Verlässlichkeit von Inhaltsanalysedaten

Reliabilitätstest, Errechnen und Interpretieren von Reliabilitätskoeffizienten für mehr als zwei Codierer

Steffen Kolb

Der Beitrag untersucht die Abhängigkeit von Reliabilitätskoeffizienten von der Anzahl der Codierer. Da alle gängigen Reliabilitätskoeffizienten auf paarweisen Übereinstimmungs- (Holsti, Scotts Pi, Cohens Kappa) oder paarweisen Abweichungsbestimmungen (Krippendorffs Alpha) beruhen, ergeben sie bei mehr als zwei Codierern keine vergleichbaren Werte für Studien mit unterschiedlichen Codiererzahlen. Darüber hinaus beleuchtet der Beitrag die theoretisch bzw. methodologisch fragwürdigen Ergebnisse von Reliabilitätsberechnungen bei gleicher Codiererzahl (größer als zwei): Alle gängigen Reliabilitätskoeffizienten sind tendenziell höher, wenn sich Fehler in wenigen Fällen häufen. Diese können als systematische Fehler bezeichnet und auf ein unpräzises Codebuch zurückgeführt werden. In gängigen Reliabilitätsberechnungen werden also Studien, die systematische Fehler aufweisen, besser bewertet als solche, in denen „nur" unsystematische, d. h. vereinzelte Fehlcodierungen vorkommen. Um diese Probleme zu überwinden, schlägt der Beitrag einen neuen Reliabilitätskoeffizienten für nominalskalierte Daten vor, der sowohl auf einer für unterschiedliche Codiererzahlen vergleichbaren Berechnung beruht als auch das Problem der Fehlerhäufungen überwindet und somit die Vergleichbarkeit von Reliabilitätsberechnungen für Studien mit unterschiedlichen Codiererzahlen streng genommen erst herstellt.

Keywords: Reliabilität, Inhaltsanalyse, Reliabilitätskoeffizient, Methoden

1. Problemstellung

Bedingt durch das wissenschaftstheoretische Ideal der Intersubjektivität bzw. intersubjektiven Nachvollziehbarkeit[1] muss jede wissenschaftliche Studie reproduzierbar *sein* oder besser *gemacht werden*[2]. Dies soll dazu führen, dass eine kritische Auseinandersetzung mit den Ergebnissen möglich wird. In Zweifelsfällen kann sogar eine regelrechte Nachprüfung durchgeführt werden, indem die Studie wiederholt wird. Um die intersubjektive Evidenz der Wissenschaft nicht zu untergraben, ist also insbesondere auf Offenlegung des Erhebungsinstrumentes und eine Kontrolle desselben zu achten (vgl. z. B. Früh, 1991: 37 f.; Mayring, 2003: 12; Merten, 1995; Neuendorf, 2002: 143 f.).

Besonders für die Analyse von Medieninhalten ergeben sich dabei spezifische Chancen und Schwierigkeiten. „Zur systematischen, intersubjektiv nachvollziehbaren Beschreibung inhaltlicher und formaler Merkmale von Mitteilungen" (Früh, 1991: 24, vgl. z. B. auch Neuendorf, 2002), also z. B. von Medieninhalten, ist die Inhaltsanalyse die ge-

1 Dies gilt zumindest für die Erkenntnistheorie, die sozialwissenschaftlichen Studien zugrunde liegt (vgl. Seiffert, 1971).

2 Damit ist gemeint, dass eine Offenlegung von Methode und Vorgehensweise streng genommen eine Studie erst nachvollziehbar macht, keine Studie also per se schon nachvollziehbarer ist als eine andere.

eignete empirische Methode.[3] Einen wesentlichen Anteil an der intersubjektiven Nachvollziehbarkeit von Inhaltsanalysen hat die Systematik der Arbeit: Eine klare und strukturierte Forschungsstrategie und deren konsequente Anwendung sind unabdingbar. Die Einführung und Beachtung möglichst eindeutiger Codierregeln, Kategorienschemata und Begriffsdefinitionen, unterstützt durch intensive Codiererschulungen, ermöglicht eine Überprüfung der Verlässlichkeit oder Reliabilität der Messung.

Dazu werden die übereinstimmenden und abweichenden Codierungen eines Teils des Untersuchungsmaterials betrachtet, das (im Idealfall) alle beteiligten Codierer und der Forscher vor Beginn der eigentlichen Untersuchung analysiert haben. Gibt es „viele“[4] Übereinstimmungen, die in der Regel anteilig zur Gesamtzahl der Codierungen dargestellt werden, so gilt die Messung als gut bzw. reliabel.[5] Implizit soll daraus geschlossen werden, dass die Codierregeln auch von anderen Personen verstanden werden können, also (eine gewisse) intersubjektive Nachvollziehbarkeit angenommen werden kann.[6] So bestimmte Übereinstimmungskoeffizienten können somit als „notwendiges, wenngleich nicht hinreichendes Kriterium für die methodische Güte von Inhaltsanalysen“ (Lauf, 2001: 57) bezeichnet werden. Ob sich diese Gütebestimmung allerdings auf die Qualität der Codierer und/oder des Codebuchs bezieht, ist wissenschaftlich nicht detailliert diskutiert.

---

3 Bei Inhaltsanalysen werden für die Analyse relevante Bedeutungsaspekte selektiert und strukturiert. Im Allgemeinen handelt es sich also nicht um eine reine Frequenzanalyse kommunikativer Akte, sondern um eine Abstraktion, die auf Bedeutungen schließen will. Als Frequenzanalysen bezeichnet Früh (1991) reines Zählen von Wort- oder Begriffshäufigkeiten (vgl. auch Mayring, 2003). Bei dieser Form von ‚Inhaltsanalysen‘ ergeben sich normalerweise kaum Probleme von mangelnder intersubjektiver Nachvollziehbarkeit, so dass sie hier nur am Rande behandelt werden. Für eine Typologie unterschiedlicher Analyseformen von Texten oder Kommunikation allgemein und eine Abstufung von ‚Schwierigkeitsgraden‘ bei Inhaltsanalysen vgl. Neuendorf (2002: 1–9).

4 Auf die Problematik des Grades der Übereinstimmung bzw. der Bedeutung von „viele“ wird im Folgenden noch konkret eingegangen.

5 Auf diskursbasierte Verfahren, die bei abweichenden Auffassungen über die Codierung diskutieren, um entweder zu einer Mehrheitsentscheidung oder zu einem einstimmigen Votum zu kommen, kann hier nur hingewiesen werden (vgl. z. B. Meyen, 2002). Dabei ist allerdings zu beachten, dass sich diese Verfahren auf abweichende Codierungen bzw. Codiervorschläge, also auf die gleiche Basis wie die hier behandelten Koeffizienten, beziehen. Auch die Diskussion von Codierungen wird in der Codiererschulung immer wieder als Vermittlungsvehikel eingesetzt, um den Codierern das Kategorienschema näher zu bringen. Der Unterschied liegt lediglich darin, ob ein in sich geschlossener Reliabilitätstest erfolgt, der dann in einen oder mehrere Koeffizienten mündet. Diese sollen dann die intersubjektive Nachvollziehbarkeit „in Kurzform“ herstellen, die bei der Offenlegung von Mitschriften der Gruppendiskussionen wegen der Datenmengen wohl eher schwierig zu erreichen sein dürfte.

6 An diesen Folgerungen könnte man generell kritisieren, dass für einen Inferenzschluss auf andere Forscher streng genommen auch Kommunikationswissenschaftler die Codierung durchführen müssten. Zudem ist in den meisten Fällen die Zahl der Codierer zu gering für solche Inferenzen oder gar den Repräsentativitätsschluss auf jeden Leser der Studie. Dieser potenziellen Kritik sei hier entgegengestellt, dass eine generelle Nachvollziehbarkeit von jedem Menschen im Zeitalter von Arbeitsteilung und Spezialisierung niemals erreicht werden kann, und dass größere Anzahlen von (meistens) Kommunikationswissenschaft studierenden Codierern, die als werdende Kommunikationswissenschaftler verstanden werden können, aus forschungsökonomischen Gründen selten möglich sein dürften.

In diesem Artikel soll verdeutlicht werden, dass Reliabilitätswert nicht gleich Reliabilitätswert ist. Dabei geht es nicht um unterschiedliche Koeffizienten wie z. B. Holsti (1969), Scotts Pi (Scott, 1955), Cohens Kappa (Cohen, 1960), Krippendorffs Alpha (Krippendorf, 1980) oder andere. Auch die Unterschiede von Reliabilitätswerten für unterschiedliche Skalenniveaus sind hier nicht gemeint, auf die hier zumindest beispielhaft in Form von Betrachtungen der Korrelation bzw. Kovariation hingewiesen werden soll (vgl. Neuendorf, 2002: 144ff.; Traub, 1994). Die zuerst aufgezählten Koeffizienten (und dieser Beitrag) liefern streng genommen nur für nominal skalierte Variablen Aussagen über die Verlässlichkeit der Codierung. Da nominales Skalenniveau in Inhaltsanalysen häufig vorkommt und da in der Forschungspraxis oftmals auch bei höheren Skalenniveaus die Reliabilitätskoeffizienten für dieses niedrigste Skalenniveau Verwendung finden, kann eine Betrachtung dieses Bereiches ausreichen und weist eine sehr große Reichweite für bereits durchgeführte Inhaltsanalysen auf: Im Folgenden wird gezeigt, dass die Reliabilitätsangaben aller Inhaltsanalysen, die mit den oben genannten Koeffizienten argumentieren und von mehr als zwei Codierern durchgeführt wurden, untereinander nicht vergleichbar sind und somit als Kriterium für eine Verlässlichkeit der wissenschaftlichen Arbeit nicht in Frage kommen.

