SSOAR

Open Access Repository
www.ssoar.info

Der Einsatz von Maßzahlen der Interkoder-Reliabilität in der Inhaltsanalyse

Müller-Benedict, Volker

Preprint / Preprint
Arbeitspapier / working paper

Empfohlene Zitierung / Suggested Citation:
Müller-Benedict, V. (1997). *Der Einsatz von Maßzahlen der Interkoder-Reliabilität in der Inhaltsanalyse.* Flensburg. https://nbn-resolving.org/urn:nbn:de:0168-ssoar-12596

Nutzungsbedingungen:
Dieser Text wird unter einer CC BY-NC-ND Lizenz (Namensnennung-Nicht-kommerziell-Keine Bearbeitung) zur Verfügung gestellt. Nähere Auskünfte zu den CC-Lizenzen finden Sie hier:
https://creativecommons.org/licenses/by-nc-nd/4.0/deed.de

Terms of use:
This document is made available under a CC BY-NC-ND Licence (Attribution-Non Comercial-NoDerivatives). For more Information see:
https://creativecommons.org/licenses/by-nc-nd/4.0

gesis
Leibniz-Institut
für Sozialwissenschaften

Mitglied der
Leibniz-Gemeinschaft

Volker Müller-Benedict

Der Einsatz von Maßzahlen der Intercoder-Reliabilität in der Inhaltsanalyse

Zusammenfassung:

Die Prüfung der Intercoder-Reliabilität bei Codierungen im Rahmen qualitativer Inhaltsanalysen steht vor der Schwierigkeit, daß zwar theoretisch verschiedene Methoden entwickelt, diese aber praktisch schlecht einsetzbar sind, da sie entweder nicht als Computerprogramm verfügbar oder auf die Probleme konkreter Codierpraxis nicht anwendbar sind. In diesem Artikel wird eine Maßzahl für die Intercoder-Reliabilität vorgeschlagen, die auf die meisten Codierungsfälle anwendbar ist. Sie ist eine Verallgemeinerung von Scotts kappa (1955). Die entsprechenden Berechnungen werden hier vorgestellt und können mit einem vom Autor entwickelten Programm durchgeführt werden. An Beispielen aus der Forschungspraxis wird dargestellt, daß der Ertrag der Anwendung dieser Methode nicht nur zur nachträglichen Messung, sondern auch heuristisch bei der Weiterentwicklung eines Kategorienschemas eingesetzt werden kann.

Summary:

The measurement of intercoder agreement in content-analytic research projects often bears difficulties: there is a variety of methods, but many of them are not implemented in computer programs or cannot be used in cases, which are typical for coding text contents. In this article a coefficient is suggested, which can be used in most common cases of coding verbal materials. It is a generalization of Scotts index (1955). The needed calculations are shown; they can be made with a computer program written by the author. Two examples of use in actual research demonstrate the yield of this method: additionally to ex-post measurement it can help to develop a system of categories.

Inhalt:

Der Einsatz von Meßzahlen der Intercoder-Reliabilität in der Inhaltsanalyse

0. Bedeutung und Notwendigkeit der Prüfung der Intercoder-Reliabilität

Ein häufiges Problem bei der Analyse von Texten aus der qualitativen Sozialforschung ist die Bildung eines Kategorienschemas, das wesentliche Eigenschaften der Texte erfaßt, und die Einordnung der Texte an Hand dieses Kategorienschemas. Dieser Vorgang der Vercodung ist die altbekannte Schwachstelle aller kategorialen Analysen: er ermöglicht einerseits die meist damit gewünschte quantitative Durchdringung des sprachlichen Materials, ist aber andrerseits anfällig gegen Kritik; er ist nämlich den subjektiven Deutungen des Kategorienschemas und der Texte ausgeliefert, die von den Codierern beim Codierungsvorgang unweigerlich ausgeführt werden müssen[1].

Um diese Fehlerquelle möglichst gering zu halten und damit Reliabilität herzustellen, wird i.A. der Codierungsvorgang von zwei oder noch mehr Codierern[2] ausgeführt; strittige Fälle werden am geeignetsten im Gesamtteam besprochen, bis eine einstimmige Einigung erzielt wird. Das ist aber nur dann möglich, wenn es so wenig zu vercodendes Material gibt, daß alle Texte von allen Codieren codiert werden können. Nun treten bei zunehmender Verbreitung der Inhaltsanalyse öfter Fälle auf, in denen viel Text codiert und deshalb das Material aus forschungspraktischen Gründen auf die Codierer aufgeteilt werden muß. Dann muß letztlich das Urteil eines einzigen Codieres für die „richtige“ Einordnung der Texte bürgen. So kam es auch zur Beschäftigung mit diesem Thema: die hier vorgestellte Methode wurde für zwei Projekte entwickelt, die mit diesem Problem konfrontiert sind.

Es handelt sich einerseits um ein Projekt[3], in dem einige tausend historische Aufsätze aus meist behördlichen und verbandsinternen Zeitschriften zur Frage der schulischen und universitären Prüfungshandhabung auf bestimmte Charakteristika hin zu kategorisieren sind; u.a. geht es z.B. um die Einordnung, ob die Tendenz eines Aufsatzes eher in Prüfungsverschärfungen oder -erleichterungen zu sehen ist und in welchem Grad, ob aus dem Aufsatz eine eher konservative Sicht der sozialen Schichtung und Mobilität spricht oder eine eher progressive und in welchem Grad etc. Im zweiten Projekt[4] geht es um die Codierung von sog. Satzergänzungstests, die an mehreren hundert Grundschulkindern aus verschiedenen Klassen durchgeführt werden. Hier müssen u.a. die spontanen Ergänzungen, die den Kindern zu Satzanfängen wie „Ich bin traurig...“ oder „Wenn ich etwas nicht schaffe,...“ einfallen, kategorisiert werden in einem Schema, das u.a. den Grad der Autonomie, Selbstständigkeit und Selbstsicherheit der Kinder messen soll.

