agreement, Scott's Pi and Cohen's Kappa (Lovejoy et al., 2016) describe the reliability of a scale with one single numeric value, assuming that the reliability is constant across the entire scale (Feng and Zhao, 2016). The Iota Concept of the first generation is based on the basic ideas of modern test theory (Ayala, 2009; Baker and Kim, 2017; Bonifay, 2020; Paek and Cole, 2020) and overcomes this limitation (Berding et al., 2022). While with the current measures, the different reliabilities in the distinct categories could not be concluded separately, it is now possible to depict the differences between the unique categories through calculating Alpha and Beta Elements as well as Iota.

The new version of the Iota Reliability Concept refines the first generation by redefining the Alpha Elements, Beta Elements, Assignment Error Matrix and by introducing a new definition of Iota. Thus, the updated concept implements its core assumptions with a clearer mathematical framework and enables a more straightforward interpretation of its results. The redefined Iota concept offers more detailed insight into a) the reliability of every single category than the first version and b) how the data generated by a coding scheme actually reflects the underlying true categories.

Providing Insights into the Production of Bias for Different Groups of Individuals/Materials Through a Coding Scheme.

The first version of the Iota Concept provides the possibility to analyze whether or not a coding scheme functions similarly for different groups of participants or materials (Berding et al., 2022, p. 18). That is, it allows to investigate if a coding scheme guides raters similarly for different groups of participants/materials. The redefinition of the Assignment Error Matrix allows a more straightforward way to do this.

Providing Rules of Thumb for Evaluating Content Analysis.

In the study that developed the first generation of the Iota Concept (Berding et al., 2022), cut-off values were derived based on the assumption that the second variable is measured with perfect reliability and that the level of reliability is equal for all categories. Furthermore, only ordinal data was considered. The current study addresses these limitations by varying the degree for reliability between the categories, by allowing different levels of reliability for both independent and dependent variables and by considering nominal and ordinal data. Additionally, the analysis focusses on a range of practically relevant strength of associations/correlations based on Cohen (1988) and uses different target variables to describe the quality of the generated data (deviation between true and estimated sample association/correlation, risk of Type I Errors, chance for correct classification of effect sizes).

The results of the second simulation study show that the predictive power of Average Iota and Minimum Iota is weaker under the condition of varying true reliability (in every category as well as for both independent and dependent variable) as in the study conducted by Berding et al. (2022). In the study of Berding et al. (2022), Average Iota performed similarly to Krippendorff's Alpha. In the current study, Average Iota and Minimum Iota perform better than Percentage Agreement but worse compared to Krippendorff's Alpha.

In contrast, the new Iota Index performs significantly better than Average Iota and Minimum Iota for all kinds of data quality (deviation, Type I Error, correct classification of effect sizes). The Iota Index also performs similarly to Krippendorff's Alpha and in some situations even better. Thus, the new concept is an actual improvement compared to the first generation.

The predictive power of the measures has a direct impact on the cut-off values for practical applications since a higher predictive power is associated with a smaller prediction interval. The second simulation study shows that for each scale value, the expected deviation from the true sample statistics can be predicted and cut-off values can be derived. The resulting cut-off values are higher than reported in the literature. For example, Krippendorff (2019, p. 356) recommends a value for Krippendorff's Alpha of .667 as minimal, of .800 as sufficient and of 1.00 as ideal for judging the quality of codings. These values were replicated and proven in the study by Berding et al. (2022). However, the

present study, which contains more realistic assumptions, implies that these cut-off vales are too low. In contrast, the analysis in study II showed that a value of at least.788 is necessary to justify the expectation of only a weak practical effect on the data. In consequence, a value of at least .963 is necessary to expect no practical effect. The certainty of these expectations is not considered in these values, as a required certainty of 95% implies that the values must be even higher, thus exceeding the possible range of values (the values must be greater 1). This implies that even a perfect value of Krippendorff's Alpha does not guarantee error-free data although the high values suggest an error-free measurement. Thus, these measures should be used with caution.

The like applies to Average Iota and Minimum Iota. The study of Berding et al. (2022, p. 17) recommend a minimum value of .474 for Average Iota and of .377 for Minimum Iota for a deviation of not more than .20. To achieve a deviation not exceeding .10, the recommended value for Average Iota is about .601 and about .478 for Minimum Iota. To achieve a deviation not exceeding .10, the current study suggests a value of at least .847 for Average Iota and of .785 for Minimum Iota. If the effect needs to be 95% certain, the values will exceed the possible range of values. A similar problem occurred for the Iota Index of the updated Iota Concept.

Thus, the currently used cut-off values do not consider that all involved variables vary in their reliability and that the reliability measures themselves are estimates that suffer from estimation errors. As a consequence, the metric of the measures should be constructed in a way that ensures that high values indeed indicate reliability. This would be the case if they guarantee low to no bias in the generated data with a high certainty.

The third simulation study addresses this issue by building in "brakes", or freeing capacities at the end of the scale to account for this degree of certainty. The third simulation study revealed two transformations of the original Iota Index. Both transformations revealed cut-off values that ensure with a certainty of 95% that the generated data suffers only from small practical effects. Generating cut-off values ensuring no practical relevant effects (error-free data) was not successful. Thus, this issue should be addressed in further research. A possible starting point could be the development of more precise estimation methods which may reduce the width of the prediction interval leading to clear cut-off values.

Another starting point could be the adaption of existing measures or the development of new measures of inter-rater reliability.

Additional Opportunities of Analysis

The updated version of the Iota Concept provides additional opportunities for content analysis that are not possible with the first generation. For example, the current measures such as Percentage Agreement, Krippendorff's Alpha and Iota of the first generation provide no answer on how to train and assess a new rater on the basis of an evaluated coding scheme, which would be important as research focuses on the development of new coding schemes (Früh, 2017, p. 185; Krippendorff, 2019, p. 394; Kuckartz, 2018, p. 95; Mayring, 2015, pp. 10–109; Schreier, 2012, pp. 152–165).

Moreover, the updated Iota Concept allows the application of an error correction based on the Assignment Error Matrix, which uses information provided by at least two raters. This improves the quality of the data and contributes to close the gap to latent modeling, ultimately allowing researchers to control for measurement errors (Geiser, 2013, p. 40; Wang & Wang, 2020, p. 1). The next section illustrates these new opportunities.

8.2 Examples for Practical Applications of iotarelr

8.2.1 Overview

Simultaneously to analyzing the new Iota Reliability Concept, an *R* package called *iotarelr* is being developed. The package aims to provide a convenient use of both the old and the new reliability concepts as it implements all calculations presented in this paper and offers an easy way to estimate the different reliability measures. Besides this basic functionality, the package provides additional tools to analyze the quality of a content analysis. These are

- tools to visualize the results similarly to Figure 4,
- tools to analyze if the coding scheme works similarly for different subgroups of materials/participants,
- tools to check the quality of new raters' coding and
- tools to increase the data quality for core studies.

The package is currently available on *github*. A release to *CRAN* is planned as this book is in its final realization stages. The packages' homepage is accessible via

https://fberding.github.io/iotarelr/ that contains the latest version, news and tutorials as well as sample applications for Iota Reliability in *R*. Some examples for the application of the Iota Reliability Concept in practice will be discussed in the following sections.

8.2.2 Checking the Quality of Codings of New Raters

Most scientific studies include the development of a new coding scheme for their study which is typically the focus of literature that introduces content analysis (Krippendorff, 2019; Kuckartz, 2018; Mayring, 2015; Schreier, 2012). In these cases, the reliability can be estimated and analyzed as described in Chapter 3. When shifting to the application of existing coding schemes (from earlier studies or other sources), however, there is less applicable literature. In these cases, the challenge is not to develop a new coding scheme, but to apply an existing one to new data, often along with new raters.

Researchers may choose to apply an existing coding scheme due to several reasons. First, studies using the same coding scheme can be directly compared, ultimately contributing to knowledge accumulation regarding a specific topic or discipline. Second, the application of an existing coding scheme provides the opportunity to prove results of prior studies by trying to reproduce them. Third, using an existing scheme saves resources, since ideally, the improvement cycle for developing (Früh, 2017, p. 185; Krippendorff, 2019, p. 394; Kuckartz, 2018, p. 95; Mayring, 2015, pp. 10–109; Schreier, 2012, pp. 152–165) is not necessary. This saves capacities which become available for other aspects in the course of the study (greater sample sizes, refining specific categories, considering more categories etc.).

Estimating the reliability of codings based on an existing coding scheme differs from developing an own coding scheme. To be specific, the development phase aims to provide a theoretically and empirically sound guide for data analysis. For that it is essential that researchers and raters steer clear of their own interpretations of categories and data and that they develop the same understanding of the categories based on theoretical and empirical evidence. Ideally, the final coding scheme is precise enough to document this shared understanding and to guide users to the same interpretation of data and the same assignments of data to categories. This shared understanding is also the basic idea behind the inter-rater reliability discussed in the literature (Mayring,

2014). Corresponding measures of inter-rater reliability try to quantify this degree of shared interpretation. The goal is that the coding scheme at least allows other people to understand how the results in a study are generated. In the best case, the coding scheme allows a replication of the study's results.

In contrast, when new raters apply an existing coding scheme, it should be avoided to incorporate the new raters' personal interpretations of data/categories. The aim is to train new raters so they understand the data/categories that are already documented in the scheme. This is important as the existing coding scheme already represents a discussed and validated understanding that new raters have to acquire in order to apply the coding scheme in the same way as in its development study or any other preliminary study.

As a consequence, reliability estimation has to consider the existence of a predefined understanding that cannot be adapted. Within the framework of the Iota Reliability Concept, this is realized as shown in Figure 22. In order to estimate how well a new rater's understanding of the items documented in a coding scheme is developed, data and material from corresponding sources must be used. The new rater assigns the material to categories and the assignments are compared with the existing assignments of the material. Based on this data, the Assignment Error Matrix for the new rater can be calculated as shown in Figure 22.

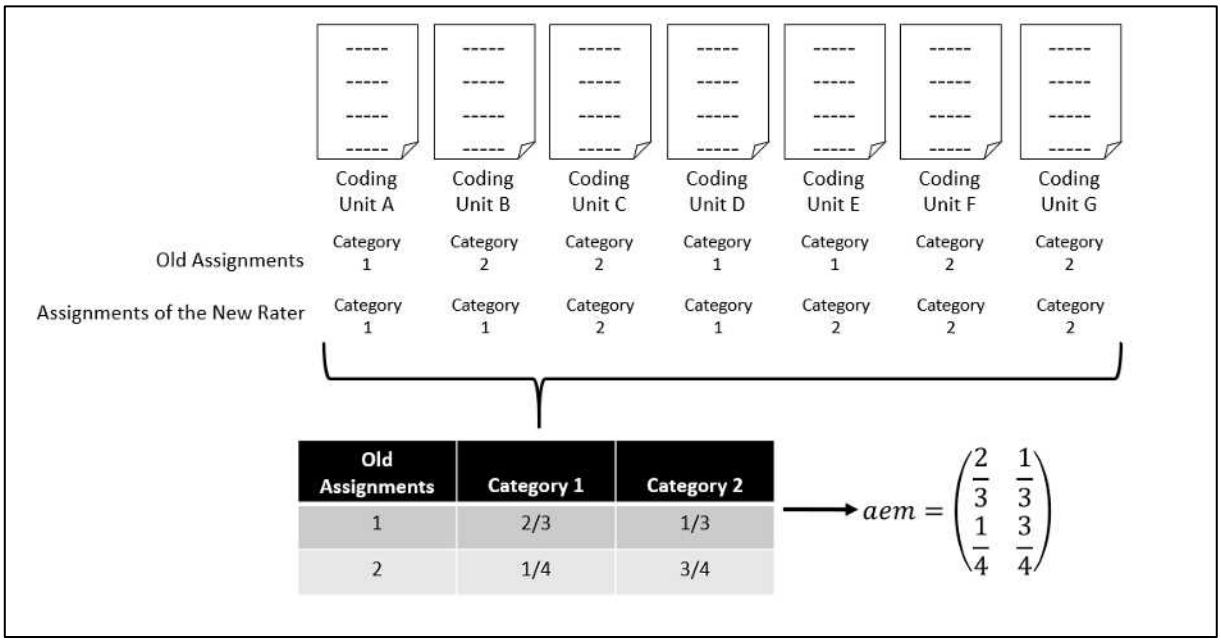


Figure 22. Example for Assessing the Reliability of an Existing Coding Scheme

In our example, the material is clustered into groups, one for each category of the scheme. The material is assigned to a group based on the *old* assignments. The next step is to count the frequency of all categories within each group that were assigned by the new rater and then divide them by the number of cases within each group (sum of the rows). The result is a table reporting the relative frequency for the assignment of a coding unit belonging to category i in preliminary sources to category j by the new rater. This table forms the input for the algorithm described in the conditioning stage of section 4. The result is the Assignment Error Matrix for that specific rater.

More specifically, in Figure 22, three documents are assigned to category 1 and four are assigned to category 2 in a *preliminary* study. Of the three documents belonging to category 1, the new rater assigned two to category 1 and one to category 2. Of the four documents truly belonging to category 2, the new rater assigned one document to category 1 and three documents to category 2. The resulting table can be read as follows: If the coding unit belongs to category 1 in the preliminary study, the new rater assigned the coding unit in 66% of cases to category 1 and in 33% to category 2. If the coding unit belongs to category 2 in the preliminary study, the new rater assigns the coding unit in 25% of cases to category 1 and in 75% of cases to category 2.