Dazu wird nach einem kurzen Überblick insbesondere der Einfluss von unterschiedlichen Codiererzahlen auf die Reliabilitätswerte untersucht. Anschließend werden aus den Problemen der Reliabilitätsbestimmung und der Diskussion der Koeffizienten als Gütekriterium für Codierer und/oder Codebuch Schlussfolgerungen gezogen, die in den Vorschlag eines alternativen Reliabilitätskoeffizienten münden. Abschließend fasst ein Fazit die Probleme und Möglichkeiten der Reliabilitätsbestimmungen zusammen. Da die Berechnung von Reliabilitätskoeffizienten insbesondere bei mehr als zwei Codierern ein großes Problem darstellt, wird in einem Exkurs auf der Website des Hans-Bredow-Instituts erläutert, wie mit dem in den Sozialwissenschaften gängigsten Softwarepaket SPSS Reliabilitäten für Multicoder-Studien errechnet werden können (vgl. Kolb, 2004).

2. Typologie von Reliabilitäten

Die Fachliteratur unterscheidet zunächst drei Typen von Reliabilitäten (vgl. Früh, 1991; Krippendorff, 1980: 130f.; Merten, 1995):

1. Intracoder-Reliabilität[7]: Übereinstimmungsmessung des vom selben Codierer zu zwei Zeitpunkten codierten, identischen Materials oder Stabilität der Messung;
2. Intercoder-Reliabilität: Übereinstimmungsmessung zwischen unterschiedlichen Codierern oder Reproduzierbarkeit der Messung;
3. Instrumentelle Reliabilität: die Übereinstimmungsmessung zwischen Forscher- und Codiererteam (vgl. Merten, 1995; Lauf, 2001: 58).[8] Früh (1991: 175 ff.) geht bei der letzten noch einen Schritt weiter, indem er die paarweisen Übereinstimmungen der

7 Die Intracoder-Reliabilität ist von besonderer Bedeutung für Langzeitstudien, da hier Lern- und Gewöhnungseffekte analysiert werden können. Viele Probleme der Vergleichbarkeit von Reliabilitätskoeffizienten tauchen hier jedoch nicht auf, so dass im Folgenden der Schwerpunkt auf Intercoder- und instrumenteller Reliabilität liegen soll.

8 Diese Messung kann technisch als Sonderfall des Intercoder-Reliabilitätstests bezeichnet werden, bei dem es mit der Codierung des Forschers eine Referenz gibt, die als im Sinne der Forschungsfrage „korrekt“ angenommen wird.

einzelnen Codierer mit dem Forscher als Validitätsprüfung bezeichnet. Dieser Typ ist ein Teil von Krippendorffs (1980: 130f.) weiter gefasstem Typ Genauigkeit („accuracy").

Die im Folgenden beleuchteten Probleme der Vergleichbarkeit der verschiedenen Koeffizienten können am Beispiel der Intercoder-Reliabilität verdeutlicht werden, da für alle Typen die Reliabilitätsüberprüfungen zur Errechnung eines Koeffizienten auf übereinstimmenden bzw. nicht übereinstimmenden Codierungen basieren. Dabei spielt es keine Rolle, ob die Codierungen desselben Codierers zu unterschiedlichen Zeitpunkten, die Codierungen verschiedener Codierer und/oder des Forschers miteinander verglichen werden. Zudem werden Intercoder-Reliabilitäten in den meisten Studien verwendet.

Darüber hinaus finden sich in mindestens einer weiteren Dimension unterschiedliche Typen von Reliabilitätskoeffizienten. Es ergibt sich eine Unterscheidungsmöglichkeit nach dem jeweiligen Aggregationsniveau der Angaben:

- auf höchstem Niveau als ein einziger Reliabilitätskoeffizient über alle Codierer, Variablen und Fälle;
- auf mittlerem Niveau als Auswertungen für jeden einzelnen Codierer[9], für jede einzelne Variable oder für jeden einzelnen Fall.[10]

Noch niedrigere Aggregationsniveaus, also z. B. Reliabilitäten für die Codierungen einer Variable von allen Codierern in einem Fall, kurz: für einzelne Codierentscheidungen, machen in der Forschungspraxis wenig Sinn, da die Fallzahlen dann zu gering werden, um gehaltvolle Ergebnisse zu produzieren. Für theoretische und methodologische Betrachtungen soll diese Vorgehensweise aber im Folgenden herangezogen werden, um die Problematik verschiedener Koeffizienten bei mehr als zwei Codierern aufzuzeigen.

3. Reliabilitätsbeeinflussende Faktoren

Die Zuverlässigkeit von Messungen (ausgedrückt in Reliabilitätskoeffizienten) hängt von mindestens fünf verschiedenen Einflussfaktoren ab (vgl. Lauf, 2001; Neuendorf, 2002):

- unterschiedlichen Aggregationsniveaus der Koeffizienten[11];
- unterschiedlichen Skalenniveaus der Variablen (vgl. Neuendorf, 2002);
- unterschiedlichen „Schwierigkeitsgraden" von Variablen (vgl. Wirth, 2001);
- formalen vs. inhaltlichen Variablen[12];
- Anzahl der Codierer (vgl. Abschnitt 4).

---

9 Dies ist natürlich nur im Vergleich zu allen anderen möglich, da in einem Intercoder-Reliabilitätstest per definitionem mindestens zwei Codierer untersucht werden müssen.

10 Dies sind in der Regel bei Inhaltsanalysen beispielsweise einzelne Zeitungsartikel, Medienbeiträge oder auch einzelne Aussagen, je nach dem welche Codiereinheit gewählt wurde.

11 Vgl. die Typologie in Abschnitt 2. In wissenschaftlichen Publikationen wird die Reliabilität häufig über alle einbezogenen Fälle, Variablen und Codierer in einem einzigen Übereinstimmungskoeffizienten angegeben. Eine solche Bestimmung und Darstellung macht das Suchen nach Ausreißern bei einzelnen Codierern, in einzelnen Fällen und in einzelnen Variablen und somit letztlich die intersubjektive Nachvollziehbarkeit unmöglich. Das kann im Extremfall bedeuten, dass Reliabilität angenommen wird, ohne sie tatsächlich, z. B. für jede einzelne Variable, hergestellt zu haben (vgl. ausführlich Lauf, 2001).

12 Dies ist als ein Sonderfall des möglicherweise unterschiedlichen „Schwierigkeitsgrades" von Variablen zu interpretieren.

Das Testen von Intercoder-Reliabilität erfolgt bekanntermaßen, indem alle am Erhebungsprozess Beteiligten dasselbe Material – eine Stichprobe aller zu codierenden Inhalte – bearbeiten. In der Regel werden die paarweisen Übereinstimmungen dann anteilig als Reliabilitätskoeffizient(en)[13] angegeben (vgl. Früh, 1991: 168 ff. und Formel 1).

Formel 1: $$CR = \frac{2 \cdot \ddot{U}}{C_1 + C_2}$$

Quelle: Früh, 1991: 170, mit CR = Codiererreliabilität, Ü = Anzahl der übereinstimmenden Codierungen, C_1 bzw. C_2 = Anzahl der Codierungen von Codierer 1 bzw. Codierer 2

In diesem Verfahren wird davon ausgegangen, dass übereinstimmende Codierungen darauf hinweisen, dass zum einen das Codebuch eine gute Anleitung für die Codierung liefert *und* dass zum anderen die Codierer die Anleitung gut befolgen. Eine große Zahl von übereinstimmenden Codierungen und somit ein hoher Reliabilitätswert muss also als Verbindung von einem präzisen und für alle Fälle gut anwendbaren Codebuch und von ausreichend geschulten und gewissenhaften Codierern interpretiert werden. Ist der Forscher am Reliabilitätstest beteiligt, so kann die Verständlichkeit sowie die Trennschärfe, die Exklusivität und die Vollständigkeit des Codebuchs anhand der Bestimmung von instrumenteller Reliabilität überprüft werden. Wenn „nur" die Intercoder-Reliabilität bestimmt wird, nimmt man für die Auswertung an, dass die Kategorie, die am häufigsten (übereinstimmend) vergeben wurde, diejenige sein dürfte, die der Forscher vergeben hätte, es sich also um eine „korrekte" Codierung handelt.[14]

Bei Bestimmung von Reliabilitäten auf der oben eingeführten, mittleren Ebene, also für einzelne Variablen, Codierer und Fälle, können die Ergebnisse verfeinert betrachtet werden: Tritt bei einem Codierer häufig eine Abweichung von den anderen auf, so kann darauf geschlossen werden, dass dieser die Codierung nicht sorgfältig durchgeführt oder das Codebuch nicht gut genug verstanden hat. Als „qualitätshemmendes" Element wäre in einem solchen Fall also der Codierer ausgemacht.