---

[1] Das trifft genauso zu für computergestützte Codierung, da die Zuordnungen, die der Computer vornimmt, ebenfalls von einem oder mehreren Codierern irgendwann einmal subjektiv festgelegt bzw. bestätigt worden sein müssen.

[2] Aus Gründen der Lesbarkeit möchte ich mich auf die männliche Form dieses Wortes beschränken.

[3] DFG -Projekt von Herrn Dr. A. N., Pädagogisches Institut, Universität L.

[4] Forschungsprojekt von Frau Prof. Dr. C. K., Pädagogisches Seminar, Universität G.

In diesen Fällen ist es wichtig, an einem Teil des Materials Pretests mit allen Codierern durchzuführen und zu bestimmen, ob die Codierer hinreichend genau das Kategorienschema in gleicher Weise anwenden. Dazu kann die sog. Intercoder-Reliabilität bestimmt werden, eine Maßzahl zwischen 0 und 1, die den Zusammenhang zwischen mehreren Vercodungen des gleichen Textteils mißt[5] und üblicherweise mit κ (kappa) bezeichnet wird.

Nun gibt es noch keine allgemein anwendbare Maßzahl für dieses Problem der Intercoder-Reliabilität. In SPSS etwa findet man „Cohens κ“, das jedoch nur unter restriktiven Bedingungen die Übereinstimmung von zwei Codieren mißt[6]. In verschiedenen Büchern über Inhaltsanalyse finden sich unterschiedliche Kritiken und Verbesserungsvorschläge der Maßzahl κ (Lisch/Kriz 1978:88-100; Merten 1983:302-307, Bakeman/Gottman 1986:74-99, Übersax 1982[7]). Dabei bleiben mehrere Probleme offen:

- mit welcher Maßzahl soll die Übereinstimmung zwischen mehr als zwei Codierern gemessen werden?
- wie ist zu verfahren, wenn die Codierer nicht dieselben Kategorien angewendet haben?
- wie kann Übereinstimmung zwischen Codierern gemessen werden, wenn diese mehrere verschiedene Kategorienschemata auf Texte angewendet haben?

Dieser Beitrag möchte die Vielfalt der hierfür möglichen Koeffizienten transparent machen und aus den verschiedenen Möglichkeiten einen bestimmten Koeffizienten vorschlagen, der den geschilderten Problemen am besten gerecht wird.

Über die oben geschilderte Problemlage hinaus bietet ein solcher Koeffizient weitere Möglichkeiten der Steigerung der Reliabilität. Er erlaubt z.B., aus einer größeren Anfangsmenge von Codierern, etwa aus einem Praxis-Seminar, eine kleinere qualifiziertere Menge für das eigentliche Forschungsprojekt auszuwählen. Weiter kann er Verbesserungen bei der Entwicklung eines Kategorienschemas meßbar machen: wenn die gleichen Codierer dieselben[8] oder vergleichbare Texte, die sie in der Pre-Test-Phase mit einem vorläufigen Kategorienschema vercodet haben, später noch einmal mit dem im Forschungsprozeß erarbeiteten endgültigen Kategorienschema codieren, sollte eine substantielle Erhöhung des Koeffizienten der Intercoder-Reliabilität nachweisbar sein.

Nach der Darstellung der verschiedenen Möglichkeiten und der Begründung für die Aus-

---

[5]Diese Maßzahl hat ähnliche Bedeutung wie andere gebräuchliche Maßzahlen von Zusammenhängen zwischen 0 und 1 (z.B. Kronbachs α, die Korrelation): 0 bedeutet, daß keinerlei Zusammenhang bzw. Übereinstimmung besteht, 1, daß der strengste Zusammenhang bzw. hier vollkommene Übereinstimmung besteht.

[6]Mit einem „Trick“, dem Stürzen der Datenmatrix (mit dem Befehl „FLIP“), kann man trotzdem die Übereinstimmung zwischen mehreren Beurteilern messen, indem man dann Kronbachs α anwendet; diese Maßzahl gilt jedoch nur für intervallskalierte Daten, also Beurteilungen mit Rationalzahlen, etwa beim Eislauf, und nicht für Kategorien (MacLennan 1993).

[7]Bakeman/Gottman wird in Schnell/Hill/Esser 1988 als relevante Quelle dieses Problems aufgeführt, auf Übersax wird wiederum in Bakeman/Gottman zur Präzisierung verwiesen. Ich werde mich im folgenden auf die genannten Texte beziehen, soweit Formeln zitiert werden, da sie bekannte Lehrbücher sind und daher die Entwicklung der Diskussion widergeben. S.a. Bos/Tarnai, wo in 3 von 20 Artikeln die Intercoder-Reliabilität erwähnt wird, wovon einer nur einen Wert angibt (203), und einer mit Cohens κ (183), einer mit Phi (dichotome Kategorisierung, 64ff.) arbeitet.

[8]mit entsprechendem zeitlichen Abstand, um „Lerneffekte“ weitgehend auszuschließen

wahl eines Koeffizienten (I.) wird die Berechnungsmethode erläutert (II.) Anschließend werden ein Beispiel (III.) und Berechnungen aus den Projekten (IV.) vorgestellt. Die Ergebnisse werden von einem für dieses Problem geschriebenen Programm erhalten, das der Autor auf diesem Wege Interessierten zur Verfügung stellen möchte.