With the help of the algorithm in the conditioning stage, the Assignment Error Matrix can be estimated, producing a result that is in line with the assumption of weak superiority. In the example above, this matrix equals the relative frequencies of that table.

To calculate and analyze the same reliability measures as in the development phase of the coding scheme, the estimated Assignment Error Matrix should be used *together* with the categorical sizes out of the *existing* sources. Assuming that the size of category 1 is 27% and of category 2 it is 73%, the Iota Index is about .455, which is quite low according to the cut-off values derived in section 6.3. Thus, the new rater has to analyze the differences between the old and their own assignments and/or has to discuss their results with other users/developers of the coding scheme in order to adapt their coding actions. Figure 23 shows the Iota values for a more detailed inspection.

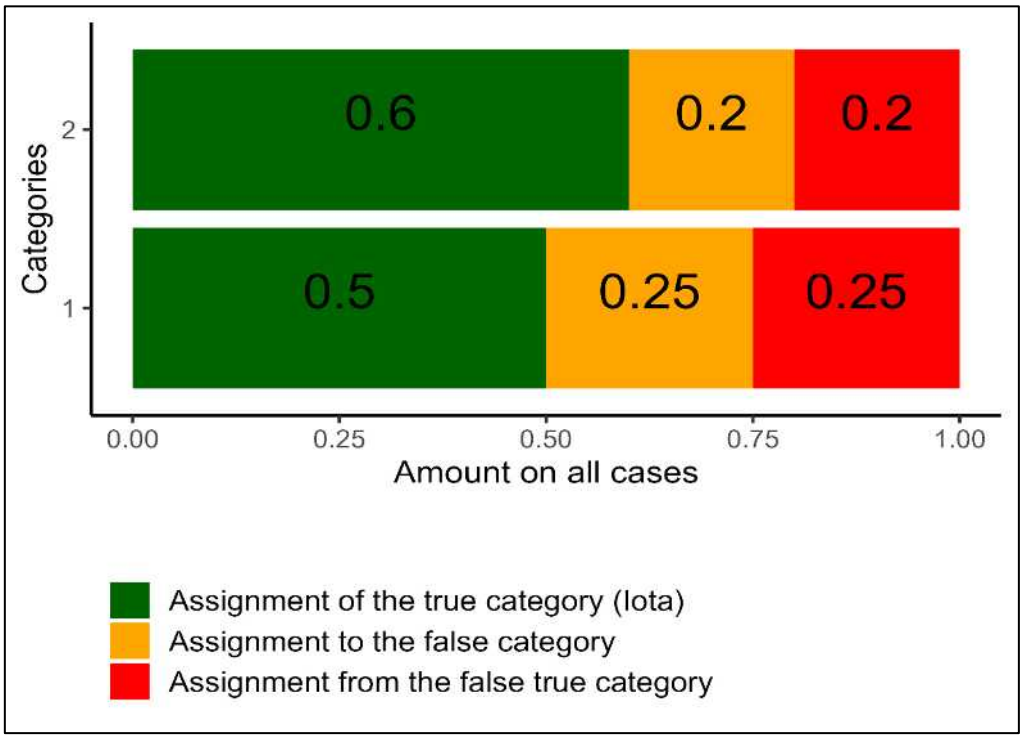


Figure 23. Example of an Illustration of Iota for a New Rater

The data generated for category 1 by the new rater is made up of about 50% of the correct coding units (green part of the bar). In 25% of the relevant data are missing coding units of category 1 (yellow part of the bar). Instead, about 25% of the data of category 1 consists of material truly belonging to category 2 (red part of the bar). A similar but slightly better result occurs for the data of category 2. Thus, the new rater needs further instructions to distinguish both categories.

Another practical situation for using this kind of reliability estimation is content analysis conducted via supervised machine learning, which typically involves a kind of artificial intelligence. Supervised machine learning means that a machine learns to transform input data into output data (Lanquillon, 2019, pp. 96–97). The machine generates a model characterizing the relationship between both kinds of data and uses this model to transform new input data into output data. During the learning process, the model is optimized for reproducing the pairs of input and output data as good as possible. Thus, the structure is the same as for human raters. An artificial intelligence has to learn from existing materials (input data) and existing assignments of the materials to categories (output data) in order to assign new material correctly. In consequence, all reliability measures proposed by the Iota Concept can be applied to this kind of content analysis in a similar way.

Using the package *iotarelr,* the corresponding function is $check\_new\_rater()$ which can be used for both human raters and artificial intelligence. The corresponding reliability measures are requested with $get\_iota2\_measures()$ and the corresponding figure is created via $plot\_iota()$.

8.2.3 Checking for Bias and Different Guidance of a Coding Scheme

The Iota Concept is inspired by item response theory, which allows to analyze to what degree items of a questionnaire or a test function similarly within different groups (Baker & Kim, 2017, pp. 38–42). Subgroup invariance is not only important for questionnaires and classical achievement/personality tests but also for content analysis. This is due to the following reasons. In the context of scientific studies, subgroup invariance results from the need of a valid measurement. If a coding scheme functions differently for specific subgroups, the generated data does not reflect the phenomena but is confounded with other constructs.

In the context of data driven decision making, an absence of subgroup invariance can lead to actions preferring or discouraging specific groups, ultimately producing wrong conclusions. For example, Seufert et al. (2021, p. 15) worked out that the use of artificial intelligence in educational settings can reproduce bias present in the training data. Similar challenges are reported by Luan et al. (2020, p. 5). Subgroup invariance is thus not only important for content analysis done by an artificial intelligence, but also for a content analysis conducted by humans. For example, if written essays are used to make judgements regarding a student's qualification (e.g., grades), the absence of subgroup invariance could imply that students with a specific gender or a specific social background earn better results than students of other groups, although this is not justified with their reported performance.

In analogy to item response theory, the Iota Reliability Concept allows for an analysis of such a bias. However, it must be known to which group the coding units belong. The idea behind the analysis is illustrated in Figure 24. The complete data set is divided into a separate data set for each group, comprising only the data for that specific group. For each group, the elements of the Iota Concept are estimated separately. If the coding scheme functions similarly for different groups, similar values for the elements should occur.

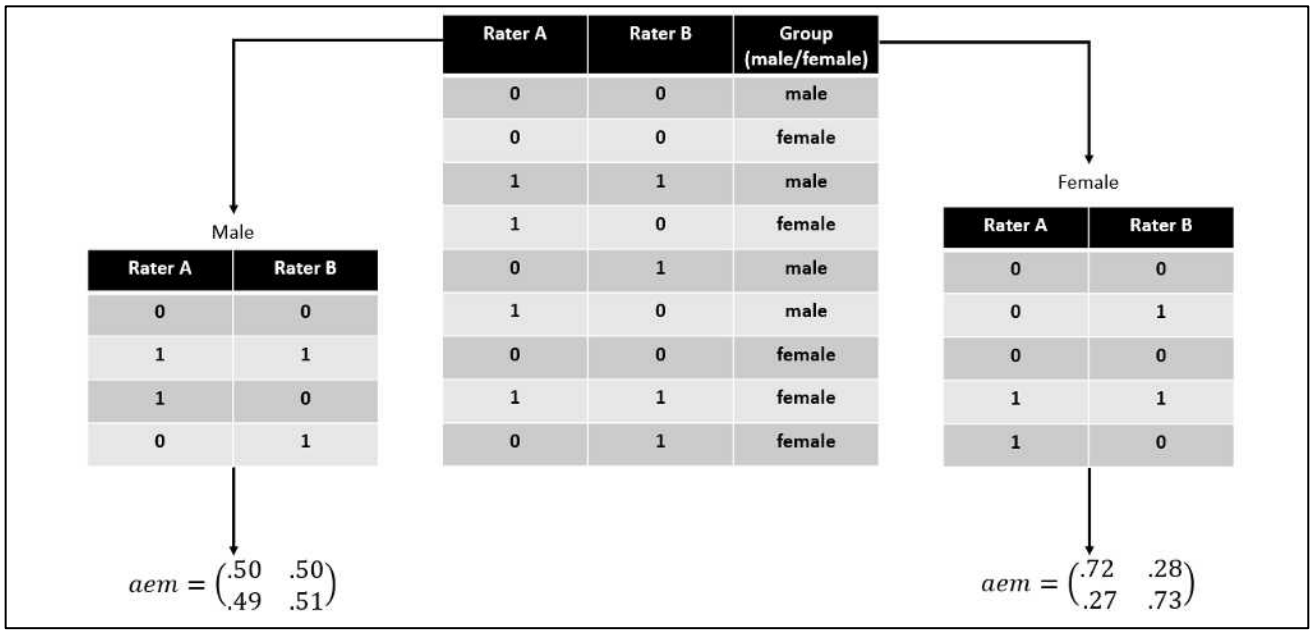


Figure 24. Example for an Analysis of Subgroup Invariance

Figure 24 illustrate the principle with an example. It is assumed that an essay has to be rated by two raters on the basis of a coding scheme with "0", indicating that the essay failed and "1", indicating that the essay passed the exam. It is further assumed that the coding should work similarly for males and females.

To analyze the subgroup invariance for sex, the complete data set is divided into two sets containing the ratings for both males and females respectively. For each group, the elements of the Iota Concept are estimated. Figure 24 presents the corresponding Assignment Error Matrices. For males, the coding shows the pattern of a random assignment since all cells in the Assignment Error Matrix are nearly .50. That is, for males the decision for failing or passing the essay equals random guessing. For females, the coding is more reliable. If the essay is truly not good, raters assign failed ("0") in about 72% of cases and passed ("1") in only 28%. If the essay is truly good enough, the raters assign passed ("1") in about 73% of cases and failed ("0") in only 27% of cases.

In consequence, the coding scheme for the essays lead to unreliable data for males. For females, the data is significantly more reliable but still requires some improvements. The ratings do not reflect the data for males correctly, leading to an increased risk of biased conclusions about the achievements of males compared to females.

Using the package *iotarelr,* this kind of analysis can be requested with $check\_dgf()$.

8.2.4 Improving the Quality of Codings

A third application case of the Iota Reliability Concept is the improvement of data quality by using all available information that can be generated through the package. This requires at least two raters. The principle is illustrated in Figure 25.

Rater A	Rater B
0	1
2	2
1	2
0	0
1	1
2	2
0	0
1	1
2	1

$$aem = \begin{pmatrix} .60 & .40 & .00 \\ .25 & .50 & .25 \\ .20 & .20 & .60 \end{pmatrix}$$

True Category 0	True Category 1	True Category 2	Expected Category
.593	.309	.099	0
.000	.148	.852	2
.000	.510	.490	1
.778	.135	.086	0
.356	.556	.089	1
.000	.148	.852	2
.778	.135	.086	0
.356	.556	.089	1
.000	.510	.490	1

Figure 25. Example for Correcting Errors in Codings

The Assignment Error Matrix allows the calculation of the probability of the true category i under the condition of the observed coding pattern. The true category with the highest probability is the category with the highest likelihood to produce the observed pattern of assignments. Thus, this is the most plausible category.

In the example from Figure 25, the assigned categories of two raters and the Assignment Error Matrix lead to the presented probabilities. If the pattern is 0-1, the true category is 0 in about 59.3% of cases, category 1 in about 30.9% of the cases and category 2 in about 9.9% of cases. Thus, it is mostly reasonable to assume that the true category of the first coding unit is category 0.

The two raters assigned the pattern 1-2 to the third coding unit. This pattern implies that the chance of the true category 0 is about 0%, of category 1 51.0% and of category 2 49.0%. Thus, it is most plausible to assign category 1 to that coding unit. However, this example shows that the certainty for this decision is quite low. In practice, this indicates a closer inspection of that coding unit. Involving more raters can help to distinguish the true categories with more clarity.

Using the package *iotarelr,* this kind of analysis can be requested with $est\_expected\_categories()$.

8.3 Limitations and Further Directions

This study is not without limitations. First, as shown in the first simulation study, the accuracy of the estimates is less precise for a low true reliability than for a high true reliability. Thus, the estimates can lead to the wrong conclusions. Further research should try to establish more accurate algorithms for parameter estimations.

Second, currently there are neither significance tests nor effect sizes established that are suitable to describe the differences between two Assignment Error Matrices. These tests and effect sizes would be helpful for judging whether or not a new rater has acquired the correct understanding of an existing coding scheme. Alike applies for the analysis of the Assignment Error Matrices for different groups of participants/materials. In these cases, significant tests and effect sizes can help to judge if the differences between groups are relevant for practice or not.

Third, the cut-off values are based only on Cramer's V and Kendall's Tau. For other applications (e.g. Pearson Correlation, Analysis of Variance, t-Test, Kruskal Wallis Test, etc.), other cut-off values may be more appropriate.

References

Afifi, A., May, S., Donatello, R. A., & Clark, V. A. (2020). *Practical Multivariate Analysis* (6th ed.). *Chapman and Hall/CRC Texts in Statistical Science Ser*. CRC Press LLC. https://ebookcentral.proquest.com/lib/kxp/detail.action?docID=5945612

Agresti, A. (2022). *Foundations of Statistics for Data Scientists: With R and Python*. *Chapman and Hall/CRC Texts in Statistical Science Ser*. CRC Press LLC. https://ebookcentral.proquest.com/lib/kxp/detail.action?docID=6809922

Andreß, H.-J., Hagenaars, J. A., & Kühnel, S. (1997). *Analyse von Tabellen und kategorialen Daten: Log-lineare Modelle, latente Klassenanalyse, logistische Regression und GSK-Ansatz*. *Springer-Lehrbuch*. Springer.