Wenn sich Abweichungen in einer Variablen häufen, so dürfen die Ergebnisse der Studie für diese Variable streng genommen nicht verwendet werden. Eine solche Situation deutet auf eine besonders schwierige Codierung der spezifischen Variablen hin. Wenn das Codebuch trennscharf, vollständig und exklusiv ist sowie genügend Beispiele und Erklärungen beinhaltet, müssten alle Codierentscheidungen für alle Variablen idealiter mehr oder weniger genau einen Schwierigkeitsgrad haben. Ähnliches gilt für die Häufung von Abweichungen in einzelnen Fällen: Wenn das Codebuch auch für die Analyse dieser offensichtlich schwierigen Fälle erarbeitet worden ist, so müsste es für alle Fälle des Untersuchungsmaterials gleich leichte Codierentscheidungen ermöglichen. Wenn sich also die Codierprobleme in einem bzw. einigen Fällen und in einer bzw. wenigen Variablen häufen, so ist von Schwächen des Codebuches für das gesamte Untersuchungsmaterial auszugehen, auch wenn die Probleme nur wenige Fälle betreffen.

Die methodologische Literatur gibt für eine gute, d. h. in diesem Zusammenhang re-

13 Streng genommen ist die relative Anzahl von paarweisen Übereinstimmungen natürlich nur ein möglicher Reliabilitätskoeffizient, dessen Entwicklung Holsti (1969) zugeschrieben wird. Da dieser besonders einfach zu bestimmen und somit am besten intersubjektiv nachzuvollziehen ist, wird die folgende Argumentation auf Holsti begründet. Auf andere Koeffizienten wird im Folgenden noch eingegangen.

14 Es spielt dabei keine Rolle, ob es eine korrekte Codierung überhaupt geben kann, eine weiterführende Diskussion dieser erkenntnistheoretischen Gedanken würde hier zu weit führen.

liable Codierung häufig faustregelartig einen Wert oder Wertebereich – z. B. 0,8 – an[15]. Solche Werte sind noch nicht einmal plausibel begründbar, geschweige denn theoretisch bzw. statistisch fundiert (vgl. Krippendorff, 1980: 133). 80 Prozent übereinstimmende Codierungen, denn nichts anderes bedeutet der Holsti-Koeffizient 0,8[16], können in bestimmten Fällen nicht als unproblematisch angesehen werden, wie im Folgenden noch gezeigt wird (vgl. Krippendorff, 1980: 134; Lauf, 2001: 59 ff. und Neuendorf, 2002: 143). Der einfache Holsti-Koeffizient birgt zudem das Risiko von zufälligen („falschen") Übereinstimmungen[17], die als reliable Codierung in den Koeffizienten eingerechnet werden.

In diesen ersten Überlegungen wird bereits deutlich, dass (Intercoder-)Reliabilitätskoeffizienten nur den Charakter von Schätzungen haben können. Es sind auch Szenarien denkbar, bei denen kein Codierer die „richtige" Codierung vorgenommen hat. Wenn es in einem solchen Fall Übereinstimmungen gibt, wären diese „falsch", schlagen sich jedoch im Koeffizienten nieder. Lässt man sich also auf die Bestimmung von Reliabilitätskoeffizienten zur Bestimmung der Güte einer Inhaltsanalyse ein, so ist die Gefahr eines Fehlschlusses bezüglich der Qualität der Codierung und des Codebuches nicht gänzlich auszuschließen. Es ist nur bei denjenigen Codierentscheidungen wahrscheinlicher, dass es sich um eine gute Codierung handelt bzw. dass „richtig" codiert wurde, bei denen möglichst viele Codierer übereinstimmend codiert haben.

4. Probleme bei mehr als zwei Codierern

Eine Auseinandersetzung mit Problemen bei Reliabilitätstests mit mehr als zwei Codierern ist m. W. bisher noch nicht erfolgt. Daher soll im Folgenden ein Beitrag zum Schließen dieser Forschungslücke geleistet werden. Neben der Komplexität der *Berechnung* von Reliabilitätskoeffizienten für Codiererteams mit mehr als zwei Codierern[18] bergen die *Ergebnisse* der Reliabilitätsauswertungen auch theoretische bzw. methodologische Probleme. Diese hängen mit der Anzahl der einzelnen, d. h. paarweisen Übereinstimmungsüberprüfungen zusammen, die mit steigender Codiererzahl ebenfalls – und dies *nicht linear* – ansteigt (vgl. Abbildung 1).

---

15 Für einen Überblick über verschiedene Faustregeln vgl. Neuendorf (2002: 142f.) Die Autorin geht davon aus, dass nur ein Wert von mindestens 0,9 immer akzeptabel ist, obwohl auch bei einem solchen Wert eine Unsicherheit bleibt. Für komplexere Koeffizienten, die die zufälligen Übereinstimmungen statistisch herausrechnen, sollten dann aber nach Neuendorf geringere Grenzen gelten. Damit spricht sie sich zwar für die komplexeren Verfahren aus, schränkt aber das Gütekriterium gleich wieder ein, so dass man bei starkem Herabsetzen des „Grenzwertes" schlussendlich den Teufel mit dem Beelzebub ausgetrieben hat.

16 Dies gilt zunächst nur für den Vergleich von zwei Codierern. Für die Probleme bei mehr als zwei Codierern wird diese Aussage im Folgenden relativiert.

17 Diese kommen z. B. zustande, wenn in einem Fall zwei Codierer einen Wert und zwei andere einen anderen Wert übereinstimmend codieren. Im Sinne der Forschungsfrage bzw. des Forschers und des theoretischen Ideals von exklusiven und trennscharfen Kategorien „richtig" kann aber nur eine der beiden übereinstimmenden Codierungen sein.

18 Vgl. zu Formelbeschreibungen von Reliabilitätsberechnungen bei mehreren Codierern Fleiss (1971). Es ist allerdings zu beachten, dass die Formel 2 (Fleiss, 1971: 379) einen Fehler beinhaltet. Das letzte „n" in der Formel muss ein „n_{ij}" sein. Verlässliche Computerprogramme zur Auswertung solcher Tests liegen m. W. bislang nicht vor.

Abbildung 1: Anzahl der nötigen Übereinstimmungsprüfungen bei einer Codierentscheidung

Anzahl Codierer	Übereinstimmungsprüfung zwischen Codierern	Anzahl der Übereinstimmungsprüfungen
2	1 und 2	1
3	1 und 2	3
	1 und 3	
	2 und 3	
4	1 und 2	6
	1 und 3	
	1 und 4	
	2 und 3	
	2 und 4	
	3 und 4	

Betrachtet man zunächst eine einzige Codierentscheidung[19] von zwei Codierern, so gibt es nur eine Übereinstimmungsprüfung. Bei drei Codierern werden für jede einzelne Codierentscheidung genau drei Überprüfungen gemacht, bei vier Codierern sind schon sechs Überprüfungen notwendig. Dadurch ergibt sich logischerweise jeweils eine andere Basis des (wie gesagt nur theoretisch interessanten) „Reliabilitätswertes" für diese einzelne Codierentscheidung.

In der folgenden Überlegung wird vom Idealfall ausgegangen, dass alle Codierer in ihrer Codierung dieses einen Falls und dieser einen Variablen übereinstimmen. Weicht nun aber ein Codierer von den übereinstimmenden Codierungen ab, so ist der Reliabilitätskoeffizient nach der grundlegenden Formel 1 für diese *einzelne* Codierentscheidung bei insgesamt zwei Codierern 0, bei dreien 0,33 und bei vieren 0,5 (vgl. auch Abbildung 2).

Abbildung 2: Reliabilitätskoeffizienten bei unterschiedlichen Codiererzahlen[20]

Anzahl Codierer	Anzahl der Abweichungen	Anzahl der Übereinstimmungen	Reliabilitätskoeffizient
2	0	1	1
	1	0	0
3	0	3	1
	1	1	0,33
	2	0	0
4	0	6	1
	1	3	0,5
	2	1	0,17
	3	0	0

19 Unter einer Codierentscheidung wird hier und im Folgenden verstanden, dass in einem Fall eine Variable codiert wird. Das bedeutet, dass jeder Codierer sich einmal entscheiden muss, ob und was codiert werden muss.

20 Bei mehr als drei Codierern können theoretisch Übereinstimmungen in mehreren Variablenausprägungen auftreten, d. h. es kann ab vier Codierern zu mehreren Pärchen von jeweils

Eine abweichende Codierung fällt also sehr unterschiedlich ins Gewicht, wenn zwei, drei oder vier Codierer beteiligt sind. Dies ist auf den ersten Blick und an dieser Stelle noch nicht weiter problematisch, da mit steigender Codiererzahl bei „nur" einer Abweichung der Reliabilitätswert ebenfalls steigt (0 bei zwei Codierern, 0,33 bei dreien und 0,5 bei vieren). Im Folgenden bildet diese triviale Erkenntnis jedoch die Grundlage der weiteren Argumentation. Wenn man den Fokus etwas weiter fasst und zwei Codierentscheidungen betrachtet, so können sich sogar bei ein und derselben Codiererzahl unterschiedliche Reliabilitätswerte für beispielsweise zwei Abweichungen ergeben. In Abbildung 3 sind für diesen Fall zwei mögliche Codeblätter dargestellt, bei denen im ersten Beispiel die zwei Abweichungen bei ein und derselben Codierentscheidung auftreten, während sie sich im zweiten Beispiel auf zwei Codierentscheidungen aufteilen.