I. Die verschiedenen Methoden, Übereinstimmung zwischen Codierern zu bestimmen

a) Die Möglichkeiten bei zwei Codierern

Das generelle Vorgehen bei der Bestimmung einer Übereinstimmungs-Maßzahl für zwei Codierer A und B ist das folgende:

1. Es wird die Anzahl der Übereinstimmungen zwischen beiden Codierern bestimmt. Jeder Text, dem beide Codierer dieselbe Kategorie zugeordnet haben, zählt als eine Übereinstimmung. Vorausetzung für dieses Verfahren ist, daß beide Codierer die gleiche Anzahl von Texten codiert haben. Da jedoch die Übereinstimmung der beiden interessiert, ist jeder Text, der nur von einem codiert wurde, für diese Fragestellung nicht relevant, und diese Voraussetzung gilt deshalb o.B.d.A.
2. Es wird der Anteil der Übereinstimmungen P_0 an der Gesamtsumme der Codierungen bestimmt[9], d.h. der Quotient aus der Summe der Übereinstimmungen und der Anzahl der codierten Texte.
3. Dieser Anteil wird bereinigt um den Anteil an Übereinstimmung P_c, der durch reine Zufallscodierung entstanden wäre[10], so daß eine Codierung nach Zufall das Maß 0 bekommt[11]. Dieser bereinigte Wert heißt dann κ und wird wie folgt berechnet:

$$(1)\ \kappa = (P_0 - P_c) / (1 - P_c)$$

Die Berechnungsmethoden von P_0 und P_c unterscheiden sich nun nach mehreren Gesichtspunkten:

(1) Bei der Berechnung der Übereinstimmung P_0 können folgende Fälle auftreten:

(a) Es gibt die Möglichkeit, daß pro zu vercodenden Text nur eine Kategorie gewählt werden kann (Bakeman, Merten) oder daß auch mehrere gewählt werden dürfen (Übersax, Lisch/Kriz:93ff.).

(b) Es gibt die Möglichkeit, daß die gewählten Kategorien bei beiden Codierern genau übereinstimmen müssen oder daß ein Codierer auch eine Kategorie gar nicht verwendet bzw. eine verwendet, die der andere nicht verwendet.

(2) Die zufällige Übereinstimmungswahrscheinlichkeit P_c kann auf mehrere Weisen bestimmt werden:

---

[9] = Z bei Kriz, C bei Merten, P_0 bei Bakeman und Übersax.

[10] = E bei Kriz, P_e bei Merten, P_c bei Bakeman und Übersax.

[11] Die Maßzahl kann also in Extremfällen im Widerspruch zum oben Gesagten auch negativ werden, wenn die Vercodung „schlechter“ als eine reine Zufallsvercodung ist, z.B. wenn bei zwei Texten und zwei Kategorien keinerlei Übereinstimmung besteht.

(a) es kann angenommen werden, daß jede Kategorie die gleiche W. hat, gewählt zu werden (Lisch/Kriz:91)

(b) es kann angenommen werden, daß die Kategorien unterschiedliche W. haben, gewählt zu werden, daß aber alle Codierer die gleiche W. haben, diese Kategorien auszuwählen (Scotts (1955) Vorschlag nach Lisch/Kriz:92, Merten:305-306)

(c) es kann angenommen werden, daß die Codierer unterschiedliche W. haben, bestimmte Kategorien zu benutzen (Bakeman/Gottman, Übersax)

Zur Auswahl eines einzigen Verfahrens aus dieser Vielfalt:

Zu (1a): In der Standardanwendung in den Sozialwissenschaften - Textteile mit einer Codenummer zu belegen - können wir annnehmen, daß wir es nur damit zu tun haben, immer genau eine Kategorie auszuwählen.

Zunächst gibt es dafür theoretische Gründe: Die Validität eines Kategorienschemas ist erstens um so höher[12], je trennschärfer die einzelnen Kategorien, je eindeutiger sie voneinander abgrenzbar sind. Wenn tatsächlich auf einen Text mehrere Kategorien zutreffen, dieser Text also mehrere Qualitäten bzw. Dimensionen aufweist, wird es sich für eine Validitätserhöhung lohnen, entweder das Kategorienschema daraufhin zu überprüfen, es auf Mischtypen zu erweitern (wenn die Simultanität des Auftretens kennzeichnend ist), oder die Texte daraufhin zu überprüfen, ob sie nicht in Teiltexte untergliedert werden müssen, die eindeutig kategorisierbar sind (wenn die Kategorien nicht simultan im Text auftreten). Zweitens ist die Validität um so höher, je umfassender das Kategorienschema ist, d.h. je wahrscheinlicher es auch auf beliebige andere Texte, die nicht im untersuchten Sampling waren, anwendbar ist und dort trennscharf bleibt. Treten gehäuft Texte auf, die nur mit mehreren Kategorien erfaßt werden können, sollte deshalb das Kategorienschema (und natürlich die dahinterstehenden theoretischen Hypothesen) generell auf seine Dimensionalität überprüft werden, weil offenbar die Kombinierbarkeit aus bisherigen Kategorien eine weitere wesentliche Qualität darstellt.

Der praktische Grund für die Vergabe von nur einer Kategorie pro Text liegt darin, daß i.A. mit Rohdatenmatrizen gearbeitet wird, in denen die Texte als „cases“ in den Zeilen stehen und die Codierer als „variables“ in den Spalten, so daß pro „Fall“ (Text) jedem Codierer nur 1 Typ zusteht. Bei mehreren Katgorien müßte so der Text zweimal erfaßt werden, wobei sich dann die Frage stellt, ob und welche Kategorie bei einem Codierer einzutragen ist, der nur eine Kategorie benutzt hat[13].