Ayala, R. J. de. (2009). *The theory and practice of item response theory*. *Methodology in the social sciences*. The Guilford Press.

Baker, F. B., & Kim, S.-H. (2017). *The Basics of Item Response Theory Using R*. Springer International Publishing. https://doi.org/10.1007/978-3-319-54205-8

Berding, F., Riebenbauer, E., Stütz, S., Jahncke, H., Slopinski, A., & Rebmann, K. (2022). Performance and Configuration of Artificial Intelligence in Educational Settings.: Introducing a New Reliability Concept Based on Content Analysis. *Frontiers in Education*, 1–21. https://doi.org/10.3389/feduc.2022.818365

Blair, G., Cooper, J., Coppock, A., Humphreys, M., & Sonnet, L. (2022). *estimatr: Fast Estimators* [Computer software]. https://CRAN.R-project.org/package=estimatr

Cain, M. K., & Zhang, Z. (2019). Fit for a Bayesian: An Evaluation of PPP and DIC for Structural Equation Modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *26*(1), 39–50. https://doi.org/10.1080/10705511.2018.1490648

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd Ed.). Taylor & Francis.

Feng, G. C., & Zhao, X. (2016). Do Not Force Agreement. *Methodology*, *12*(4), 145–148. https://doi.org/10.1027/1614-2241/a000120

Früh, W. (2017). *Inhaltsanalyse: Theorie und Praxis* (9., überarbeitete Auflage). *UTB Medien- und Kommunikationswissenschaft, Psychologie, Soziologie: Vol. 2501*. UVK Verlagsgesellschaft mbH. http://www.blickinsbuch.de/item/7639ebd5a2b1dab9fcae73f36314f7f8

Gamer, M., Lemon, J., & Fellows Puspendra Singh, I. (2019). *irr: Various Coefficients of Interrater Reliability and Agreement* [Computer software]. https://CRAN.R-project.org/package=irr

Geiser, C. (2013). *Data analysis with Mplus*. Guilford Press.

Hallquist, M. N., & Wiley, J. F. (2018). Mplusautomation: An R Package for Facilitating Large-Scale Latent Variable Analyses in Mplus. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*(4), 621–638. https://doi.org/10.1080/10705511.2017.1402334

Hayes, A. F. (2005). *Statistical methods for communication science*. *LEA'S communication series*. Lawrence Erlbaum. http://www.loc.gov/catdir/enhancements/fy0668/2005040570-d.html

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/10.1080/19312450709336664

Heck, R. H., & Thomas, S. L. (2020). *An introduction to multilevel modeling techniques: Mlm and SEM approaches* (Fourth edition). *Quantitative methodology series*. Routledge.

Hesse, I., & Latzko, B. (2011). *Diagnostik für Lehrkräfte* (2. Auflage). Budrich.

Hoofs, H., van de Schoot, R., Jansen, N. W. H., & Kant, I. (2018). Evaluating Model Fit in Bayesian Confirmatory Factor Analysis With Large Samples: Simulation Study Introducing the BRMSEA. *Educational and Psychological Measurement*, *78*(4), 537–568. https://doi.org/10.1177/0013164417709314

Hove, D. ten, Jorgensen, T. D., & van der Ark, L. A. (2018). On the Usefulness of Interrater Reliability Coefficients. In M. Wiberg, S. Culpepper, R. Janssen, J. González, & D. Molenaar (Eds.), *Springer Proceedings in Mathematics & Statistics. Quantitative Psychology* (Vol. 233, pp. 67–75). Springer International Publishing. https://doi.org/10.1007/978-3-319-77249-3_6

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Hußmann, S., Leuders, T., & Prediger, S. (2007). Schülerleistungen verstehen - Diagnose im Alltag. *Praxis Der Mathematik in Der Schule*, *49*(15), 1–8.

Ingenkamp, K., & Lissmann, U. (2008). *Lehrbuch der pädagogischen Diagnostik* (6. Auflage). Beltz.

Kleesiek, J., Murray, J. M., Strack, C., Kaissis, G., & Braren, R. (2020). Wie funktioniert maschinelles Lernen? *Der Radiologe*, *60*, 24–31. https://doi.org/10.1007/s00117-019-00616-x

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

Krippendorff, K. (2019). *Content Analysis: An Introduction to Its Methodology* (4th Ed.). SAGE.

Kuckartz, U. (2018). *Qualitative Inhaltsanalyse: Methoden, Praxis, Computerunterstützung* (4. Auflage). *Grundlagentexte Methoden*. Beltz Juventa. http://www.beltz.de/de/nc/verlagsgruppe-beltz/gesamtprogramm.html?isbn=978-3-7799-3682-4

Lanquillon, C. (2019). Grundzüge des maschinellen Lernens. In S. Schacht & C. Lanquillon (Eds.), *Blockchain und maschinelles Lernen: Wie das maschinelle Lernen und die Distributed-Ledger-Technologie voneinander profitieren* (pp. 89–142). Springer.

Larusson, J. A., & White, B. (2014). Introduction. In J. A. Larusson & B. White (Eds.), *Learning Analytics: From Research to Practice* (pp. 1–12). Springer New York. https://doi.org/10.1007/978-1-4614-3305-7_1

Leuders, T. (2010). Kompetenzorientierte Aufgaben im Unterricht. In W. Blum (Ed.), *Bildungsstandards Mathematik: konkret: Sekundarstufe I: Aufgabenbeispiele, Unterrichtsanregungen, Fortbildungsideen.* (1st ed., pp. 81–95). Cornelsen Verlag Scriptor.

Lovejoy, J., Watson, B. R., Lacy, S., & Riffe, D. (2016). Three Decades of Reliability in Communication Content Analyses. *Journalism & Mass Communication Quarterly*, *93*(4), 1135–1159. https://doi.org/10.1177/1077699016644558

Luan, H., Geczy, P., Lai, H., Gobert, J., Yang, S. J. H., Ogata, H., Baltes, J., Guerra, R., Li, P., & Tsai, C.-C. (2020). Challenges and Future Directions of Big Data and Artificial Intelligence in Education. *Frontiers in Psychology*, *11*, 1–11. https://doi.org/10.3389/fpsyg.2020.580820

Lüdecke, D. (2018). *Sjstats: Statistical Functions For Regression Models* [Computer software]. Zenodo.

Mayring, P. (2014). *Qualitative content analysis: theoretical foundation, basic procedures and software solution*. https://nbn-resolving.org/urn:nbn:de:0168-ssoar-395173

Mayring, P. (2015). *Qualitative Inhaltsanalyse: Grundlagen und Techniken* (12., überarbeitete Auflage). Beltz.

Muthén Linda, K., & Muthén, B. O. (2022). *Mplus* (Version 8.8) [Computer software].

R Core Team. (2021). *R: A language and environment for statistical computing [*Computer software]. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

Rasch, B., Friese, M., Hofmann, W. J., & Naumann, E. (2010). *Quantitative Methoden 1: Einführung in die Statistik für Psychologen und Sozialwissenschaftler* (3. Auflage). Springer.

Schreier, M. (2012). *Qualitative Content Analysis in Practice*. SAGE.

Seufert, S., Guggemos, J., & Ifenthaler, D. (2021). Zukunft der Arbeit mit intelligenten Maschinen: Implikationen der Künstlichen Intelligenz für die Berufsbildung. In S. Seufert, J. Guggemos, D. Ifenthaler, H. Ertl, & J. Seifried (Eds.), *Zeitschrift für Berufs- und Wirtschaftspädagogik Beiheft: Vol. 31. Künstliche Intelligenz in der beruflichen Bildung: Zukunft der Arbeit und Bildung mit intelligenten Maschinen?!* (pp. 9–27). Franz Steiner Verlag.

Sjuts, J. (2007). Kompetenzdiagnostik im Lernprozess - auf theoriegeleitete Aufgabengestaltung und -auswertung kommt es an. *Mathematica Didactica*, *30*(2), 33–52.

Sjuts, J. (2010). Unterrichtliche Gestaltung und Nutzung kompetenzorientierter Aufgaben in diagnostischer Hinsicht. In W. Blum (Ed.), *Bildungsstandards Mathematik: konkret: Sekundarstufe I: Aufgabenbeispiele, Unterrichtsanregungen, Fortbildungsideen.* (1st ed., pp. 96–112). Cornelsen Verlag Scriptor.

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

Wang, J., & Wang, X. (2020). *Structural equation modeling: Applications using Mplus* (2nd Ed.). *Wiley series in probability and statistics*. Wiley.

Zhao, X., Feng, G. C., Liu, J. S., & Deng, K. (2018). We agreed to measur o measure agreement - Redefining r eement - Redefining reliability de-justifies eliability de-justifies Krippendorff's alpha. *China Media Research*, *14*(2), 1–15.

Zhao, X., Liu, J. S., & Deng, K. (2013). Assumptions behind Intercoder Reliability Indices. *Annals of the International Communication Association*, *36*(1), 419–480. https://doi.org/10.1080/23808985.2013.11679142

Zyphur, M. J., & Oswald, F. L. (2015). Bayesian Estimation and Inference. *Journal of Management*, *41*(2), 390–420. https://doi.org/10.1177/0149206313501200

Appendix A –Confidence Intervals

Primary Parameters

	True Iota Index	Median	CI75	CI90	CI95	CI99
1	0	0.0541	0.2821	0.5034	0.6758	0.886
2	0.05	0.0541	0.2821	0.5034	0.6758	0.886
3	0.1	0.0657	0.2067	0.441	0.5573	0.7822
4	0.15	0.0664	0.1619	0.3293	0.4491	0.6428
5	0.2	0.0635	0.13	0.2283	0.3123	0.4982
6	0.25	0.0608	0.1204	0.1981	0.2544	0.3994
7	0.3	0.0588	0.1181	0.1922	0.2423	0.3564
8	0.35	0.0556	0.1153	0.1911	0.2421	0.348
9	0.4	0.0505	0.1107	0.1915	0.2462	0.3571
10	0.45	0.0445	0.1032	0.1879	0.2476	0.3651
11	0.5	0.0354	0.0872	0.1716	0.2355	0.3592
12	0.55	0.03	0.0775	0.1611	0.2251	0.3507
13	0.6	0.0234	0.0588	0.1288	0.1883	0.3203
14	0.65	0.0205	0.0543	0.1248	0.1852	0.3049
15	0.7	0.0183	0.0494	0.1112	0.1619	0.2817
16	0.75	0.0153	0.0422	0.1006	0.1513	0.2568
17	0.8	0.0143	0.0386	0.0929	0.1444	0.2743
18	0.85	0.0117	0.0304	0.0751	0.1167	0.2178
19	0.9	0.0118	0.0327	0.0694	0.104	0.2473
20	0.95	0.0099	0.0254	0.0496	0.0756	0.1681
21	1	0.0064	0.0165	0.0329	0.0598	0.1804

Alpha Reliability

	True Iota Index	Median	CI75	CI90	CI95	CI99
1	0	0.0236	0.0815	0.2696	0.4633	0.4973
2	0.05	0.0236	0.0815	0.2696	0.4633	0.4973
3	0.1	0.0374	0.1023	0.2388	0.3806	0.4892
4	0.15	0.0491	0.1065	0.2185	0.3256	0.478
5	0.2	0.0511	0.1008	0.1774	0.2506	0.4361
6	0.25	0.0502	0.0965	0.1633	0.2198	0.3813
7	0.3	0.0494	0.094	0.1553	0.2065	0.3497
8	0.35	0.0474	0.0921	0.1517	0.1993	0.3336
9	0.4	0.0449	0.0906	0.1522	0.2018	0.337
10	0.45	0.041	0.0867	0.1496	0.2007	0.3398
11	0.5	0.0339	0.0773	0.144	0.1974	0.335
12	0.55	0.028	0.0693	0.1387	0.1943	0.329
13	0.6	0.0217	0.0544	0.1166	0.1691	0.3155
14	0.65	0.0184	0.0493	0.1153	0.1683	0.3049
15	0.7	0.0161	0.0451	0.1038	0.1609	0.2994
16	0.75	0.0133	0.0398	0.1009	0.1592	0.2752
17	0.8	0.0126	0.039	0.1022	0.1556	0.3121
18	0.85	0.0091	0.0271	0.0748	0.1278	0.2334
19	0.9	0.0099	0.0327	0.0767	0.1274	0.2811
20	0.95	0.007	0.0223	0.0491	0.0833	0.2046
21	1	0.0041	0.0139	0.0338	0.0783	0.1988

Beta Reliability

	True Iota Index	Median	CI75	CI90	CI95	CI99
1	0	0	0	0	0	0
2	0.05	0	0	0	0	0
3	0.1	0	0.0011	0.0467	0.0856	0.2429
4	0.15	0.0043	0.0419	0.0925	0.1388	0.2774
5	0.2	0.0265	0.0604	0.1072	0.1466	0.264
6	0.25	0.0327	0.0651	0.1093	0.1464	0.2509
7	0.3	0.0352	0.0689	0.1141	0.1508	0.2496
8	0.35	0.0364	0.0731	0.1224	0.1617	0.2682
9	0.4	0.0356	0.0763	0.1332	0.1788	0.2968
10	0.45	0.0314	0.0758	0.1422	0.1963	0.3435
11	0.5	0.0218	0.0648	0.1372	0.1999	0.4036
12	0.55	0.0054	0.0437	0.1118	0.1758	0.3671
13	0.6	0	0.0176	0.0765	0.1358	0.3104
14	0.65	0	6.00E-04	0.0493	0.1416	1
15	0.7	0	0	0.0379	0.1062	1
16	0.75	0	0	0.0358	0.1219	1
17	0.8	0	0	0.0072	0.0683	1
18	0.85	0	0	0.0471	1	1
19	0.9	0	0	0	1	1
20	0.95	0	0	0	1	1
21	1	0	0	1	1	1