Abbildung 3: Codierblätter für vier Codierer C1-C4, zwei Codierentscheidungen und zwei abweichende Codierungen (abweichende Codierungen kursiv)

Beispiel 1					
Codierentscheidung	C1	C2	C3	C4	Reliabilitätskoeffizient
1	3	3	3	3	1
2	*1*	*2*	*3*	2	0,17
gesamt:					0,58
Beispiel 2					
Codierentscheidung	C1	C2	C3	C4	Reliabilitätskoeffizient
1	*2*	3	3	3	0,5
2	2	*1*	2	2	0,5
gesamt:					0,5

Es gibt in beiden Beispielfällen also innerhalb einer Variablen die gleiche Anzahl an abweichenden Codierungen. Berechnet man für diese beiden Beispiele den Reliabilitätswert für jede Codierentscheidung auf der Basis paarweiser Übereinstimmungen und bildet danach – wie es für die Bestimmung von Reliabilitäten einzelner Variablen üblicherweise gemacht wird – den Mittelwert, so erhält man unterschiedliche Reliabilitätswerte, nämlich 0,58 für Beispiel 1 und 0,5 für Beispiel 2 (vgl. Abbildung 3).

In Abbildung 4 werden zur Vertiefung beispielhaft für vier Codierer die möglichen Reliabilitätswerte angegeben, die bei ein bis sechs abweichenden Codierungen und zwei Codierentscheidungen auftreten können.

Die Höhe des Reliabilitätskoeffizienten hängt davon ab, ob die abweichenden Codierungen bei einer oder bei mehreren Codierentscheidungen aufgetreten sind. Hinzu kommt, dass höhere Reliabilitätswerte erreicht werden, wenn viele Fehler in einem Fall auftreten. Verteilen sich die Abweichungen auf viele Fälle, so erhält man niedrigere Werte. Wie in Abbildung 4 ersichtlich ist, kann es sogar ohne Auswirkung auf den Reliabilitätswert bleiben, wenn ein „Fehler" mehr gemacht wird: Der Übereinstimmungswert

übereinstimmenden Codierungen kommen. Bei vier Codierern kann z. B. die Darstellungsform eines Beitrags zweimal als Meldung und zweimal als Nachricht codiert werden. Dieser Sonderfall beeinflusst die Anzahl der Übereinstimmungen, der Abweichungen und somit auch den Koeffizienten. Auch dieses Problem kann im Folgenden durch den neuen Koeffizienten verkleinert werden.

Abbildung 4: Reliabilitätskoeffizienten bei vier Codierern und zwei Codierentscheidungen

Anzahl der Abweichungen	bei x Codierentscheidungen	Reliabilitätskoeffizient
1	1	0,75[21]
	2	nicht möglich
2	1	0,58[22]
	2	0,5
3	1	0,5
	2	0,33
4	1	nicht möglich
	2	0,17 oder 0,25[23]
5	1	nicht möglich
	2	0,08
6	1	nicht möglich
	2	0

Abbildung 5: Beschreibung der Beispiele 3 und 4

Allgemeines	Angenommen, es liegen für eine Variable 50 Fälle im Reliabilitätstest vor, die alle von vier Codierern bearbeitet worden sind.
	Ein nach gängigen Faustregeln als akzeptabel einzustufender Holsti-Koeffizient von 0,8 bedeutet für dieses Beispiel, dass zwischen 20 (Beispiel 3) und 30 (bzw. 40[24], Beispiel 4) abweichende Codierungen vorgenommen worden sind:
Beispiel 3	Wenn alle Abweichungen in unterschiedlichen Fällen auftreten, „erlaubt" der Grenzwert von 0,8 lediglich 20 Abweichungen:
	Dieser Wert ergibt sich aus 30 vollständig übereinstimmenden Codierungen und 20 Fällen, in denen jeweils ein Codierer abweicht. Berechnung: $(30 \cdot 1 + 20 \cdot 0{,}5)/50 = 0{,}8$.
Beispiel 4	Bei Häufungen in wenigen Fällen werden bis zu 30 Fehler mit einem Reliabilitätskoeffizienten von 0,8 zugelassen:
	Dieser Wert wird bestimmt durch 40 vollständig übereinstimmende Codierungen. In den restlichen zehn Fällen weichen alle Codierer voneinander ab, so dass sich insgesamt 30 Abweichungen ergeben. Berechnung: $(40 \cdot 1 + 10 \cdot 0)/50 = 0{,}8$.

21 Der Wert ergibt sich als Mittelwert aus 0,5 (eine Abweichung bei einer Codierentscheidung) und 1 (keine Abweichung bei der zweiten Codierentscheidung).

22 Die Werte werden hier zur einfacheren Darstellung ohne mögliche, zufällige Übereinstimmungen ausgewiesen. So können z. B. bei zwei ermittelten Abweichungen bei einer Codierentscheidung beide abweichenden Codierer denselben Code vergeben haben, also untereinander übereinstimmen. Dadurch würde der einfache Holsti-Koeffizient höher als hier angegeben: 0,67 statt 0,58.

23 Die unterschiedlichen Werte ergeben sich hier daraus, dass entweder bei beiden Codierentscheidungen zwei Abweichungen oder bei einer Entscheidung eine und bei der anderen Entscheidung drei Abweichungen auftreten können.

24 Die abweichenden Codierungen haben pro codierten Fall nur drei Freiheitsgrade: Wenn drei Codierungen abweichen, muss die vierte automatisch auch abweichen. Im Folgenden sollen nur höchstens drei Abweichungen pro Fall betrachtet werden.

für zwei Abweichungen in zwei Fällen und für drei Abweichungen in einem Fall liegt jeweils bei 0,5. Um die Tragweite dieser problematischen Erkenntnis zu verdeutlichen, werden hier weitere Beispiele angeführt, die praxisnäher sind als zwei Codierentscheidungen (vgl. Abbildung 5).

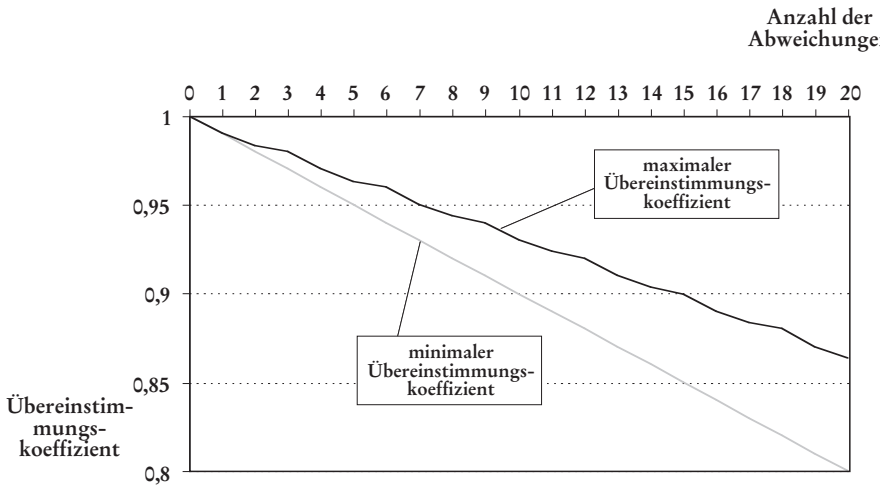
Bilden die Reliabilitätskoeffizienten für die Beispiele 3 und 4 aber tatsächlich gleichwertige bzw. gleich gute Codierungen ab? Das Beispiel 4 beschreibt letztlich, dass in (zehn) bestimmten Fällen, also in der Regel bei bestimmten Artikeln, die erforderlichen Codierentscheidungen offenbar nicht eindeutig getroffen werden *können*: Die Codierer können also z. B. nicht entscheiden, ob Kernenergie als Thema unter Umwelt oder unter Energie gefasst werden muss. Man könnte von systematischen Fehlern bei bestimmten, schwierigen Codierungen sprechen. Dies weist recht deutlich auf ein unzureichendes Codebuch hin, das für bestimmte Fälle keine oder nur mangelhafte Entscheidungskriterien liefert, also nicht trennscharf ist.

Schlechtere Werte ergeben sich, wie in Abbildung 4 verdeutlicht wurde, wenn in verschiedenen Fällen je ein Fehler auftritt. Beispiel 3 veranschaulicht das mit der geringeren Anzahl (20) von Abweichungen bei identischem Reliabilitätskoeffizienten im Vergleich zu Beispiel 4. Diese unsystematischen Fehler können selbst bei gewissenhaftester Codierung auftauchen, es ist keine spezifische Schwäche des Codebuches zu erkennen. Wenn die Abweichungen immer beim gleichen Codierer vorkommen, könnte es sich – wie angedeutet – um ein mangelndes oder fehlerhaftes Verständnis des Codebuchs oder schlampige Arbeit handeln.

Die gängigen Verfahren begünstigen also Mängel in der Trennschärfe, in der Exklusivität und in der Vollständigkeit des Codierbuchs gegenüber vereinzelten Abweichungen der Codierer. Dies erscheint insbesondere problematisch, weil der Reliabilitätstest vor Beginn der eigentlichen Untersuchung zusammen mit Codiererschulungen zu einem möglichst gut verständlichen, intersubjektiv nachvollziehbaren Codebuch führen soll. Daran schließt sich die normative Frage an, ob in Reliabilitätsüberprüfungen nun die Qualität der Codierer oder die des Kategoriensystems gemessen werden bzw. ob für eine Inhaltsanalyse die Qualität der Codierer oder die des Codebuchs wichtiger sein soll. In beiden Fällen ist m. E. eindeutig dem Kategoriensystem bzw. dem Codebuch Vorrang zu geben, da in Studien mit mehreren Codierern ja implizit von einer Austauschbarkeit der Codierenden ausgegangen wird und nur die systematische Anleitung zu verbesserten Codierungen führen kann.