---

[12]S. die von Holsti (1969) und u.a. von Merten (1983) begründeten Forderungen an inhaltsanalytische Kategorienschemata

[13]Mit den kombinatorischen Formeln von Übersax scheint das Problem der Zuweiseung mehrerer Kategorien lösbar zu sein. Es finden sich aber dort gravierende Fehler, die zu contraintuitiven Resultaten führen und prinzipiell nicht behebbar sind. Weil dann pro Text mehrere Kombinationen von Kategorien vorkommen, summiert Übersax folgerichtig immer über alle Kombinationen von paarweise gewählten Kategorien, nicht von Codiererpaaren. Im Falle, daß z.B. Codierer A für einen Text die Kategorien 4 und 5 , Codierer B nur die Kategorie 4 gewählt hat, zählen für seine Berechnungen die paarweisen Codierungen (4,4), (5,4), und (4,5). Die Vercodung von B zählt also doppelt (das ist ein vielleicht zu lösendes Problem), und - was m.E. ein echter Fehler ist - die zwei Codierungen desselben Codieres zählen als ein mögliches Kategorienpaar! So ergibt sich nach den Formeln von Übersax für diesen Fall ein P_0 von 1/3 , während man bei Nichtbeachtung dieser „Codierung mit sich selbst“ 1/2 erhält, was intuitiv richtiger ist. Auch bei der Berechnung von P_e er-

Aus diesen Gründen sollte der Koeffizient so gestaltet sein, daß er nur eine Kategorie pro Text und Codierer zuläßt.

Zu (1b): Diese Möglichkeit ist der Grund dafür, warum Cohens κ (auf das sich Bakeman/Gottman und auch Übersax beziehen) nur beschränkt angewendet werden kann. Cohens κ bestimmt die „zufällig mögliche Übereinstimmung“ in einer Kategorie immer mit Hilfe der Produktwahrscheinlichkeiten der Benutzung dieser Kategorie durch die beiden Codierer (s. unten (2c)). Für die Berechnung dieser Maßzahl ist deshalb notwendig, daß jede zur Verfügung stehende Kategorie von jedem Codierer auch mindestens einmal benutzt worden ist. Hat einer von beiden eine Kategorie gar nicht verwendet, wird in diesem Fall auch die Wahrscheinlichkeit, gemeinsame Übereinstimmung zufällig zu erzielen, Null.

Das ist aber nicht sinnvoll, wenn die fragliche Kategorie beiden Codierern prinzipiell zur Verfügung gestanden hat. Setzt man sie Null, so würde man nämlich ein Modell voraussetzen, in dem ein Codierer, der eine vorhandene Kategorie nicht benutzt, sie - aus welchen Gründen auch immer - auch tatsächlich nicht benutzen kann: dann wäre in der Tat die Wanhrscheinlichkeit für gemeinsame Benutzung Null. Das ist aber nicht der Fall, wenn Texte kategorisiert werden sollen: der volle Kategoriensatz steht allen Codieren immer zur Verfügung. Die Nichtcodierung bei Textanalysen durch einen Codierer muß deshalb als bewußter Vorgang gewertet werden, d.h. in einer entsprechenden, gegenüber anderen gemeinsamen Vercodungen kleinen Wahrscheinlichkeit zum Ausdruck kommen. Deshalb muß für Textkategorisierungen zugelassen werden, daß Kategorien auch nicht verwendet werden dürfen[14].

Zu (2a): Diese Annahme ist sicherlich zu stark, denn sie bedeutet, daß die im Kategorienschema repräsentierten Qualitäten der Texte quantitativ in etwa gleich großer Häufigkeit anzutreffen sind, was selten der Fall sein wird und mögliche Resultate über die quantitative Verteilung schon vorwegnimmt.

Zu (2b) und (2c): Bakeman und Übersax als Sozialpsychologen und Psychologen gehen davon aus, daß körperliche oder charakterliche Eigenheiten der Vercoder dazu führen, daß bestimmte Kategorien je Codierer unterschiedlich angewendet werden. Ein Beispiel bei Bakeman/Gottman behandelt das Einordnen der Piepser von Hühnerküken in vier verschiedene Kategorien (Phioo, Soft Peep, Peep and Screech). Ebenso kann man sich vorstellen, daß bei psychologischen Beurteilungen z.B. von Patientenäußerungen der (feststellbare) Charaktertyp des Beurteilers eine systematische Verzerrung der Beurteilung darstellen kann. In diesen Fällen muß man also davon ausgehen, daß die Wahrscheinlichkeit, bestimmte Kategorien auszuwählen, pro Codierer unterschiedlich ist, da individuelle physische oder psychologische Barrieren bestehen, die das Erkennen bestimmter Kategorien hemmen.

---

geben sich bei Übersax ähnliche, weitere Fehler.

[14]Craig gibt eine ähnliche Begründung: „Since, in most content analyses, some „final“ coding decision is made for every content unit, wether by using one coders decision, by negotiation between coders to resolve disagreements, or by majority rule of more than two coders, Scotts procedure is usable in most content analyses“ (1981:261)

Für die Zwecke der Beurteilung von Texten kann man dagegen davon ausgehen, daß keine in der Person der Beurteilenden liegenden Gründe vorhanden sind, die zu systematisch abweichender Vergabe von Kategorien führen. Weder körperliche noch charakterliche Eigenschaften sollten bei einer Textinterpretation eine Rolle spielen. Im Gegenteil: die den Text charakterisierenden Kategorisierungen sollten für jedermann/jede Frau nachvollziehbar und intersubjektiv teilbar sein! Deshalb möchte ich im folgenden o.B.d.A. von Annahme (2b) ausgehen: die Kategorien haben zwar eine unterschiedliche W., gewählt zu werden, aber diese Wahrscheinlichkeit ist gleich für alle Codierer; sie wird nur von der Qualität der Kategorie und von ihrem Auftreten im Textmaterial bestimmt[15].