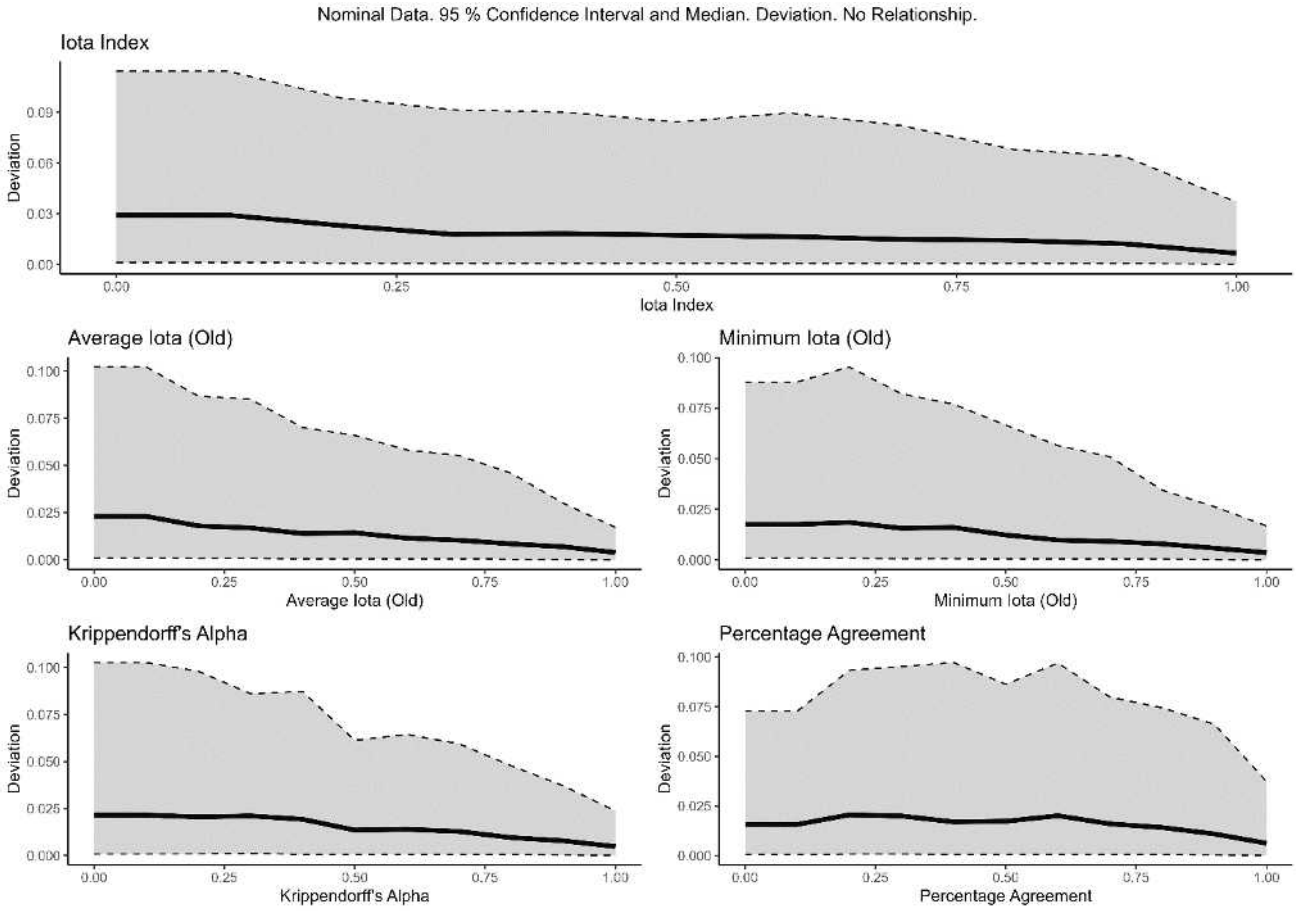
Iota

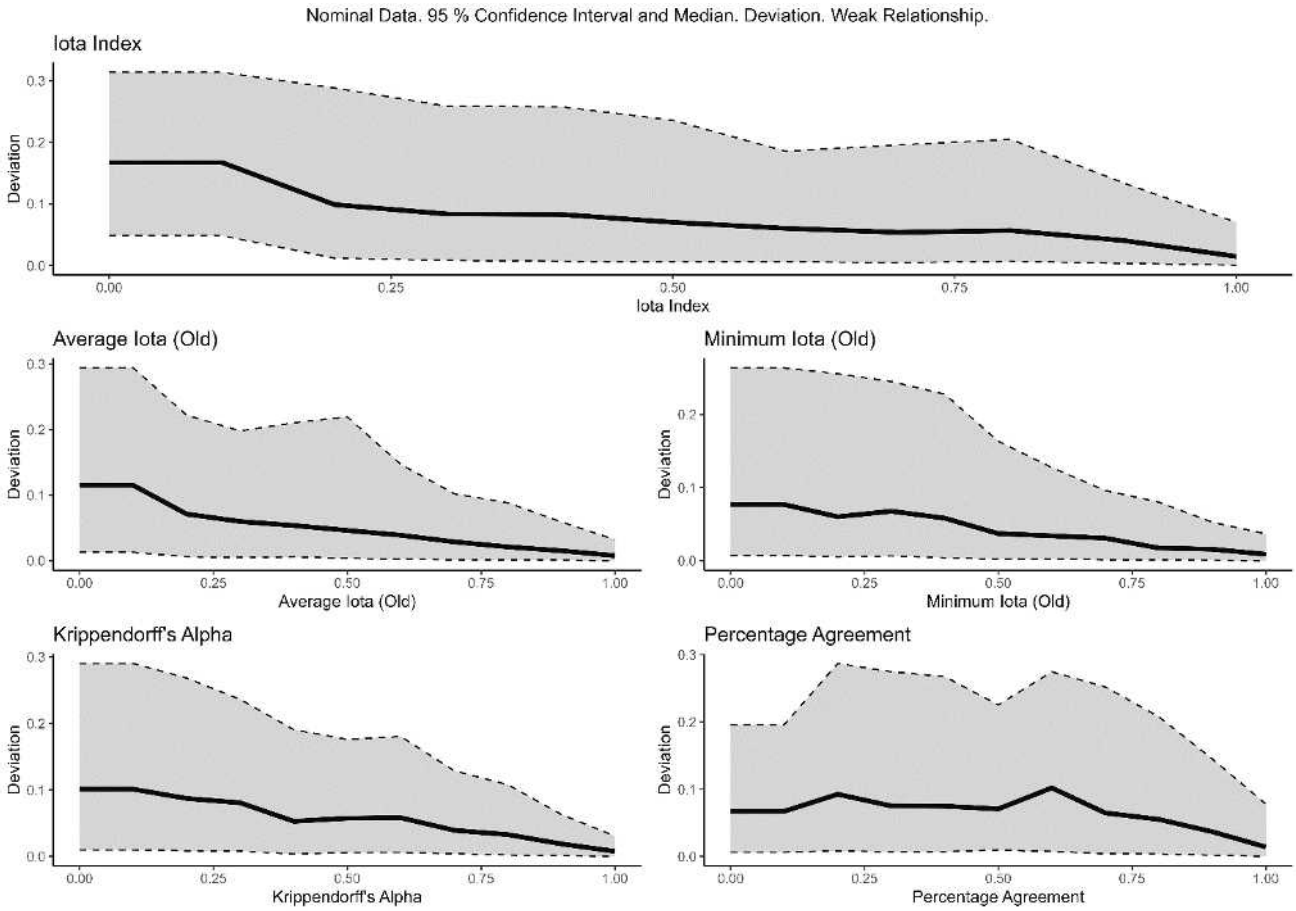
	True Iota Index	Median	CI75	CI90	CI95	CI99
1	0	0.1562	0.2837	0.3825	0.4278	0.468
2	0.05	0.1562	0.2837	0.3825	0.4278	0.468
3	0.1	0.115	0.218	0.3203	0.3757	0.4476
4	0.15	0.0908	0.1679	0.2545	0.311	0.4029
5	0.2	0.0692	0.1267	0.1938	0.2417	0.3484
6	0.25	0.0599	0.1093	0.1663	0.2061	0.2925
7	0.3	0.0563	0.105	0.1612	0.1994	0.28
8	0.35	0.0532	0.1028	0.1622	0.2019	0.2832
9	0.4	0.0483	0.0993	0.1626	0.2056	0.2939
10	0.45	0.0423	0.0922	0.1583	0.2046	0.3035
11	0.5	0.0319	0.0748	0.1417	0.1911	0.3033
12	0.55	0.0255	0.0613	0.118	0.1633	0.2777
13	0.6	0.0196	0.0437	0.0892	0.1308	0.2573
14	0.65	0.0174	0.0388	0.0782	0.1135	0.2184
15	0.7	0.015	0.0333	0.067	0.098	0.1831
16	0.75	0.0137	0.0293	0.0545	0.0736	0.1474
17	0.8	0.0124	0.0279	0.0549	0.0805	0.146
18	0.85	0.0107	0.0221	0.0406	0.0606	0.1283
19	0.9	0.0099	0.0218	0.0436	0.0643	0.131
20	0.95	0.0072	0.0155	0.0305	0.0481	0.1125
21	1	0.0045	0.0116	0.0316	0.0667	0.1388