Es stellt sich nun die grundsätzliche Frage, ob jede einzelne Abweichung nicht für eine bessere Transparenz der Reliabilitätswerte mit *gleichem* Gewicht in die Auswertung eingehen sollte. Dagegen könnte ins Feld geführt werden, dass eine Konzentration von Fehlern in wenigen Fällen für die anstehende Studie große Vorteile mit sich bringt. Geht man von der übereinstimmenden als der korrekten bzw. gewünschten Codierentscheidung aus, so könnten im Beispiel 3 schlimmstenfalls nur 30 Fälle als richtig codiert und somit als aussagekräftig bezeichnet werden, während im Beispiel 4 mit 40 richtig codierten Fällen weniger Fehler aufgetreten sind. Allerdings unterscheiden sich in den beiden Beispielen für die 20 bzw. 10 übrigen Fälle die Wahrscheinlichkeiten für eine richtige Codierung deutlich. Während in Beispiel 3 drei von vier Codierern eine richtige Entscheidung getroffen haben, die Wahrscheinlichkeit also bei 75 Prozent liegt, so ist die Wahrscheinlichkeit für eine korrekte Codierung im Beispiel 4 gleich der zufälligen Codierwahrscheinlichkeit der richtigen Ausprägung, also 20 Prozent. Zufallsverteilt ergäben sich im Beispiel 3 also drei Viertel von 20 fraglichen Fällen oder mit großer Wahrscheinlichkeit 15 weitere korrekt codierte Fälle, insgesamt also 45 von 50. Für das Beispiel 4 lassen sich zusätzlich zu den 40 korrekten Codierungen nur weitere 2 richtig co-

Abbildung 6: Maximale und minimale Reliabilitätskoeffizienten nach Anzahl der abweichenden Codierungen (vier Codierer und 0–20 Abweichungen)



dierte Fälle annehmen, nämlich 20 Prozent von 10 fraglichen Fällen. Man käme also auf 42 richtig codierte Fälle in diesem vierten Beispiel.[25]

In Abbildung 6 wird zur Unterstützung der Argumentation gegen ein Vorziehen des unpräzisen Codebuches und zur Veranschaulichung beispielhaft an den Daten von Beispiel 3 und 4 dargestellt, wie sich die Reliabilitätskoeffizienten bei zunehmender Zahl von Abweichungen entwickeln.

Es öffnet sich mit schlechter werdenden Koeffizienten eine Schere zwischen maximaler und minimaler Fehlerzahl, die bei einem bestimmten Reliabilitätswert erreicht werden. Umgekehrt zeigt sich in der Grafik deutlich, dass bei 20 Abweichungen von 4 Codierern in 50 Fällen der Reliabilitätswert zwischen 0,8 und 0,86 schwanken kann. Dieser Effekt könnte nun aufgrund des simplen Holsti-Koeffizienten zustande kommen, aber die beiden gängigen, von zufälligen Übereinstimmungen „bereinigten" Koeffizienten Cohen's Kappa und Scott's Pi[26] ergeben auch identische Werte für die Beispiele 3 und 4, weil sie auf paarweisen Übereinstimmungen beruhen (vgl. Carletta, 1996): Beide ergeben für 20 und 30 Abweichungen einen etwas niedrigeren Wert von 0,75[27].

25 Da diese Wahrscheinlichkeitsargumentationen stark davon abhängen, wie viele Ausprägungen die untersuchte Variablen haben und ob die unterschiedlichen Ausprägungen die gleiche Wahrscheinlichkeit haben, codiert zu werden, soll diese hier nicht weiter ausgeführt werden.

26 Gwet (2002b) kritisiert Kappa- und Pi-Auswertungen auch für zwei Codierer als instabil und schwer interpretierbar. Das statistische Herausrechnen von zufälligen Übereinstimmungen ist stark abhängig von den Wahrscheinlichkeiten, mit denen jede Codierausprägung im Datensatz vorkommt (trait prevalence).

27 Mit dem in Kolb (2004) erläuterten SPSS-Verfahren stellt sich heraus, dass sich die Kappa-Werte marginal unterscheiden: Beispiel 3: 0,74985930, Beispiel 4: 0,75000000. Wenn man die Kappa-Werte wirklich so genau betrachten möchte, stellt sich heraus, dass das weniger problematische Beispiel 3 dabei sogar den schlechteren Wert bekommt.

Spearman's Rho ergibt ebenfalls identische Werte für beide Szenarien: 0,998. Die korrelationsbasierten Werte von Pearson und Lin sind streng genommen nicht anwendbar, da sie intervall- oder metrisch skalierte Daten erfordern (vgl. Neuendorf, 2003). Fasst man die fünfstufige Beispielvariable – wie z. B. bei Likert-Skalen gängig – als intervallskaliert auf, so ergeben diese Koeffizienten einen Unterschied in der dritten Nachkommastelle (0,957 gegenüber 0,955) und weisen tatsächlich der problematischeren Codierung mit 30 Abweichungen den (leicht) schlechteren Wert zu.[28] Sogar der sehr angesehene, aber extrem kompliziert zu berechnende Koeffizient von Krippendorff (1980) ergibt für beide Beispiele identische Alphawerte von 0,75[29], obwohl er nicht auf paarweisen Übereinstimmungen beruht. Die Alpha-Berechnung erfolgt allerdings durch die Betrachtung von paarweisen Abweichungen, so dass das Problem lediglich verlagert wird.[30]

Als ein Spezialfall solcher Interpretationsschwierigkeiten können die Werte für instrumentelle Reliabilität gesehen werden: Für ein Codiererteam von vier Studentinnen und Studenten plus Forscher ergeben sich bei paarweisen Vergleichen für einen Intercoder-Reliabilitätstest theoretisch aus zehn Vergleichen auch zehn mögliche Übereinstimmungen innerhalb einer Variablen. Die Abweichung eines Codierers (von allen vier anderen) bedeutet also eine Reduktion der paarweisen Reliabilität von 1, d. h. zehn Übereinstimmungen, um vier Übereinstimmungen oder 0,4 auf 0,6 (sechs Übereinstimmungen), während sie in der instrumentellen Reliabilitätsprüfung durch die neue Basis von vier Vergleichen mit der Referenzcodierung des Forschers nur zu einer Reduktion von 1 auf 0,75 führt (vgl. Lauf, 2001: 58f.). Daher kann m. E. für beide nicht dieselbe Faustformel gültig sein.[31]

---

28 Die Berechnungen für die komplexeren Reliabilitätskoeffizienten wurden mit dem frei verfügbaren Tool PRAM (download von http://www.geocities.com/skymegsoftware/pram.html) durchgeführt. Das Tool ist unter Mitarbeit von Kimberley Neuendorf entwickelt worden und verwendet Excel-Daten zur Berechnung. Leider ist es sehr fehleranfällig und hat zum Teil einschränkende Voraussetzungen. Es unterstützt keine Berechnungen für codierte Werte mit Dezimalstellen oder Werte, die größer als 32768 sind. Die integrierte Readme-Datei gibt allerdings die Empfehlung, keine Werte größer 100 zu verwenden. Außerdem verarbeitet PRAM höchstens 8192 Zellen, die sich aus dem Produkt der Anzahlen von Codierern, Variablen und Fällen errechnet. Eine Beschreibung findet sich auch bei Neuendorf (2003).

29 Auch bei Krippendorffs Alpha ergaben sich wieder marginale Unterschiede: hier in der vierten Nachkommastelle (0,75125 vs. 0,75111). Die Berechnung von Krippendorffs Alpha wurde mit einer eigens entwickelten SPSS-Syntax durchgeführt, die online dokumentiert und erläutert wird (vgl. Kolb, 2004).

30 Vgl. Abbildungen 2 und 4: In unserem Beispiel von vier Codierern ergibt die erste abweichende Codierung in einem Fall logischerweise drei paarweise Abweichungen, während die zweite abweichende Codierung nur zu zwei weiteren paarweisen Abweichungen und die dritte Abweichung, die gleich bedeutend mit der vierten Abweichung ist, nur zum Verschwinden der letzten paarweisen Übereinstimmung führt. Daraus ergibt sich auch der oben geschilderte Fall, dass 20 und 30 Abweichungen den gleichen Reliabilitätskoeffizienten haben können: die erste Abweichung in einem Fall und die zweite und dritte (gleichzeitig auch die vierte) Abweichung in einem Fall reduzieren den Reliabilitätswert pro Codierentscheidung in gleichem Umfang, nämlich um 0,5.

31 Darüber hinaus wäre zu diskutieren, ob die Voraussetzung einer solchen Messung, dass der Forscher quasi unfehlbar bei der Codierung ist, unproblematisch ist. Darauf kann hier nicht weiter eingegangen werden.