Mit dieser Auswahl der Annahmen ist vorgezeichnet, daß der Koeffizient nicht in gängiger Software vorhanden ist und selbst berechnet werden muß. Er hat jedoch den Vorteil, in der Praxis ein weiteres Anwendungsfeld zu besitzen als Cohens κ[16].

b) Übereinstimmung unter mehreren Codierern und bei mehreren Kategoriensätzen

Die Aufhebung der Beschränkung auf zwei Codierer und ein Kategorienschema führt zu weiteren Entscheidungen für die Berechnung von κ:

(3) Bei mehreren Codierern kann sowohl ein „Durchschnitt“ der Übereinstimmungen aller Codierer-Paare (Übersax) als auch ein Wert für die gemeinsame Übereinstimmung aller bzw. eines angebbaren Teils der Codierer gewählt werden (Merten, Craig).

(4) Oft ist es in der Praxis so, daß auf die vorhandenen Textmaterialien nicht nur ein Kategorienschema, sondern mehrere angewendet werden, die jeweils unterschiedlich viele Typen bzw. Kategorien beinhalten, deren Vercodung jedoch von denselben Codieren vorgenommen wird. In den obigen Beispielen etwa wurden für die Anfänge der verschiedenen Sätze je nach den zu erwartenden kindlichen Fortsetzungen unterschiedliche große Kategorienschemata entwickelt; ebenso werden die verschiedenen inhaltlichen Dimensionen der Aufsätze, wie Prüfungsverschärfungsvorschläge, politische Einordnung der Verfasser etc., ebenfalls durch unterschiedliche Kategorienschemata erfaßt. Wie soll die Übereinstimmung der Codierer bzgl. mehrerer Kategorienschemata gemessen werden?

Zu (3): Wird die Übereinstimmung mehrerer Codierer, d.h. die Wahl derselben Kategorie durch mehr als zwei Codierer, zur Berechnung von P_0 verwendet, so werden die Werte dieses Maßes in jedem Fall[17] kleiner sein als ein Durchschnitt aus allen paarweisen Codierungen. Deshalb wären die Werte eines solchen Maßes nicht vergleichbar mit den Werten, die für je zwei aus der vorhandenen Codierermenge gemessen werden. Bei mehreren Codierern stellt

---

[15]Hubert behandelt diesen Fall als „Levenes“ Modell und bemerkt dazu: „Although Levenes model is somewhat contrary to the inference schemes used in psychology, an almost identical result has been used in measuring nominal scale agreement in sociology for some time (see Krippendorf 1970; Scott 1957). In fact, even though no variance term has been available to researchers up to now, Levenes notion may be generally more popular in the social sciences than either of the two matching concepts presented earlier“ (295)

[16] Dessen Anwendung z.B. in SPSS oft mit „cannot be comuputed ...“ endet. S.a. Craig 1981

[17]außer im Idealfall der kompletten Übereinstimmung aller Codierer.

sich aber i.A. nicht die Frage[18], wie hoch die absolute Übereinstimmung zwischen allen ist, sondern, ob durch das Hinzutreten weiterer Codierer eine Verschlechterung der gesamten Codierleistung eintritt und ob einzelne Codierer in ihrer Codierleistung zu weit von anderen abweichen, d.h. mit den meisten anderen weniger Übereinstimmung erzielen. Deshalb muß eine Maßzahl gefunden werden, die die Eigenschaft hat, bei Hinzutreten eines weiteren Codierers, der die gleiche paarweise Übereinstimmung mit den bisherigen Codieren erzielt, wie diese schon untereinander hatten, ebenfalls gleichbleibt. Eine solche Maßzahl ist dadurch gegeben, daß ein - nach der Anzahl der jeweils vorhandenen paarweisen Codierungen gewichteter - Durchschnitt aller paarweisen Übereinstimmungen gebildet wird.

Ein weitere Überlegung führt ebenfalls auf den Durchschnitt als Maßzahl der Übereinstimmung zwischen mehreren Codierern: das Maß an Übereinstimmung zwischen drei Codieren A,B,C sollte genauso groß sein wie das Maß für das Codiererpaar A und B, das sich ergäbe, wenn ihrer gemeinsamen Codierung die Codepaare je Text, die von den Paaren A und C bzw. B und C codiert wurden, hinzugefügt würden. Genau das aber leistet ein gewichteter Durchschnitt der paarweisen Übereinstimmung.

Zu (4): Das Maß für die Übereinstimmung, das sich ergeben sollte, wenn ein Codiererpaar zwei verschiedene Kategorienschema angewendet hat, sollte dasselbe sein, wie wenn die Codierung mit dem zweiten Schema von einem ganz anderen Paar vorgenommen worden wäre, da die Codierungen zweier inhaltlich verschiedener Systeme unabhängig voneinander sind. In diesem Fall hat man es also wieder mit zwei unabhängigen Paaren zu tun, so daß dieselben Überlegungen wie zu (3) zutreffen.

II. Die Berechnung der Maßzahl κ

a) Berechnung von κ_{Cohen} und κ_{Scott}

Mit den obigen Annahmen läßt sich mit den folgenden Schritten ein κ - Wert errechnen, der die oben bezeichneten Eigenschaften hat und hier als κ_{Scott} bezeichnet werden soll.

Zunächst ist von je zwei Codierern die Übereinstimmung P_0 bzgl. eines Kategorienschemas auszuzählen. Dazu wird eine Kreuztabelle erstellt, deren Zeilenköpfe aus den von Codierer A gewählten Kategorien und deren Spaltenköpfe aus den von Codierer B gewählten Kategorien besteht. Dabei sollten die Kategorien, soweit möglich, in der gleichen Reihenfolge gewählt werden. In jeder Zelle werden die Texte gezählt, die von beiden Codierern mit den der Zelle entsprechenden Kategorien codiert wurden. Mit diesem Verfahren stehen in den Zellen der Hauptdiagonalen, soweit sie in Zeile und Spalte dieselbe Kategorie repräsentieren[19], genau die Anzahl der Übereinstimmungen, in den restlichen Zellen die der Nichtüber-

[18] Ich gehe hier weiter von dem Fall aus, daß die Codierer einzeln arbeiten müssen, da nur dann eine Überprüfung der Intercoderreliabilität auf diese Weise notwendig ist. Wenn alle alles codieren können, ist selbstverständlich die Einigung über die strittigen Fälle auf dem Wege der inhaltlichen Diskussion der beste Weg zur Reliabilität, weitere Berechnungen von Koeffizienten sind nicht nötig.