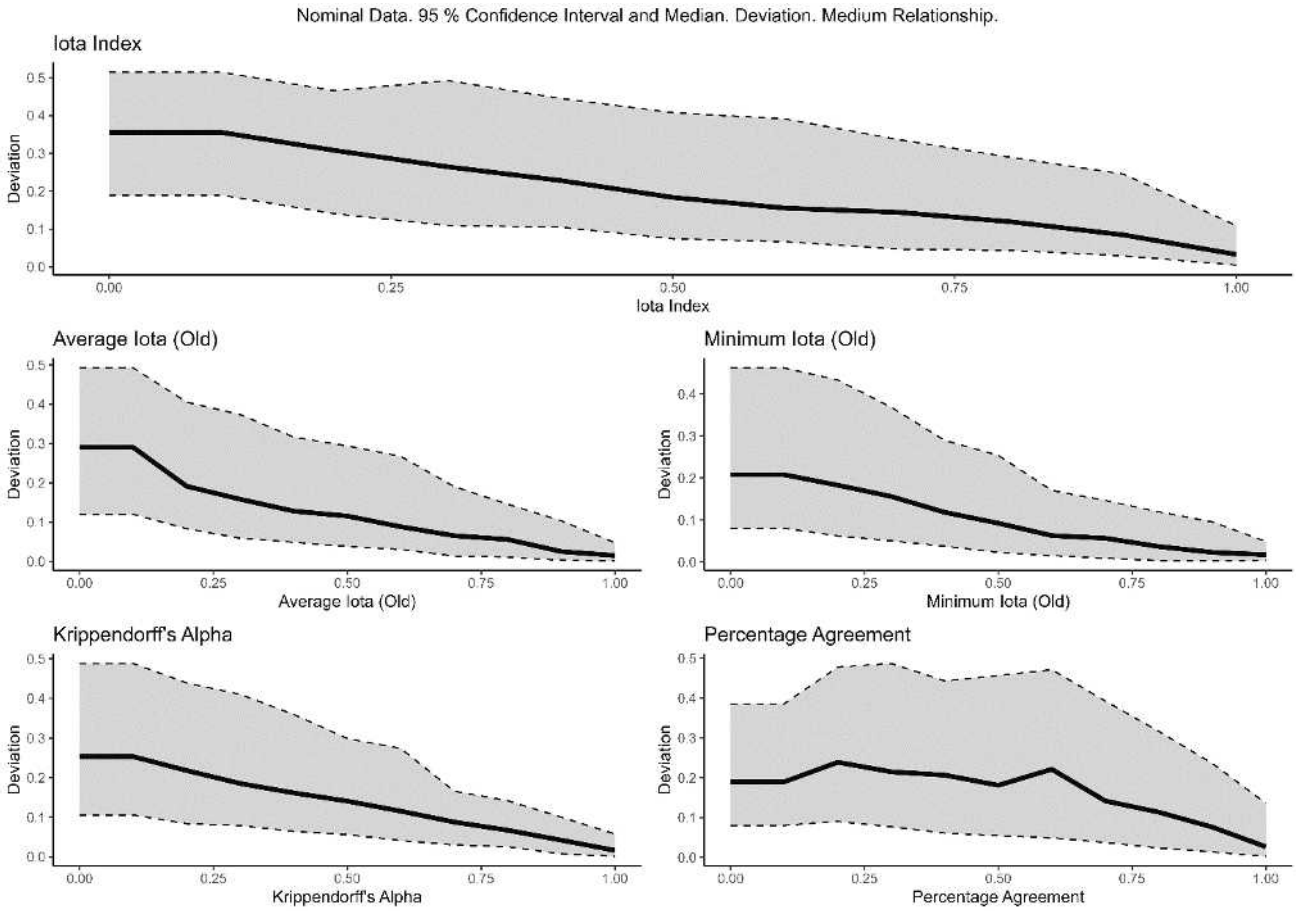
Iota Index

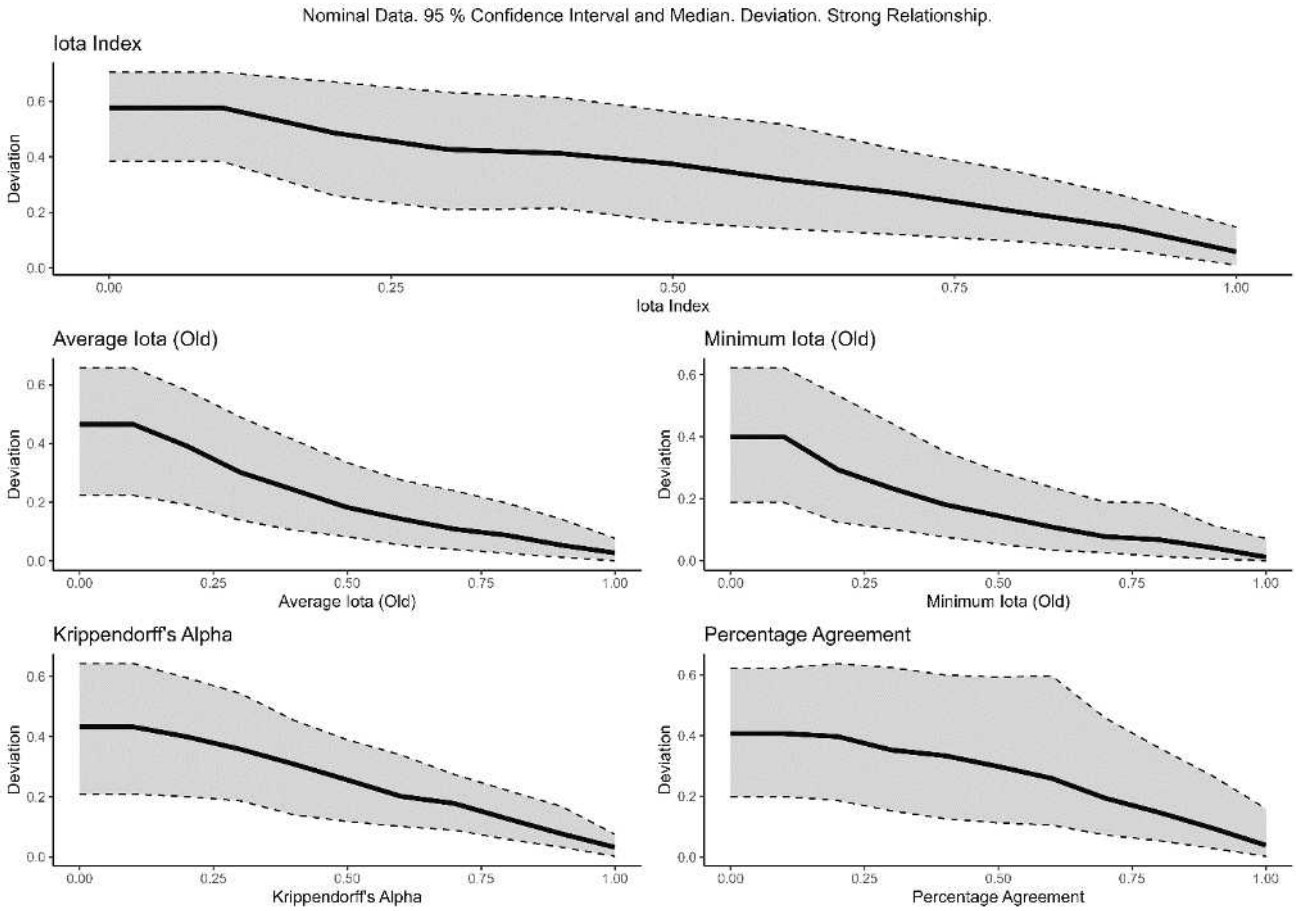
	True Iota Index	Median	CI75	CI90	CI95	CI99
1	0	0.0235	0.0561	0.126	0.1783	0.2976
2	0.05	0.0235	0.0561	0.126	0.1783	0.2976
3	0.1	0.0337	0.0618	0.113	0.1628	0.2663
4	0.15	0.0434	0.0768	0.114	0.1474	0.2499
5	0.2	0.0402	0.0738	0.116	0.1483	0.2386
6	0.25	0.0336	0.0631	0.1022	0.1324	0.2018
7	0.3	0.0283	0.0534	0.0872	0.1143	0.1793
8	0.35	0.0247	0.0474	0.0782	0.1038	0.1685
9	0.4	0.0228	0.0445	0.0751	0.1012	0.1643
10	0.45	0.0219	0.0445	0.0782	0.1056	0.1758
11	0.5	0.0193	0.0413	0.0772	0.1093	0.1812
12	0.55	0.0203	0.0425	0.0802	0.1129	0.176
13	0.6	0.0176	0.0358	0.0669	0.0926	0.1526
14	0.65	0.0172	0.0354	0.0676	0.0949	0.1514
15	0.7	0.016	0.0318	0.0584	0.0835	0.1326
16	0.75	0.0147	0.0287	0.0513	0.0707	0.1102
17	0.8	0.0123	0.025	0.0445	0.0631	0.1014
18	0.85	0.011	0.0212	0.0386	0.0515	0.0867
19	0.9	0.0091	0.0185	0.0334	0.0439	0.0784
20	0.95	0.0066	0.0131	0.0213	0.0278	0.0483
21	1	0.0037	0.0068	0.0133	0.0193	0.0333

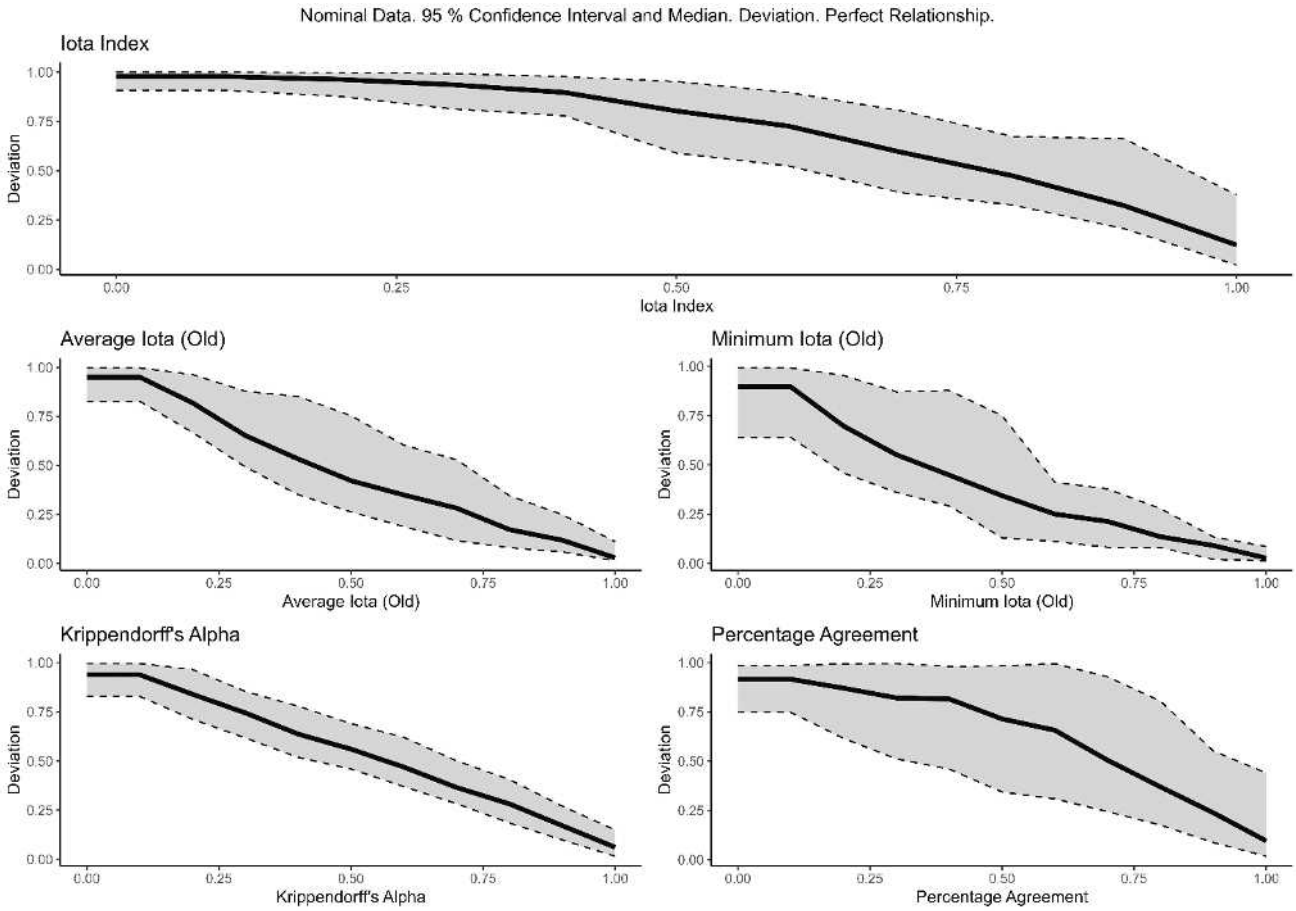
Appendix B – Illustrations of the Relationship Between Reliability and the Deviation Between the True and Estimated Association/Correlation