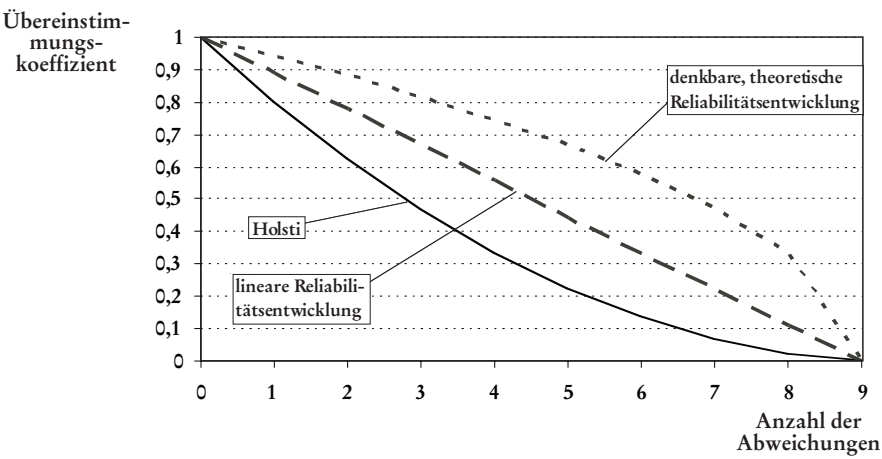
5. Schlussfolgerungen und Lösungsansätze

Schon die Berechnung oder Auszählung paarweiser Übereinstimmungen stellt bei umfangreichen Inhaltsanalysen und/oder großen Codiererteams einen gewaltigen Arbeitsaufwand dar. Unkomplizierte und fehlerfrei arbeitende Reliabilitätsberechnungs-Software gibt es insbesondere für mehr als zwei Codierer so gut wie nicht. Darüber hinaus verwenden die meisten Programme unterschiedliche Datenformate (ASCII, Excel, SPSS, SAS etc.) und -formatierungen (Fallebene, Variablenebene).

Generell lassen sich daneben für alle gängigen Reliabilitätskoeffizienten unterschiedlicher Komplexität eine ganze Reihe von mehr oder weniger großen Schwierigkeiten für die Verwendung und Interpretation der Werte konstatieren. Man sollte daraus schlussfolgern, dass sogar ein Ausweisen verschiedener Koeffizienten und aller weiteren von Lauf (2001: 67) geforderten Mindestanforderungen an empirische Studien keine leicht zu interpretierenden Werte ergibt. Auch die Vergleichbarkeit von Reliabilitäten einzelner Variablen ist nur sehr eingeschränkt gegeben, wenn – wie gezeigt werden konnte – z. T. „bessere" Codierungen bzw. Codebücher „schlechtere" Reliabilitätswerte aufweisen. In jedem Falle müssen bei der Dateninterpretation die Fehlerwahrscheinlichkeiten und ihr „Durchschlag" im Koeffizienten berücksichtigt werden. Denkbar zur Lösung dieser Probleme wären gewichtete oder auf anderer Basis berechnete Reliabilitätskoeffizienten. Leider können auch vorliegende gewichtete Kappa-Statistiken nicht als gute Lösung angesehen werden, da sie ebenfalls auf paarweisen Übereinstimmungen beruhen. Auch die neueren Weiterentwicklungen auf dem Gebiet der Reliabilitätskoeffizienten z. B. von Gwet (2001, 2002a, b, c) konzentrieren sich zumeist auf die Verbesserung der Gewichtung bzw. des Herausrechnens der zufälligen Übereinstimmungen.

Für ein besseres Überschauen des Problemfeldes kann ein Blick auf die Reliabilitäts- bzw. Abweichungsentwicklung bei größeren Codiererteams geworfen werden. In Abbildung 7 sind unterschiedliche Reliabilitätstypen beispielhaft für 10 Codierer in einem

Abbildung 7: Reliabilitätswerte bei 10 Codierern in einer Codierentscheidung



Fall, also bei einer Codierentscheidung, die von 10 Codierern getroffen wurde, abgetragen. Die durchgezogene Linie zeigt dabei den Verlauf des Holsti-Koeffizienten, der – wie gezeigt wurde – auch den komplexeren Koeffizienten zugrunde liegt. Es zeigt sich, dass solche Koeffizienten eine Reliabilitätskurve unterhalb der linearen (lang gestrichelten) Linie aufweisen. Das bedeutet, dass die ersten Fehler stärker ins Gewicht fallen als die späteren, weil die Kurve erst sehr stark fällt und das Gefälle (die negative Steigung) zum Ende hin (im Betrag) geringer wird. Theoretisch wären zwei Szenarien für die Reliabilitätsentwicklung denk- und begründbar. Für eine möglichst große Transparenz der Reliabilitätswerte sollte jede Abweichung gleich viel zählen, oder genauer: den Reliabilitätskoeffizienten um den gleichen Betrag reduzieren. Dieser Fall ist mit der linearen Reliabilitätsentwicklung in Abbildung 7 lang gestrichelt dargestellt. Eine weitere denkbare Auffassung der Entwicklung von Reliabilitäten mit zunehmenden Fehlern pro Fall könnte beinhalten, dass jeder weitere Fehler stärker ins Gewicht fällt. Dies ist plausibel, weil die Codierung des Falles desto problematischer wird, je mehr abweichende Codierungen vorkommen. Ein Vorschlag für eine solche gewichtete Kurve ist kurz gestrichelt in Abbildung 7 eingezeichnet. Dieser Vorschlag beruht auf einer einfachen Gewichtung der linearen Kurve durch Wurzelziehen.

Die lineare Kurve lässt sich auf den ersten Blick leicht bestimmen. Man muss nur auf die Daten schauen und die Zahl der abweichenden Codierungen zählen. Dies setzt – so könnte man kritisieren – allerdings voraus, dass die korrekte Codierung als Referenz bekannt ist. Die gleiche Einschränkung gilt jedoch auch für die paarweise Übereinstimmungsmessung, da auch hier implizit davon ausgegangen wird, dass die Codierung mit den meisten Übereinstimmungen die „richtige" ist. Die *Berechnung* des linearen Reliabilitätskoeffizienten für mehr als zwei Codierer[32] ist nicht sofort evident. Hierzu muss für jeden einzelnen Codierer in einem Fall geprüft werden, mit wie vielen anderen Codierern er übereinstimmt. Wenn dieser Wert 0 annimmt, der Codierer also mit niemandem übereinstimmt, so erhält man eine abweichende Codierung. Ist dieser Codierer in obigem Beispiel der Einzige mit 0 Übereinstimmungen zu anderen Codierern, so werden alle anderen Codierer idealiter 8 übereinstimmende Codierungen zu den jeweils anderen haben.

Auch dieser Koeffizient birgt das Risiko, dass die Codierer mit mehr als 0 Übereinstimmungen sich z. B. in zwei Gruppen teilen, d. h. 4 von ihnen in einer Codierungsausprägung und 5 in einer anderen Ausprägung übereinstimmen. Dies bedeutet aber keine Verschlechterung zum Status quo, da die Übereinstimmungskoeffizienten dieses Problem auch haben. Allerdings wird dies für den neuen ‚Abweichungskoeffizienten' problematisch, wenn sich alle Codierer in Gruppen aufteilen, das heißt jeder mit mindestens einem anderen übereinstimmt. Man kann jedoch, um dieses Problem zu lösen, den Maximalwert der Übereinstimmungen der Codierer mit allen anderen bestimmen. Dieser müsste in unserem Beispiel bei 9 liegen, wenn es keinen Codierer mit 0 Übereinstimmungen gibt. Denkbar wäre auch, dass von den 10 Codierern 3 miteinander in einer Ausprägung und 7 in einer anderen übereinstimmen. Betrachtet man dann den Maximalwert der Übereinstimmungen je Codierer, so wäre dieser 6 – bei einem theoretischen Maximalwert 9 (Anzahl der *m* Codierer – 1 Freiheitsgrad *df*). Ein neuer linearer Reliabilitätskoeffizient α_1 für einen Fall lässt sich also berechnen, indem die maximale

32 Bei zwei Codierern gleicht sie der allgemein üblichen paarweisen Berechnung von Übereinstimmungen.

Anzahl der übereinstimmenden Codierungen ($\ddot{U}_{max\ emp}$) erhoben wird und durch den theoretischen Maximalwert dividiert wird (s. Formel 2).[33]

Formel 2: $$\alpha_1 = \frac{\ddot{U}_{max\,emp}}{m - df}$$

Für die Gesamtreliabilität α_g über alle n Fälle gilt Formel 3.

Formel 3: $$\alpha_g = \frac{\sum_1^n \alpha_1}{n} = \frac{1}{n} \cdot \frac{\ddot{U}_{1max\,emp} + \ddot{U}_{2max\,emp} + \ddot{U}_{3max\,emp} + \ldots + \ddot{U}_{nmax\,emp}}{m - df}$$

Für die viel zitierten Beispiele mit vier Codierern, einer Variable mit fünf möglichen Ausprägungen und 50 Fällen ergibt sich folgendes Bild. Im Beispiel 3, bei dem in 20 Fällen ein Codierer abweichend codiert hat und in 30 Fällen alle Codierer übereinstimmen, ergibt sich aus Formel 2 für jeden einzelnen der 20 „fehlerhaften" Fälle das Ergebnis 1: 0,67.