[19] Die Unterschiede zu Cohens κ sind deutlich: dort muß diese Kreuztabelle quadratisch sein und alle Hauptdiagonalenzellen müssen dieselbe Kategorie repräsentieren.

einstimmungen. Der Quotient mit der doppelten Anzahl der gemeinsam vercodeten Texte ist dann P_0.

Von allen diesen paarweisen Übereinstimmungen wird weiter ein gewichteter Durchschnitt gebildet, wobei die Gewichte der jeweilige Anteil der je paarweisen Vercodungen an der Gesamtzahl aller Vercodungen aller Paare bildet. Gibt es z.B. 3 Codierer A,B,C, und haben A und B 60 Texte gemeinsam vercodet, A und C 50, B und C 70, so ist P_0 (A,B,C) = 1/3 * P_0(A,B) + 5/18* P_0(A,C) + 7/18 * P_0(B,C).

Dann ist das Maß zu bestimmen, das durch zufällige Kategorisierung entstanden wäre. Hier sind beide Methoden (2b) nach Cohen und (2c) nach Scott möglich. Nach Cohen ist das Maß an Übereinstimmung, das sich für eine Kategorie durch Zufall ergeben würde, gleich der Wahrscheinlichkeit, mit der diese Kategorie von beiden gleichzeitig gewählt wird, also die Produktwahrscheinlichlichkeit aus den Anteilen, mit der die beiden Codierer sie gewählt haben. Habe etwa Codierer A die Kategorie i 30, B sie 40 mal gewählt, und beide haben jeweils 80 Texte vercodet, so ist sie (30/80)*(40/80) = 3/16 = 48/256. P_c ist dann die Summe dieser Werte über alle Kategorien (s. unten Formel 2(b)). Für mehrere Codiererpaare ist die zufällige Übereinstimmung nach Cohen dann wieder ein entsprechender gewichteter Durchschnitt der paarweisen P_c - Werte (Übersax), da die Wahrscheinlichkeiten der Wahl einer Kategorie von den Codierern abhängen und diese unabhängig sind.

Für den in (2c) beschriebenen, an Scott orientierten Koeffizienten ist für jede Kategorie i die Wahrscheinlichkeit, gewählt zu werden, anders zu bestimmen. Für zwei Codierer ist diese gleich der Häufigkeit p_i, mit der die Kategorie von beiden zusammen gewählt worden ist. Habe etwa Codierer A die Kategorie i 30, B sie 40 Mal gewählt, und beide haben jeweils 80 Texte vercodet, so ist p_i = (30+40)/(80+80) = 7/16. Das Maß an Übereinstimmung P_c, das eine zufällige Codierung erbringen würde, ergibt sich wie oben daraus, daß für beide Codierer angenommen wird, daß sie die Kategorien mit dieser Wahrscheinlichkeit wählen, und ist deshalb p_i^2 = 49/256, und für alle Kategorien

(2a) $P_c = \sum_i p_i^{\,2}$ (nach Scott) (zum Vergleich (2b) $P_c = \sum_i p_{i.} \cdot p_{.i}$ nach Cohen)

Die nach Scott zufällig zustandekommenden Übereinstimmungen können nun bei mehreren Codierern nicht als Durchschnitt von Paaren berechnet werden, da die Wahrscheinlichkeiten der Wahl einer Kategorie nach diesem Modell von allen Codieren gemeinsam beeinflußt wird. Man könnte zunächst meinen, daß einfach für jede Kategorie folgender Quotient berechnet wird, der für mehrere Codierer genauso wie für zwei Codierer zustandekommt: im Zähler die Summe aus allen Wahlen dieser Kategorie über alle Codierer, und im Nenner die Summe der Wahlen aller Codierer. Habe z.B. bei drei Codierern A,B,C, von denen A 10 Texte, B ebenfalls 10 und C 8 Texte codiert hat, Codierer A die Kategorie „3“ 5 Mal, Codierer B dieselbe Kategorie 4 Mal und Codierer C 2 Mal verwendet, so ist p_3 = {(5 + 4 + 2) / (10 + 10 + 8)}^{2} = (11/28)^2 = .3939^2. Damit verschenkt man aber Informationen: man weiß nämlich nicht nur, wie oft Codierer A bei 10 Texten die Kategorie „3“ vergibt, sondern auch wie oft er

sie bei 8 Texten (als Paar mit Codierer C) vergibt, ebenso für B. Nehmen wir also im obigen Beispiel weiter an, daß in der Paarung (A,C) (nur 8 gemeinsame Texte) A dreimal Kategorie „3“ vergeben hat und in der Paarung (B,C) (ebenfalls 8 gemeinsame Texte) B dreimal. Dann ist $p_3 = \{(5 + 4 + 3 + 2 + 3 + 2) / (10 + 10 + 8 + 8 + 8 + 8)\}^2 = (19/52)^2 = .3654^2$. Damit hat man für A den „Mittelwert“ von „5 mal bei 10 Texten“ und „3 mal bei 8 Texten vergeben“ benutzt und damit die zusätzliche Information über die gemeinsame Codierung der Codierer A und C bzw. B und C berücksichtigt.