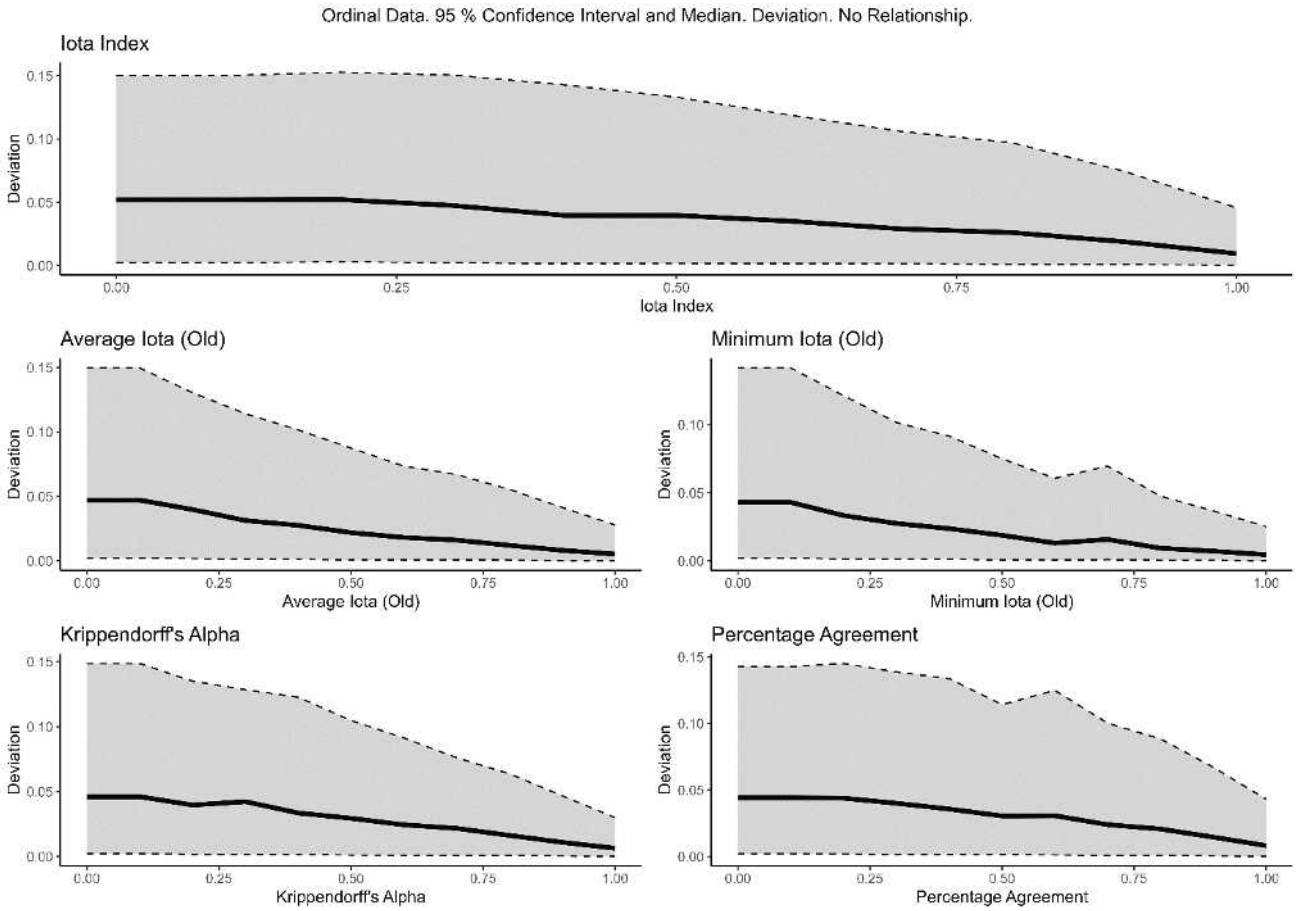


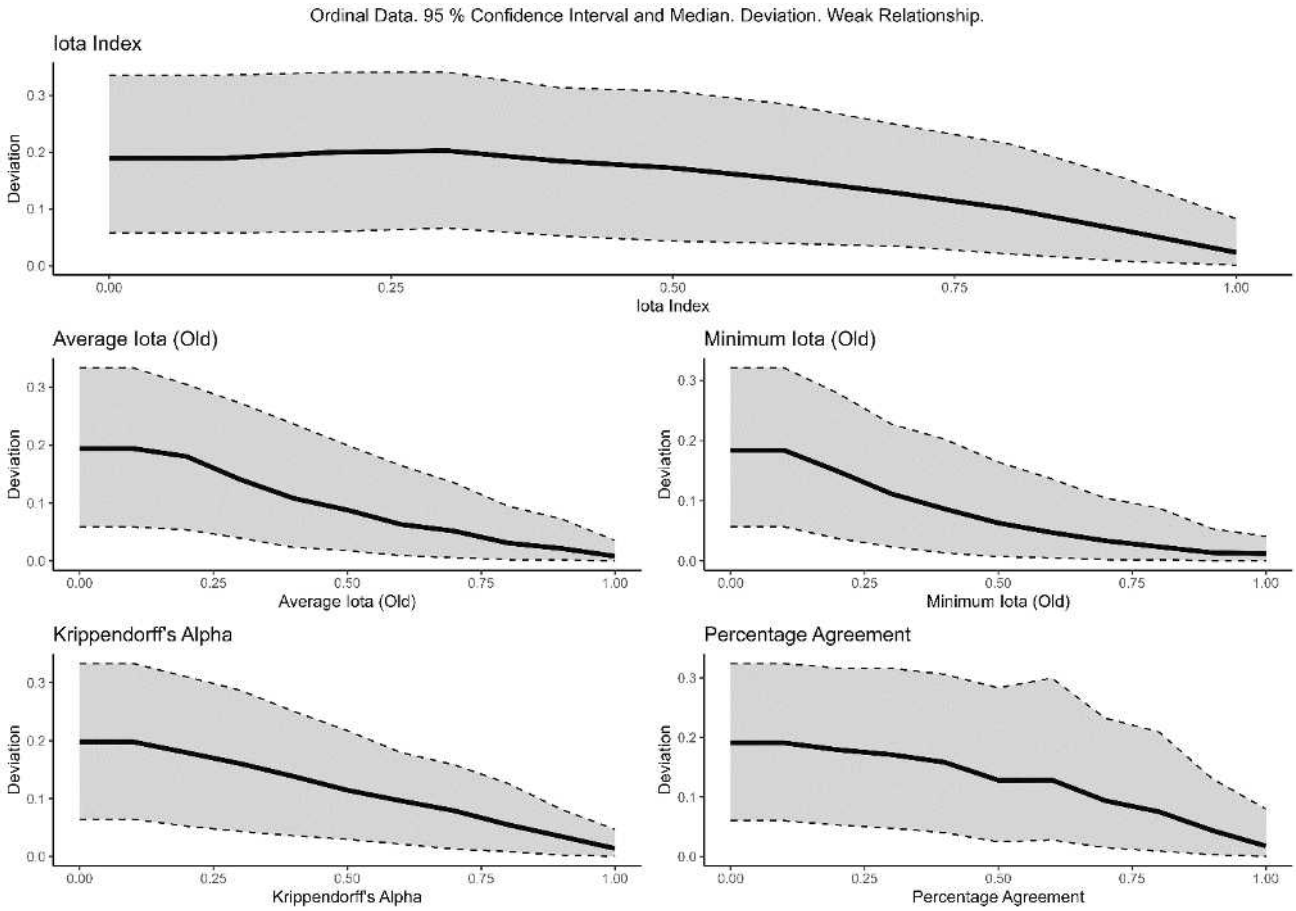

Nominal Data. 95 % Confidence Interval and Median. Deviation. Weak Relationship.
Iota Index
Deviation
0.0
0.1
0.2
0.3
0.00
0.25
0.50
0.75
1.00
Iota Index
Average Iota (Old)
Deviation
Average Iota (Old)
Minimum Iota (Old)
Deviation
Minimum Iota (Old)
Krippendorff's Alpha
Deviation
Krippendorff's Alpha
Percentage Agreement
Deviation
Percentage Agreement

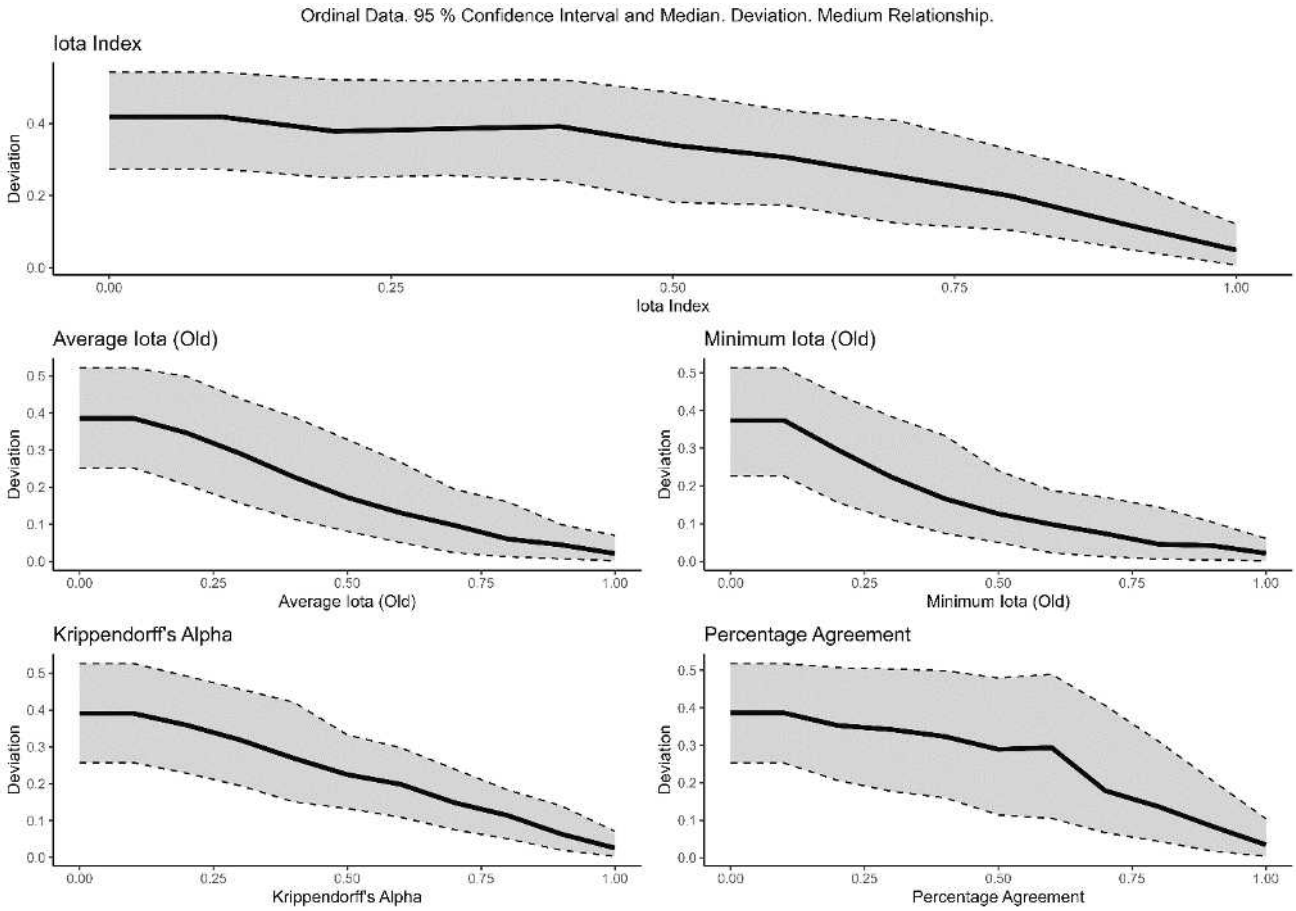

Nominal Data. 95 % Confidence Interval and Median. Deviation. Medium Relationship.
Iota Index
Deviation
Iota Index
Average Iota (Old)
Minimum Iota (Old)
Krippendorff's Alpha
Percentage Agreement

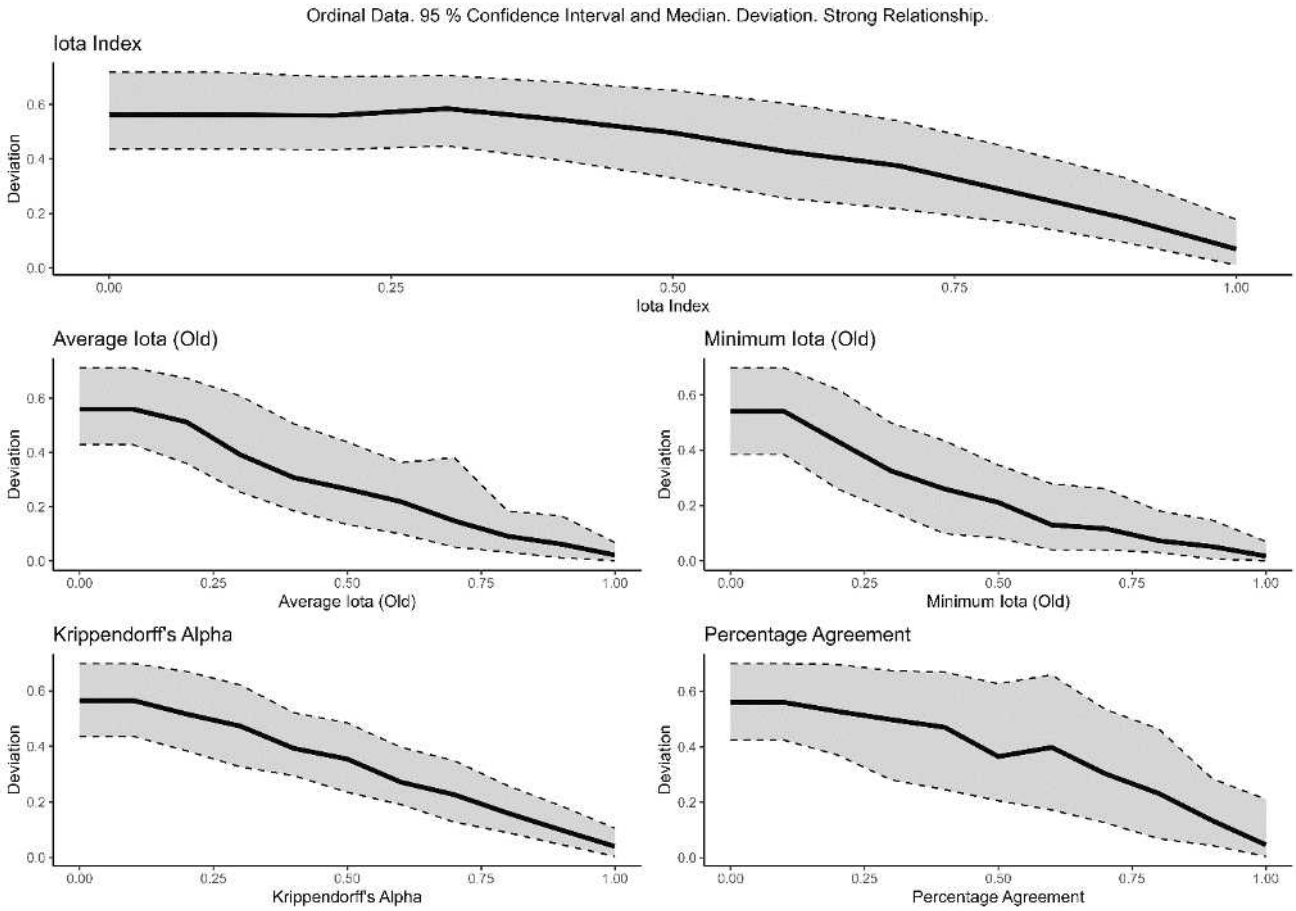

Nominal Data. 95 % Confidence Interval and Median. Deviation. Strong Relationship.
Iota Index
Deviation
0.0
0.2
0.4
0.6
0.00
0.25
0.50
0.75
1.00
Iota Index
Average Iota (Old)
Average Iota (Old)
Minimum Iota (Old)
Minimum Iota (Old)
Krippendorff's Alpha
Krippendorff's Alpha
Percentage Agreement
Percentage Agreement