Ergebnis 1: $$\alpha_{a1} = \frac{2}{4-1} = 0,\bar{6}$$

Für jeden der 30 komplett übereinstimmenden Fälle ergibt sich aus Formel 2 logischerweise wie bei den gängigen Berechnungsarten der Wert 1 (vgl. Ergebnis 2).

Ergebnis 2: $$\alpha_{b1} = \frac{3}{4-1} = 1$$

Für das vierte Beispiel ergeben sich die Reliabilitäten der Einzelfälle, die in den Ergebnisformeln 3 und 4 dargestellt werden.

Ergebnis 3: $$\alpha_{a2} = \frac{0}{4-1} = 0$$

Ergebnis 4: $$\alpha_{b2} = \frac{3}{4-1} = 1$$

Summiert man diese (Teil-) Ergebnisse 3 und 4 nun nach Formel 3 auf, so ergibt sich als Ergebnis für das Beispiel 3 – wie in Ergebnis 5 gezeigt – 0,87 und für Beispiel 4 – wie in Ergebnis 6 dargestellt – 0,8.

Ergebnis 5: $$\alpha_{g1} = \frac{20 \cdot \alpha_{a1} + 30 \cdot \alpha_{b1}}{50} = \frac{20 \cdot 0,\bar{6} + 30 \cdot 1}{50} = 0,8\bar{6}$$

33 Damit lässt sich auch das Problem der doppelten Übereinstimmungen beheben. Der Holsti-Koeffizient für den beschrieben Fall beträgt nämlich nur 0,53. Der neu berechnete Übereinstimmungskoeffizient dagegen 0,67, was ausgehend von der Annahme, dass 7 Codierer richtig liegen, der plausibler zu interpretierende Wert ist.

Ergebnis 6: $\alpha_{g2} = \frac{10 \cdot \alpha_{a2} + 40 \cdot \alpha_{b2}}{50} = \frac{10 \cdot 0 + 40 \cdot 1}{50} = 0{,}8$

Diese Ergebnisse sind nach den Beschreibungen der Daten (20 Abweichungen vs. 30 bzw. 40 Abweichungen) besser interpretierbar, das weniger problematische Beispiel 3 bekommt den höheren Reliabilitätswert.

Dieser einfache Reliabilitätskoeffizient muss allerdings analog zur Kritik des Holsti-Koeffizienten kritisiert werden: Zufällige Übereinstimmungen werden hier nicht wie bei Scotts Pi, Cohens Kappa oder Krippendorffs Alpha herausgerechnet. Allerdings bietet er die gleichen Möglichkeiten des statistischen Herausrechnens wie andere Koeffizienten. Hier gibt es grundsätzlich zwei verschiedene Herangehensweisen, von denen Krippendorff (1980) die komplexere und genauere wählt. Diese soll im zweiten Schritt präsentiert werden. Im ersten Schritt lässt sich die zufällige Übereinstimmungswahrscheinlichkeit auch mehr oder weniger grob schätzen. Man kann den Codierprozess wahrscheinlichkeitstheoretisch als „Ziehung mit Zurücklegen" auffassen, da bei jedem Codiervorgang alle Ausprägungen zur Verfügung stehen.[34] Wenn eine Variable k Ausprägungen hat, so ist die Wahrscheinlichkeit p_k, dass jede weitere Ziehung den gleichen Wert ergibt, in Formel 4 ersichtlich:

Formel 4: $p_k = \frac{1}{k}$

In unserem Beispiel mit einer Variable mit fünf Ausprägungen folgt also eine zufällige Übereinstimmungswahrscheinlichkeit von 0,2 (s. Ergebnis 7).

Ergebnis 7: $p_5 = \frac{1}{5} = 0{,}2$

Eine Gewichtung des einfachen Alphawertes müsste folgerichtig die zufällige Übereinstimmungswahrscheinlichkeit vom ermittelten Reliabilitätswert abziehen. Dadurch verliert der Koeffizient aber seine Normierung auf den Wertebereich von 0 bis 1. Um diese (zumindest zum Teil) wieder herzustellen, muss der bereinigte Wert durch die zufällige Wahrscheinlichkeit der Abweichung geteilt werden. Vollständige Übereinstimmungen ergeben dann den Wert 1. Die untere Normierungsgrenze fällt trotzdem weg, dies ist allerdings leicht interpretierbar, da mögliche negative Reliabilitätskoeffizienten einen Wert bedeuten, der unterhalb der zufälligen Übereinstimmung liegt (vgl. Formel 5).

Formel 5: $\alpha_{gew} = \frac{\alpha - p_k}{1 - p_k}$

Für unsere zwei Beispielfälle ergeben sich also einfach gewichtete Alpha-Werte von 0,83 (Beispiel 3, vgl. Ergebnis 8) bzw. 0,75 (Beispiel 4, vgl. Ergebnis 9).

---

34 Vgl. ursprünglich z. B. Bennett, Alpert und Goldstein (1954); Holley und Guilford (1964) oder Maxwell (1977) sowie den Überblick von Gwet (2001). Dieser konstante Wert gibt die maximal mögliche Wahrscheinlichkeit für zufällige Übereinstimmungen an. Es sind Fälle denkbar, in denen sich dieser Wert verändert, wenn die Variable nicht vollständig zufallsverteilt ist. Vgl. dazu Gwet (2001) und s. unten.

Ergebnis 8: $\alpha_{gew1} = \frac{\alpha_{g1} - p_k}{1 - p_k} = \frac{0{,}8\bar{6} - 0{,}2}{1 - 0{,}2} = \frac{0{,}6\bar{}}{0{,}8} = 0{,}8\bar{3}$

Ergebnis 9: $\alpha_{gew2} = \frac{\alpha_{g2} - p_k}{1 - p_k} = \frac{0{,}8 - 0{,}2}{1 - 0{,}2} = \frac{0{,}6}{0{,}8} = 0{,}75$

Streng genommen kann diese Gewichtung allerdings nur vorgenommen werden, wenn die einzelnen Ausprägungen tatsächlich die gleiche Wahrscheinlichkeit haben, codiert zu werden. Diese Voraussetzung kann insbesondere bei Codierungen, die eine Selektions- *und* eine Klassifikationsleistung beinhalten, häufig verletzt werden.

Krippendorff (1980: 142, 4. Formel) schlägt zur Gewichtung ebenfalls die Bestimmung der Abweichungswahrscheinlichkeit vor. Er wendet ein empirisch orientiertes Verfahren an, das hier adaptiert werden kann. Krippendorff stellt dazu die empirisch beobachteten Übereinstimmungen und Abweichungen in einer symmetrischen Matrix dar, die in einer Kreuztabelle jeweils die Ausprägungen der zu untersuchenden Variable als Zellen und Spalten aufweist[35] (vgl. Krippendorff, 1980: 141 und für unser Beispiel Abbildung 8).

Abbildung 8: Krippendorffsche Matrix für Beispielfall 4 (Vier Codierer, eine Variable mit fünf Ausprägungen und 50 Fällen: in 10 Fällen drei Abweichungen, in 40 Fällen keine Abweichung)

Ausprägung	1	2	3	4	5	Randsumme
1	96	0	3	0	15	114
2	0	96	3	12	3	114
3	3	3	96	9	12	123
4	0	12	9	96	3	120
5	15	3	12	3	96	129
Randsumme	114	114	123	120	129	600

In der Diagonale von links oben nach rechts unten finden sich dann die (faktorisierten Anzahlen der) Übereinstimmungen, während die restlichen Zellen die Abweichungen zeigen. Über die Randsummen (R_1, R_2, R_3, R_4, R_5 und R_g) können nun unter Einbezug der empirischen Codierungshäufigkeit der einzelnen abweichenden Ausprägungen deren Wahrscheinlichkeiten berechnet werden.[36]

35 Er untersucht allerdings alle Übereinstimmungen und Abweichungen unter Einbezug der Reihenfolge der Nennung, so dass in der Tabelle in den Randsummen (um den Faktor m Codierer –1) höhere Werte als die tatsächlichen Übereinstimmungen auftreten.

36 Diese Berechnung ähnelt der Bestimmung von Chi-Quadrat-Tabellen mit erwarteten Werten. Ähnlich wie bei diesen Tabellen ergeben sich Probleme für die Gewichtung des Reliabilitätskoeffizienten, wenn einige Zellen nicht besetzt sind. Das Verfahren berechnet z. B. für eine Vierfeldertabelle, bei der alle Codierpaare (zu codieren 0 und 1) in einer Zelle (also z. B. bei 1 und 1) zu finden sind, die zufällige Übereinstimmungswahrscheinlichkeit 1. Das hängt damit zusammen, dass die empirisch bestimmten Randsummen als „Grundgesamtheiten" behandelt werden und das Gewichtungsverfahren somit als Grundannahme hat, dass nur die Ausprä-

Für eine Gesamtabweichungswahrscheinlichkeit (D_e)[37] gibt Krippendorff (1980: 142) folgende Formel 6 an:

Formel 6: $$D_e = \frac{2 \cdot \left(R_1 \cdot R_2 + R_1 \cdot R_3 + R_1 \cdot R_4 + R_1 \cdot R_5 + R_2 \cdot R_3 + R_2 \cdot R_4 + R_2 \cdot R_5 + R_3 \cdot R_4 + R_3 \cdot R_5 + R_4 \cdot R_5\right)}{R_g \cdot \left(R_g - \left(m-1\right)\right)}$$

Bei Einsetzen der entsprechenden Werte ergibt Ergebnis 10 für unser Beispiel 4 eine Gesamtabweichungswahrscheinlichkeit von 0,804.