Auf diese Weise berechnet man die Wahrscheinlichkeiten des Auftretens der Kategorien aus den Informationen über alle Paare und dann nach Formel (2) das Maß P_c für zufällige Codierung. Aus P_0 und P_c wird dann nach (1) κ_{Scott} berechnet, das nun ein Maß für die Übereinstimmung zwischen allen Codieren darstellt[20].

b) Berechnung von Varianzen und Tests

Zur Berechnung der Varianzen der beiden Maßzahlen muß eine Annahme über die Verteilung der Randsummen der Übereinstimmungsmatrizen, d.h. der Wahrscheinlichkeiten, mit denen ein Codierer einem Text eine der Kategorien zuordnet, gemacht werden[21]. Es bietet sich an, sie als (multinomial je nach Anzahl k der Kategorien) Bernoulli-verteilt B(n, $p_1,...,p_i,...,p_k$) anzunehmen, mit der entweder durch die jeweilige Randsummenhäufigkeit $p_{i.}$ bzw. $p_{.i}$ (Cohens Methode) gegebenen oder durch die Häufigkeit der Kategorie in bezug auf (beide bzw. alle) Codierer p_i (= $(p_{i.} + p_{.i})/2$; Scotts Methode) gegebenen Wahrscheinlichkeit als Parameter $p_{i,}$ und n der Zahl der Texte. Die absolute Zahl der Übereinstimmung R_0 bei zwei Codierern ist dann wieder B(n,p) - verteilt mit $p = P_c$ nach (2a) oder (2b), und für den Anteil $P_0 = (1/n)R_0$ gilt

$$(3)\ E(P_0) = \sum_i \left(p_{i.} \cdot p_{.i} \right) = P_c$$

$$\text{Var}(P_0) = (1/n) \cdot P_c \cdot (1 - P_c) ,$$

entsprechend bei Scotts Methode unter dem Summenzeichen p_i^2. Die Verteilung der Maßzahl κ selbst ist dann schwieriger zu berechnen, weil sowohl P_0 als auch P_c Zufallsvariable sind (Fleiss et al 1969). Für einen Test darauf, ob überhaupt Übereinstimmung vorhanden ist, gilt die Nullhypothese $\kappa = 0$. Dann ist natürlich $E(\kappa) = 0$, und es gilt

$$(4)\ \text{Var}(\kappa) = (1/N)\cdot(1/(1-P_c)^2)\cdot \left(P_c - \sum_i \left(p_{i.} \cdot p_{.i}\left(p_{i.} + p_{.i}\right) \right) + P_c^2 \right)$$

(Hubert: 292). Mit diesen Angaben lassen sich Konfidenzintervalle für P_0 berechnen und ein Test auf $\kappa \neq 0$ durchführen[22], wenn man die Bedingungen für die übliche Approximation der Multinomial-Verteilung durch die Normalverteilung als gegeben annimmt.

---

[20] Nur mit diesem Vorgehen ist definierbar, was z.B. ein κ von 0 für alle Codierer genau bedeutet. Wird hier z.B. der Median aller κ - Werte genommen (wie bei Lange/Willenberg:184), kommt beim Codieren nach Zufall keineswegs genau 0 heraus.

[21] Für das Folgende s. besonders Hubert 1977.

[22] Möchte man für beliebiges κ ein Konfidenzintervall angeben, ist eine andere Formel für die Varianz zu benutzen (Fleiss et al 1969, Bakeman/Gottman:81)

Da alle Codierer unabhängig sind, sind es auch die paarweisen Übereinstimmungen; damit ist die Varianz von P_0 über alle Paare die Summierung der entsprechend gewichteten Varianzen. Da bei Scotts Methode die p_i gleich sind für alle Codierer, läßt sich auch κ_{Scott} über alle Paare nach (4) berechnen[23].

c) Vergleich von Cohens und Scotts κ, Güte von κ

Wenn zwei Codierer dieselben Kategorien benutzt haben, also die Vorausetzungen für Cohens Methode vorliegen, ergibt Scotts κ immer leicht geringere Werte als Cohens. Der Grund ist der, daß die Produktwahrscheinlichkeiten ungleicher Randverteilungen immer kleiner sind als die der daraus gemittelten Produkte[24], so daß auch P_c nach Cohen dann kleiner ist.

Für den Fall, daß ein Codierer auch Kategorien benutzt hat, die der andere nicht benutzt hat, sinkt Scotts κ weiter gegenüber Cohens κ ab, da diese Fälle bei der Berechnung von Cohens κ ausgeschlossen werden, aber in jedem Fall eine weitere Nichtübereinstimmung anzeigen. Scotts κ ist damit generell konservativer als Cohens (Krippendorf: 145)

Ist durch den obigen Test klargestellt, daß Codierer eine Übereinstimmung erzielt haben, so fragt sich, welcher Grad der Übereinstimmung als hinreichend für eine akzeptable Codierung angesehen werden soll. Im Prinzip mißt κ, in wieviel Prozent des Bereichs von zufälliger Codierung bis zur völlig gleichen Codierung Übereinstimmung besteht. Seien z.B. 100 Texte zu codieren, und sei auf Grund der gegebenen Kategorien eine zufällige Übereinstimmung von 30 zu erwarten, bedeutet ein κ von 0.6 eine Übereinstimmung in 0,6* (100 - 30) + 30 ≈ 72 Fällen. Ob man damit zufrieden ist, kann nur im Einzelflall entschieden werden. Wie bei vielen derartigen Maßzahlen zwischen 0 und 1 wird i.A. ein Wert über 0.7 als gut angesehen, Werte über 0.5 als noch akzeptabel[25].