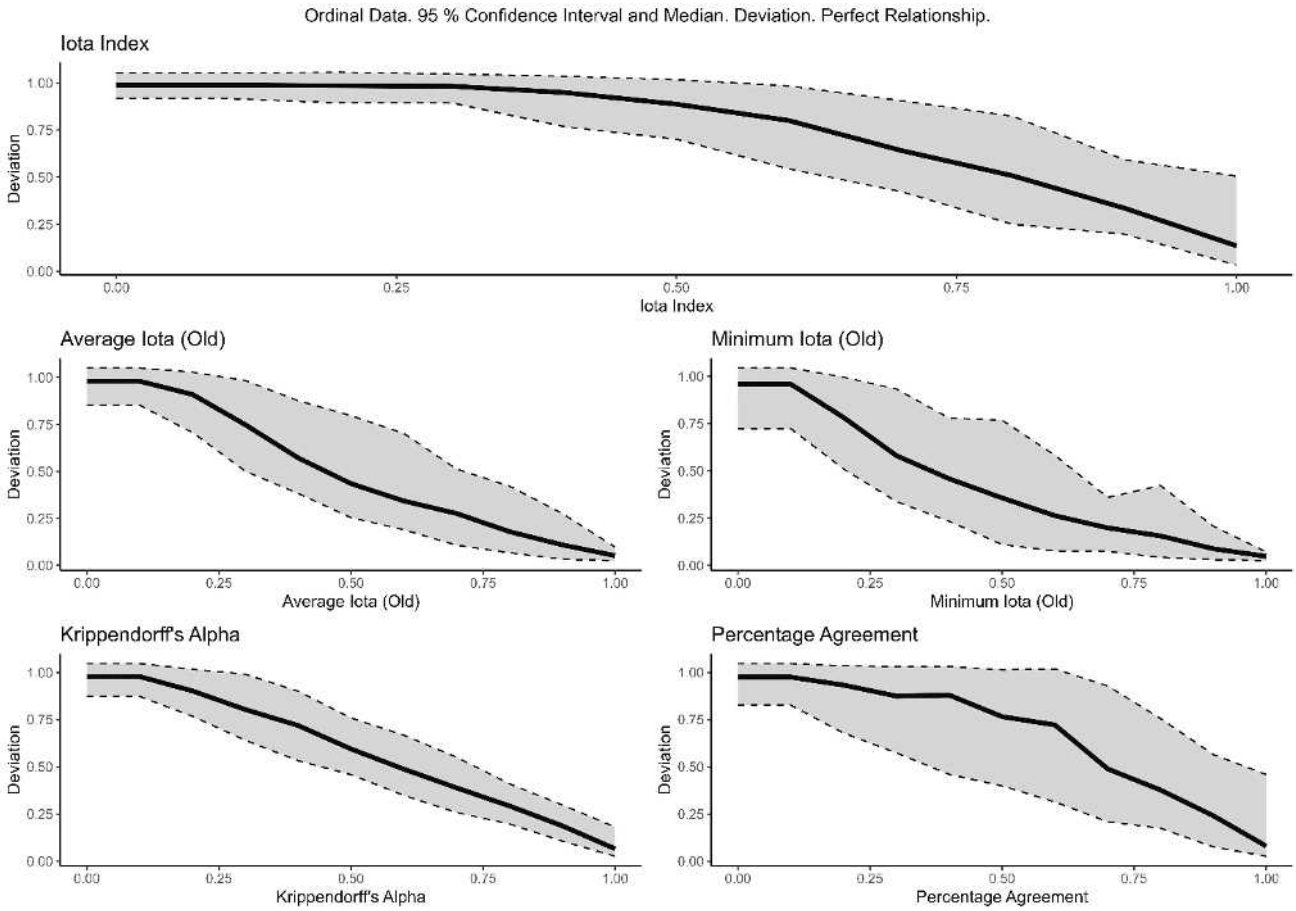

Nominal Data. 95 % Confidence Interval and Median. Deviation. Perfect Relationship.
Iota Index
Deviation
Iota Index
Average Iota (Old)
Minimum Iota (Old)
Krippendorff's Alpha
Percentage Agreement
0.00
0.25
0.50
0.75
1.00


Ordinal Data. 95 % Confidence Interval and Median. Deviation. No Relationship.
Iota Index
Deviation
0.15
0.10
0.05
0.00
0.00
0.25
0.50
0.75
1.00
Iota Index
Average Iota (Old)
Average Iota (Old)
Minimum Iota (Old)
Minimum Iota (Old)
Krippendorff's Alpha
Krippendorff's Alpha
Percentage Agreement
Percentage Agreement


Ordinal Data. 95 % Confidence Interval and Median. Deviation. Weak Relationship.
Iota Index
Deviation
0.0
0.1
0.2
0.3
0.00
0.25
0.50
0.75
1.00
Iota Index
Average Iota (Old)
Average Iota (Old)
Minimum Iota (Old)
Minimum Iota (Old)
Krippendorff's Alpha
Krippendorff's Alpha
Percentage Agreement
Percentage Agreement


Ordinal Data. 95 % Confidence Interval and Median. Deviation. Medium Relationship.
Iota Index
Deviation
Iota Index
Average Iota (Old)
Minimum Iota (Old)
Krippendorff's Alpha
Percentage Agreement


Ordinal Data. 95 % Confidence Interval and Median. Deviation. Strong Relationship.
Iota Index
Deviation
0.0
0.2
0.4
0.6
0.00
0.25
0.50
0.75
1.00
Iota Index
Average Iota (Old)
Deviation
Average Iota (Old)
Minimum Iota (Old)
Deviation
Minimum Iota (Old)
Krippendorff's Alpha
Deviation
Krippendorff's Alpha
Percentage Agreement
Deviation
Percentage Agreement


Ordinal Data. 95 % Confidence Interval and Median. Deviation. Perfect Relationship.
Iota Index
Deviation
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
Iota Index
Average Iota (Old)
Average Iota (Old)
Minimum Iota (Old)
Minimum Iota (Old)
Krippendorff's Alpha
Krippendorff's Alpha
Percentage Agreement
Percentage Agreement

Appendix C – Global Indices of Model Fit in Simulation Study II

RMSEA

Nominal Data – Deviation					
	No Relationship	Weak Relationship	Medium Relationship	Strong Relationship	Perfect Relationship
Iota Index	0	0	0	0	0
Average Iota	0	0	0	0	0
Minimum Iota	0	0	0	0	0
Krippendorff's Alpha	0	0	0	0	0
Percentage Agreement	0	0	0	0	0
Nominal Data – Type I Error					
	No Relationship	Weak Relationship	Medium Relationship	Strong Relationship	Perfect Relationship
Iota Index	0.001	0	0	0	0
Average Iota	0	0	0	0	0.003
Minimum Iota	0.001	0.002	0.002	0.001	0.03
Krippendorff's Alpha	0	0	0	0	0.003
Percentage Agreement	0	0	0	0	0.035
Nominal Data – Type II Error					
	No Relationship	Weak Relationship	Medium Relationship	Strong Relationship	Perfect Relationship
Iota Index	0	0			
Average Iota	0	0			
Minimum Iota	0	0			
Krippendorff's Alpha	0.001	0.001			
Percentage Agreement	0.001	0			
Nominal Data – Correct Classification of Effect Sizes					
	No Relationship	Weak Relationship	Medium Relationship	Strong Relationship	Perfect Relationship
Iota Index	0.004	0.001	0	0.017	NA
Average Iota	0.005	0.003	0	0.01	NA
Minimum Iota	0.006	0.004	0	NA	NA
Krippendorff's Alpha	0.007	0.004	0	0.002	NA
Percentage Agreement	0.005	0	0	NA	NA

Ordinal Data – Deviation					
	No Relationship	Weak Relationship	Medium Relationship	Strong Relationship	Perfect Relationship
Iota Index	0	0	0	0	0
Average Iota	0	0	0	0	0
Minimum Iota	0	0	0	0	0
Krippendorff's Alpha	0	0	0	0	0
Percentage Agreement	0	0	0	0	0
Ordinal Data – Type I Error					
	No Relationship	Weak Relationship	Medium Relationship	Strong Relationship	Perfect Relationship
Iota Index	0.002	0	0.003	0	0.002
Average Iota	0.001	0.001	0	0.002	0.001
Minimum Iota	0	0	0	0	0.001
Krippendorff's Alpha	0.003	0.001	0.001	0	0
Percentage Agreement	0.001	0	0.001	0.001	0.001
Ordinal Data – Type II Error					
	No Relationship	Weak Relationship	Medium Relationship	Strong Relationship	Perfect Relationship
Iota Index	0	0			
Average Iota	0	0			
Minimum Iota	0	0			
Krippendorff's Alpha	0	0			
Percentage Agreement	0	0			
Ordinal Data – Correct Classification of Effect Sizes					
	No Relationship	Weak Relationship	Medium Relationship	Strong Relationship	Perfect Relationship
Iota Index	0.001	0	0	0.004	0.004
Average Iota	0.001	0	0.002	0.001	NA
Minimum Iota	0.002	0	0.003	0	NA
Krippendorff's Alpha	0	0	0.002	0.003	0.015
Percentage Agreement	0	0.001	0	0	NA

SRMR (Maximal Value of Within and Between Level)

Nominal Data – Deviation					
	No Relationship	Weak Relationship	Medium Relationship	Strong Relationship	Perfect Relationship
Iota Index	0.001	0.01	0.006	0.005	0.002
Average Iota	0.005	0.017	0.008	0.003	0.002
Minimum Iota	0.003	0.019	0.011	0.004	0.004
Krippendorff's Alpha	0.005	0.024	0.008	0.002	0.001
Percentage Agreement	0.008	0.009	0.009	0.003	0.001
Nominal Data – Type I Error					
	No Relationship	Weak Relationship	Medium Relationship	Strong Relationship	Perfect Relationship
Iota Index	0.005	0.001	0	0	0.001
Average Iota	0.002	0.002	0.002	0.001	0.006
Minimum Iota	0.004	0.003	0.004	0.002	0.012
Krippendorff's Alpha	0.004	0.002	0.001	0	0.005
Percentage Agreement	0.001	0.001	0.001	0	0.017
Nominal Data – Type II Error					
	No Relationship	Weak Relationship	Medium Relationship	Strong Relationship	Perfect Relationship
Iota Index	0.012	0.005			
Average Iota	0.01	0.003			
Minimum Iota	0.006	0.004			
Krippendorff's Alpha	0.013	0.011			
Percentage Agreement	0.012	0.001			
Nominal Data – Correct Classification of Effect Sizes					
	No Relationship	Weak Relationship	Medium Relationship	Strong Relationship	Perfect Relationship
Iota Index	0.011	0.003	0.001	0.009	NA
Average Iota	0.014	0.005	0.003	0.004	NA
Minimum Iota	0.015	0.008	0.003	NA	NA
Krippendorff's Alpha	0.018	0.007	0.001	0.001	NA
Percentage Agreement	0.012	0.001	0.002	NA	NA

Ordinal Data – Deviation					
	No Relationship	Weak Relationship	Medium Relationship	Strong Relationship	Perfect Relationship
Iota Index	0.002	0.007	0.004	0.002	0.002
Average Iota	0.004	0.009	0.008	0	0.001
Minimum Iota	0.009	0.008	0.011	0.004	0
Krippendorff's Alpha	0.002	0.007	0.008	0.001	0.003
Percentage Agreement	0.004	0.011	0.007	0.002	0
Ordinal Data – Type I Error					
	No Relationship	Weak Relationship	Medium Relationship	Strong Relationship	Perfect Relationship
Iota Index	0.006	0	0.004	0.002	0.004
Average Iota	0.004	0.003	0.001	0.003	0.002
Minimum Iota	0.001	0.001	0	0.001	0.003
Krippendorff's Alpha	0.007	0.002	0.002	0.001	0.001
Percentage Agreement	0.004	0.002	0.003	0.003	0.003
Ordinal Data – Type II Error					
	No Relationship	Weak Relationship	Medium Relationship	Strong Relationship	Perfect Relationship
Iota Index	0.005	0.008			
Average Iota	0.007	0.008			
Minimum Iota	0.006	0.011			
Krippendorff's Alpha	0	0.015			
Percentage Agreement	0.005	0.004			
Ordinal Data – Correct Classification of Effect Sizes					
	No Relationship	Weak Relationship	Medium Relationship	Strong Relationship	Perfect Relationship
Iota Index	0.004	0.001	0.001	0.015	0.004
Average Iota	0.005	0.002	0.005	0.006	NA
Minimum Iota	0.007	0.002	0.006	0.002	NA
Krippendorff's Alpha	0	0	0.005	0.012	0.004
Percentage Agreement	0.003	0.003	0.003	0.001	NA

CFI

Nominal Data – Deviation					
	No Relationship	Weak Relationship	Medium Relationship	Strong Relationship	Perfect Relationship
Iota Index	1	1	1	1	1
Average Iota	1	1	1	1	1
Minimum Iota	1	1	1	1	1
Krippendorff's Alpha	1	1	1	1	1
Percentage Agreement	1	1	1	1	1
Nominal Data – Type I Error					
	No Relationship	Weak Relationship	Medium Relationship	Strong Relationship	Perfect Relationship
Iota Index	0.999	1	1	1	1
Average Iota	1	1	1	1	0.999
Minimum Iota	0.999	1	1	1	0.996
Krippendorff's Alpha	1	1	1	1	0.999
Percentage Agreement	1	1	1	1	0.993
Nominal Data – Type II Error					
	No Relationship	Weak Relationship	Medium Relationship	Strong Relationship	Perfect Relationship
Iota Index	1	1			
Average Iota	1	1			
Minimum Iota	1	1			
Krippendorff's Alpha	0.986	0.998			
Percentage Agreement	0.98	1			
Nominal Data – Correct Classification of Effect Sizes					
	No Relationship	Weak Relationship	Medium Relationship	Strong Relationship	Perfect Relationship
Iota Index	0.809	1	1	0.999	NA
Average Iota	0.933	1	1	1	NA
Minimum Iota	0.976	0.999	1	NA	NA
Krippendorff's Alpha	0.787	0.999	1	1	NA
Percentage Agreement	0.984	1	1	NA	NA