Ergebnis 10: $$D_e = \frac{2 \cdot (114 \cdot 114 + 114 \cdot 123 + 114 \cdot 120 + 114 \cdot 129 + 114 \cdot 123 + 114 \cdot 120 + 114 \cdot 129 + 123 \cdot 120 + 123 \cdot 129 + 120 \cdot 129)}{600 \cdot \left(600 - (4-1)\right)}$$

$$= \frac{2 \cdot 143919}{358200}$$

$$= \frac{287838}{358200}$$

$$= 0{,}804$$

Dieser Wert unterscheidet sich von der geschätzten Gesamtabweichungswahrscheinlichkeit nur in der dritten Stelle hinter dem Komma, so dass sich das korrigiert gewichtete Alpha nicht stark vom einfach gewichteten Alpha unterscheidet (0,75 vs. 0,751, vgl. Ergebnisse 9 und 11).

Ergebnis 11: $$\alpha_{korr2} = \frac{\alpha_{g1} - p_k}{D_e} = \frac{0{,}8 - 0{,}2}{0{,}804} = \frac{0{,}6}{0{,}804} = 0{,}751$$

Wenn sich die Verteilungen auf die unterschiedlichen Ausprägungen der zu untersuchenden Variable stärker voneinander unterscheiden, kann eine komplexe Gewichtung Sinn machen. Die einfach geschätzte Gewichtung unseres Alpha-Koeffizienten kann jedoch in Fällen mit großer Streuung zur empirischen Bestimmung verwendet werden. Es bleibt festzuhalten, dass die hier vorgestellte Alpha-Berechnung für diesen Beispielfall der von Krippendorff (1980) gleicht, da die empirischen Abweichungskoeffizienten gleich sind (jeweils 0,2). Sie unterscheiden sich jedoch für den Beispielfall 3 deutlich (Krippendorff: 0,2, Kolb 0,133). Mit Krippendorffs Alpha-Berechnung kommt man folgerichtig zum (fast[38]) gleichen Wert wie im Beispielfall 4 (0,751). Die hier vorgestellte Gewichtung ermittelt jedoch einen korrigiert gewichteten Alpha-Wert von 0,834[39].

6. Fazit

Generell muss bei anspruchsvollen Reliabilitätsberechnungen leider ein hoher Arbeitsaufwand in Kauf genommen werden. Dieser Aufwand und die Probleme der etablierten Koeffizienten sollten jedoch nicht zur desillusionierten Aufgabe von Reliabilitätsüber-

---

gung 1 vorkommen konnte, somit also keine Wahl bestanden hat. Dieses Problem versucht Gwet (2001) mit anderen Gewichtungsverfahren zu umgehen.

37 D_e steht für „expected disagreement", also erwartete Abweichungen (Krippendorff, 1980: 142).

38 Vgl. Fußnote 29.

39 Vgl. auch hier den in Ergebnis 8 ermittelten, einfach gewichteten Wert von 0,833, der sich nur marginal von dem mit der Krippendorffschen Gewichtung unterscheidet.

prüfungen führen. Die hier vorgestellten alternativen Alpha-Berechnungen stellen insbesondere für die Berechnung von Reliabilitätswerten für mehr als zwei Codierer auf Nominalskalenniveau eine Weiterentwicklung dar, die besser vergleichbare Reliabilitätswerte liefert, da diese auf äquivalenter Grundlage berechnet werden. Dies stellt streng genommen erst die Basis für eine Interpretation und wissenschaftliche Diskussion solcher Werte her: Nur wenn die Reliabilitätskoeffizienten für unterschiedliche Codiererzahlen und Variablenausprägungen wirklich vergleichbar sind, kann die Qualität von Inhaltsanalysen fundiert diskutiert werden.[40]

Der eigentliche Reliabilitätstest beginnt m. E. – unabhängig davon, welcher Koeffizient verwendet wird – erst nach der Berechnung. Die ermittelten Werte können und sollen publiziert werden, insbesondere da die hier vorgestellte alternative Berechnungsart die Interpretierbarkeit der Koeffizienten deutlich verbessert. Diese Koeffizienten können im Endeffekt aber nur als Hinweise auf die Notwendigkeit einer genaueren Datenanalyse im wörtlichen Sinne verwendet werden: ein Durchsehen der einzelnen Variablen und Fälle nach Auffälligkeiten und Fehlerhäufungen. Diese genaue Datenanalyse kann dem Forschungsteam genaueren Aufschluss über die Probleme des Codebuches geben, die es in Nachbesserungen weitestgehend zu eliminieren gilt.

7. Literatur und Quellen

Bennett, E. M.; Alpert, R. & Goldstein, A. C. (1954): Communications through limited response questionning. In: Public Opinion Quarterly: 303–308, zitiert nach Gwet (2001).

Carletta, Jean (1996): Assessing Agreement on Classification Tasks: the Kappa Statistics. In: Computational Linguistics, 2/1996, 249–254.

Cohen, Jacob (1960): A Coefficient of Agreement for Nominal Scales. In: Educational and Psychological Measurement, 1/1960, 37–46.

Fleiss, Joseph L. (1971): Measuring Nominal Scale Agreement among many Raters. In: Psychological Bulletin, 5/1971 (Vol. 76), 378–382.

Früh, Werner (1991): Inhaltsanalyse: Theorie und Praxis. München: Ölschläger.

Gwet, Kilem (2001): Handbook of Inter-Rater Reliability. How to Estimate the Level of Agreement Between Two or Multiple Raters. Gaithersburg, MD: Stataxis Publishing Company.

Gwet, Kilem (2002a): Computing Inter-Rater Reliability with the SAS System. In: Statistical Methods for Inter-Rater Reliability Assessment, No. 3, Download von http://www.stataxis.com/ vom 26.5.2003.

Gwet, Kilem (2002b): Inter-Rater Reliability: Dependancy on Trait Prevalence and Marginal Homogenity. In: Statistical Methods for Inter-Rater Reliability Assessment, No. 2, Download von http://www.stataxis.com/ vom 26.5.2003.

Gwet, Kilem (2002c): Kappa Statistic is not Satisfactory for Assessing the Extent of Agreement Between Raters. In: Statistical Methods for Inter-Rater Reliability Assessment, No. 1, Download von http://www.stataxis.com/ vom 26.5.2003.

Holley, J. W. & Guilford, J. P. (1964): A note on the G index of agreement. In: Educational and Psychological Measurement: 749–753, zitiert nach: Gwet (2001).

Holsti, Ole R. (1969): Content Analysis for the Social Sciences and Humanities. Reading, MA: Addison-Wesley cop.

Kolb, Steffen (2004): Berechnung von Multicoder-Reliabilitäten mit SPSS, http://www.hans-bredow-institut.de/publikationen/muk/muk043exkurskolb.pdf

40 Durch ihre bessere Vergleichbarkeit könnte man nach mehreren Anwendungen eine erneute Grenzwertdiskussion beginnen, die hier nicht geführt werden kann, da noch keine Daten für den neuen Koeffizienten vorliegen.

Krippendorff, Klaus (1980): Content Analysis: An Introduction to its Methodology. Beverly Hills, CA: Sage.
Lauf, Edmund (2001): „.96 nach Holsti". Zur Reliabilität von Inhaltsanalysen und deren Darstellung in kommunikationswissenschaftlichen Fachzeitschriften. In: Publizistik, 1/2001, 57–68.
Maxwell, A. E. (1977): Coefficients of agreement between observers and their interpretation. In: British Journal of Psychiatry: 79–83, zitiert nach Gwet (2001).
Mayring, Phillipp (2003): Qualitative Inhaltsanalyse. Grundlagen und Techniken. (8. Auflage) Weinheim, Basel: Beltz Verlag.
Merten, Klaus (1995): Inhaltsanalyse: Einführung in Theorie, Methode und Praxis. Opladen: Westdeutscher Verlag.
Meyen, Michael (2002): Die Quelle Mensch: Biographische Interviews als Weg zu einer Geschichte der Mediennutzung in der DDR. Vortragsabstract zur Jahrestagung der DGPuK-Fachgruppe Methoden der empirischen Kommunikationsforschung in Mainz. In: http://www.dgpuk.de/fg_meth/, Download vom 09.10.2003.
Neuendorf, Kimberly A. (2002): The Content Analysis Guidebook. Thousand Oaks, CA: Sage.
Neuendorf, Kimberly A. (2003): The Content Analysis Guidebook Online. In: http://academic.csuohio.edu/kneuendorf/content/, Download vom 26.05.2003.
Scott, William A. (1955): Reliability of Content Analysis: The Case of Nominal Scale Coding. In: Public Opinion Quarterly, Fall/1955, 321–325.
Seiffert, Helmut (1971): Einführung in die Wissenschaftstheorie. Band 1 und 2. München: CH Beck.
Traub, Ross E. (1994): Reliability for the Social Sciences: Theory and Applications. Thousand Oaks, CA: Sage.
Wirth, Werner (2001): Der Codierprozess als gelenkte Rezeption. Bausteine für eine Theorie des Codierens. In: Lauf, Edmund & Wirth, Werner (Hg.): Inhaltsanalysen. Perspektiven, Probleme, Potentiale. Köln: Halem Verlag: 157–182.