Ordinal Data – Deviation					
	No Relationship	Weak Relationship	Medium Relationship	Strong Relationship	Perfect Relationship
Iota Index	1	1	1	1	1
Average Iota	1	1	1	1	1
Minimum Iota	1	1	1	1	1
Krippendorff's Alpha	1	1	1	1	1
Percentage Agreement	1	1	1	1	1
Ordinal Data – Type I Error					
	No Relationship	Weak Relationship	Medium Relationship	Strong Relationship	Perfect Relationship
Iota Index	0.999	1	1	1	1
Average Iota	0.999	1	1	1	1
Minimum Iota	1	1	1	1	1
Krippendorff's Alpha	0.998	1	1	1	1
Percentage Agreement	0.999	1	1	1	1
Ordinal Data – Type II Error					
	No Relationship	Weak Relationship	Medium Relationship	Strong Relationship	Perfect Relationship
Iota Index	1	1			
Average Iota	1	1			
Minimum Iota	1	1			
Krippendorff's Alpha	1	0.998			
Percentage Agreement	1	1			
Ordinal Data – Correct Classification of Effect Sizes					
	No Relationship	Weak Relationship	Medium Relationship	Strong Relationship	Perfect Relationship
Iota Index	0.996	1	1	0.999	1
Average Iota	0.983	1	1	1	NA
Minimum Iota	0.813	1	1	1	NA
Krippendorff's Alpha	1	1	1	0.999	0.999
Percentage Agreement	1	1	1	1	NA

Appendix D – Global Indices of Model Fit in Simulation Study III

RMSEA

Nominal Data – Deviation					
	No Relationship	Weak Relationship	Medium Relationship	Strong Relationship	Perfect Relationship
Iota Index	0	0	0	0	0
$d = 1.5$	0	0	0	0	0
$d = 2.0$	0	0	0	0	0
$d = 3.0$	0	0	0	0	0
$d = 4.0$	0	0	0	0	0
$d_{dyn} = 0.5$	0	0	0	0	0
$d_{dyn} = 2.0$	0	0	0	0	0
Nominal Data – Type I Error					
	No Relationship	Weak Relationship	Medium Relationship	Strong Relationship	Perfect Relationship
Iota Index	0.007	0.013	0.003	0.006	0.004
$d = 1.5$	0.007	0.016	0.004	0.006	0.003
$d = 2.0$	0.007	0.017	0.005	0.005	0.002
$d = 3.0$	0.008	0.017	0.005	0.004	0
$d = 4.0$	0.008	0.016	0.006	0.003	0
$d_{dyn} = 0.5$	0.007	0.017	0.005	0.006	0.003
$d_{dyn} = 2.0$	0.007	0.015	0.005	0.006	0.004
Nominal Data – Correct Classification of Effect Sizes					
	No Relationship	Weak Relationship	Medium Relationship	Strong Relationship	Perfect Relationship
Iota Index	0.001	0.014	0.005	0.05	NA
$d = 1.5$	0.001	0.015	0.006	0.033	NA
$d = 2.0$	0.001	0.015	0.006	0.018	NA
$d = 3.0$	0.001	0.015	0.007	NA	NA
$d = 4.0$	0	0.015	0.008	NA	NA
$d_{dyn} = 0.5$	0.001	0.016	0.007	0.023	NA
$d_{dyn} = 2.0$	0.001	0.015	0.006	0.012	NA

Ordinal Data – Deviation					
	No Relationship	Weak Relationship	Medium Relationship	Strong Relationship	Perfect Relationship
Iota Index	0	0	0	0	0
$d = 1.5$	0	0	0	0	0
$d = 2.0$	0	0	0	0	0
$d = 3.0$	0	0	0	0	0
$d = 4.0$	0	0	0	0	0
$d_{dyn} = 0.5$	0	0	0	0	0
$d_{dyn} = 2.0$	0	0	0	0	0
Ordinal Data – Type I Error					
	No Relationship	Weak Relationship	Medium Relationship	Strong Relationship	Perfect Relationship
Iota Index	0.002	0.011	0	0.002	0.002
$d = 1.5$	0.002	0.013	0	0.002	0.003
$d = 2.0$	0.001	0.013	0	0.002	0.003
$d = 3.0$	0	0.014	0.001	0.002	0.004
$d = 4.0$	0	0.013	0	0.002	0.005
$d_{dyn} = 0.5$	0.002	0.014	0	0.002	0.004
$d_{dyn} = 2.0$	0.002	0.013	0	0.003	0.004
Ordinal Data – Correct Classification of Effect Sizes					
	No Relationship	Weak Relationship	Medium Relationship	Strong Relationship	Perfect Relationship
Iota Index	0.004	0	0	0.002	NA
$d = 1.5$	0.003	0.002	0	0.002	0.007
$d = 2.0$	0.002	0.003	0	0.002	NA
$d = 3.0$	0.002	0.004	0.001	0.001	0.007
$d = 4.0$	0.001	0.005	0.001	0.001	NA
$d_{dyn} = 0.5$	0.003	0.003	0	0.001	0.005
$d_{dyn} = 2.0$	0.004	0.002	0	0.001	NA

SRMR (Maximal Value of Within and Between Level)

Nominal Data – Deviation					
	No Relationship	Weak Relationship	Medium Relationship	Strong Relationship	Perfect Relationship
Iota Index	0.026	0.023	0.004	0.008	0.004
$d = 1.5$	0.027	0.024	0.003	0.007	0.003
$d = 2.0$	0.027	0.024	0.003	0.006	0.003
$d = 3.0$	0.028	0.024	0.003	0.006	0.003
$d = 4.0$	0.028	0.023	0.004	0.005	0.004
$d_{dyn} = 0.5$	0.028	0.024	0.003	0.006	0.004
$d_{dyn} = 2.0$	0.028	0.024	0.004	0.007	0.004
Nominal Data – Type I Error					
	No Relationship	Weak Relationship	Medium Relationship	Strong Relationship	Perfect Relationship
Iota Index	0.016	0.019	0.004	0.007	0.007
$d = 1.5$	0.018	0.021	0.005	0.008	0.005
$d = 2.0$	0.019	0.022	0.006	0.008	0.004
$d = 3.0$	0.02	0.023	0.007	0.007	0.003
$d = 4.0$	0.021	0.024	0.009	0.006	0.003
$d_{dyn} = 0.5$	0.018	0.022	0.006	0.008	0.006
$d_{dyn} = 2.0$	0.018	0.02	0.006	0.008	0.007
Nominal Data – Correct Classification of Effect Sizes					
	No Relationship	Weak Relationship	Medium Relationship	Strong Relationship	Perfect Relationship
Iota Index	0.005	0.024	0.013	0.028	NA
$d = 1.5$	0.005	0.026	0.014	0.016	NA
$d = 2.0$	0.005	0.027	0.014	0.008	NA
$d = 3.0$	0.005	0.028	0.016	NA	NA
$d = 4.0$	0.004	0.029	0.017	NA	NA
$d_{dyn} = 0.5$	0.005	0.028	0.015	0.011	NA
$d_{dyn} = 2.0$	0.006	0.027	0.015	0.005	NA

Ordinal Data – Deviation					
	No Relationship	Weak Relationship	Medium Relationship	Strong Relationship	Perfect Relationship
Iota Index	0.014	0.002	0.004	0.005	0.004
$d = 1.5$	0.011	0.003	0.004	0.005	0.005
$d = 2.0$	0.009	0.004	0.005	0.005	0.006
$d = 3.0$	0.007	0.005	0.005	0.005	0.007
$d = 4.0$	0.006	0.006	0.005	0.004	0.008
$d_{dyn} = 0.5$	0.011	0.004	0.004	0.005	0.007
$d_{dyn} = 2.0$	0.013	0.004	0.004	0.005	0.006
Ordinal Data – Type I Error					
	No Relationship	Weak Relationship	Medium Relationship	Strong Relationship	Perfect Relationship
Iota Index	0.006	0.016	0.001	0.004	0.004
$d = 1.5$	0.005	0.017	0.001	0.003	0.005
$d = 2.0$	0.004	0.018	0.002	0.003	0.006
$d = 3.0$	0.002	0.019	0.002	0.003	0.008
$d = 4.0$	0.002	0.02	0.002	0.004	0.009
$d_{dyn} = 0.5$	0.005	0.019	0.002	0.003	0.007
$d_{dyn} = 2.0$	0.006	0.018	0.002	0.004	0.006
Ordinal Data – Correct Classification of Effect Sizes					
	No Relationship	Weak Relationship	Medium Relationship	Strong Relationship	Perfect Relationship
Iota Index	0.011	0.002	0.002	0.008	NA
$d = 1.5$	0.009	0.004	0.003	0.008	0.008
$d = 2.0$	0.008	0.005	0.003	0.007	NA
$d = 3.0$	0.006	0.007	0.003	0.006	0.004
$d = 4.0$	0.005	0.008	0.003	0.005	NA
$d_{dyn} = 0.5$	0.009	0.006	0.003	0.006	0.005
$d_{dyn} = 2.0$	0.01	0.005	0.003	0.007	NA

CFI

Nominal Data – Deviation					
	No Relationship	Weak Relationship	Medium Relationship	Strong Relationship	Perfect Relationship
Iota Index	1	1	1	1	1
$d = 1.5$	1	1	1	1	1
$d = 2.0$	1	1	1	1	1
$d = 3.0$	1	1	1	1	1
$d = 4.0$	1	1	1	1	1
$d_{dyn} = 0.5$	1	1	1	1	1
$d_{dyn} = 2.0$	1	1	1	1	1
Nominal Data – Type I Error					
	No Relationship	Weak Relationship	Medium Relationship	Strong Relationship	Perfect Relationship
Iota Index	0.983	0.998	1	0.999	0.999
$d = 1.5$	0.978	0.996	1	0.999	0.999
$d = 2.0$	0.972	0.995	1	0.999	0.999
$d = 3.0$	0.961	0.994	0.999	0.999	1
$d = 4.0$	0.954	0.993	0.999	0.999	1
$d_{dyn} = 0.5$	0.975	0.995	0.999	0.999	0.999
$d_{dyn} = 2.0$	0.979	0.997	1	0.999	0.999
Nominal Data – Correct Classification of Effect Sizes					
	No Relationship	Weak Relationship	Medium Relationship	Strong Relationship	Perfect Relationship
Iota Index	0.983	0.994	0.999	0.991	NA
$d = 1.5$	0.996	0.993	0.999	0.996	NA
$d = 2.0$	0.998	0.991	0.998	0.999	NA
$d = 3.0$	1	0.988	0.998	NA	NA
$d = 4.0$	1	0.987	0.997	NA	NA
$d_{dyn} = 0.5$	0.997	0.991	0.998	0.998	NA
$d_{dyn} = 2.0$	0.989	0.992	0.998	1	NA

Ordinal Data – Deviation					
	No Relationship	Weak Relationship	Medium Relationship	Strong Relationship	Perfect Relationship
Iota Index	1	1	1	1	1
$d = 1.5$	1	1	1	1	1
$d = 2.0$	1	1	1	1	1
$d = 3.0$	1	1	1	1	1
$d = 4.0$	1	1	1	1	1
$d_{dyn} = 0.5$	1	1	1	1	1
$d_{dyn} = 2.0$	1	1	1	1	1
Ordinal Data – Type I Error					
	No Relationship	Weak Relationship	Medium Relationship	Strong Relationship	Perfect Relationship
Iota Index	0.998	0.998	1	1	1
$d = 1.5$	0.999	0.997	1	1	0.999
$d = 2.0$	1	0.996	1	1	0.999
$d = 3.0$	1	0.994	1	1	0.998
$d = 4.0$	1	0.994	1	1	0.997
$d_{dyn} = 0.5$	0.999	0.996	1	1	0.998
$d_{dyn} = 2.0$	0.999	0.997	1	1	0.999
Ordinal Data – Correct Classification of Effect Sizes					
	No Relationship	Weak Relationship	Medium Relationship	Strong Relationship	Perfect Relationship
Iota Index	0.923	1	1	1	NA
$d = 1.5$	0.952	1	1	1	0.997
$d = 2.0$	0.969	1	1	1	NA
$d = 3.0$	0.986	0.999	1	1	0.999
$d = 4.0$	0.994	0.999	1	1	NA
$d_{dyn} = 0.5$	0.959	1	1	1	1
$d_{dyn} = 2.0$	0.938	1	1	1	NA

In educational settings, analyzing textual data via content analysis is a popular research method. The data is a valuable source of information as it offers deep insights into learning and learning outcomes. In practice, it can be used to improve classroom diagnostics and instruction. Nowadays, technology such as learning analytics can be used for the same cause. For both purposes, reliable research instruments are needed. Content analysis, often the measure of choice, is required to meet quality criteria such as objectivity, reliability and validity. However, some of the reliability measures most frequently used have lately been discussed controversially, indicating that there is room for improvement. The first generation of the Iota concept caters to the idea of improved reliability measures for content analysis done by humans or artificial intelligences. In this book, the authors introduce a refined measure: The Iota concept of the second generation. In contrast to pre-existing measures, second generation Iota can for example a) provide insights into the reliability of every single category of a scale and how a coding scheme may produce bias, b) provide rules of thumb for evaluating content analysis and c) provide possibilities for data replication and error-corrected data.

This book is structured as a guide for researchers that want to learn more about the mechanics and details of the Iota concept or use it as the reliability measure of choice in their research.

Logos Verlag Berlin

ISBN 978-3-8325-5581-8

ISSN 2629-3137