III. Ein einfaches Beispiel

Der Autor hat ein Programm entwickelt, das die vorstehenden Schritte ausführt und die entsprechenden Ergebnisse ausgibt[26]. Das folgende einfache Beispiel demonstriert die obigen Berechnungen an Hand der Ausgabe dieses Programms[27]. Es gebe drei Codierer C1, C2 und C3. 100 Texte sind codiert, davon hat C3 nur 80 bearbeitet. Es stehen 4 Kategorien zur Einordnung der Texte zur Verfügung, die die Codes 1, 2, 3, und 4 haben.

---

[23]Die Varianz des κ -Werts nach Cohen für alle Paare ist bei Hubert (296ff) unter der besonderen Annahme angegeben, daß die Randverteilungen keine Zufallsvariable, sondern fix vorgegeben sind (sog. „matching model"), d.h. P_c ist keine Zufallsvariable, sondern Konstante. Dann gilt allgemein
$Var(\kappa) = (1 / (N (1 - P_c))^2) Var (P_0)$.

[24]Ein Beispiel: hat A 50 mal eine bestimmte Kategorie gewählt, B 70 mal, bei 200 Texten insgesamt, so zählt nach Scott $((50+70)/(200+200))^2 = (60/200) * (60/200)$, nach Cohen (50/200) * (70/200) als Maß für zufällige Übereinstimmung in dieser Kategorie.

[25]"our own inclination, based on using kappa with a number of different coding schemes, is to regard kappa less than .7, even when significant, with some concern, but this is only an informal rule of thumb. Fleiss, for example, characterizes kappas of .40 to .60 as fair, .60 to .75 as good, and over .75 as excelant" (Bakeman/Gottman:82). Die Reliabilität von Interviews wird nicht besser eingeschätzt (König 1962:175)

[26]Das Programm „Intercod" kann vom Autor unentgeltlich bezogen werden.

[27]Das Beispiel ist angelehnt an Bakeman/Gottman:80ff

Zunächst soll angenommen werden, daß nur die Kategorien 1 und 2 verwendet werden:

```
1.tes Paar von Codierern C1, C2

Code   1   2  Sum
   1   7   2   9
   2   1  90  91
  Sum  8  92 100

Übereinstimmungen:  97
p0 (Cohen)          = 0.97;
            Varianz = 0.001314
Summe Randprodukte  = 8444
pc (Cohen)          = 0.8444
kappa (Cohen)       = 0.8072;
            Varianz = 0.0099
   z-Wert (kappa=0) = 8.808872

p0 (Scott)          = 0.97
            Varianz = 0.001314
pc (Scott)          = 0.84445
kappa (Scott)       = 0.8071362
            Varianz = 0.01
   z-Wert (kappa=0) = 8.07136
---------------------------------
3.tes Paar von Codierern: C2, C3

Code   1   2   Sum
   1   7   0   7
   2   1  72  73
  Sum  8  72  80

Übereinstimmungen:  79
p0 (Cohen)          = 0.9875
            Varianz = 0.00176
Summe Randprodukte  = 5312
pc (Cohen)          = 0.83
kappa (Cohen)       = 0.9264706
            Varianz = 0.0124
   z-Wert (kappa=0) = 8.30911

p0 (Scott)          = 0.9875
            Varianz = 0.00176
pc (Scott)          = 0.8300781
kappa (Scott)       = 0.9264368
            Varianz = 0.0125
   z-Wert (kappa=0) = 8.2863
```

```
2.tes Paar von Codierern C1, C3

Code   1   2  Sum
   1   6   1   7
   2   1  72  73
  Sum  7  73  80

Übereinstimmungen:  78
p0 (Cohen)          = 0.975
            Varianz = 0.00167
Summe Randprodukte  = 5378
pc (Cohen)          = 0.8403125
kappa (Cohen)       = 0.8434444
            Varianz = 0.0125
   z-Wert (kappa=0) = 7.544

p0 (Scott)          = 0.975
            Varianz = 0.00167
pc (Scott)          = 0.8403125
kappa (Scott)       = 0.8434444
            Varianz = 0.0125
   z-Wert (kappa=0) = 7.544
---------------------------------
Alle Codiererpaare:

p0 (Cohen) alle Paare  = 0.97692
                Varianz = 0.00052
pe (Cohen) alle Paare  = 0.83871
kappa(Cohen)alle Paare = 0.85692

Saemtliche Kategorien:    1   2
Globale Haeufigkeiten:   46 474
Summe aller paarw. Beob.:   520

p0 (Scott) alle Paare  = 0.97692
                Varianz = 0.00052
pe (Scott) alle Paare  = 0.83873
kappa(Scott)alle Paare = 0.85691
                Varianz = 0.00385
      z-Wert (kappa=0) = 13.8172
```

Wie oben bemerkt, sind für die Paare (C1, C2) und (C2, C3) die κ - Werte von Scott geringfügig kleiner als die von Cohen, während sie bei der identischen Randverteilung beim Paar (C1, C3) sogar gleich sind. Die κ - Werte für alle Paare liegen unter den Werten des Paares (C2, C3), aber über allen Paaren mit C1, anzeigend, daß C1 „schlechter“ vercodet. Zusätzlich werden nun die Kategorien 3 und 4 verwendet, jedoch nur von C1 bzw. C3.

```
1.tes Paar von Codierern: C1; C2
--> Hinweis: Anzahl der benutzen
Kategorien stimmt nicht überein!

Code   1   2   3  Sum
   1   7   2   0   9
   2   1  80  10  91
  Sum  8  82  10 100
```

```
Übereinstimmungen:  87
p0 (Cohen)          = 0.96667
            Varianz = 0.00158
Summe Randprodukte  = 6714
pc (Cohen)          = 0.82889
kappa (Cohen)       = 0.80519
            Varianz = 0.01106
   z-Wert (kappa=0) = 7.6549

p0 (Scott)          = 0.87
```

```
            Varianz = 0.00183
pc (Scott)          = 0.75795
kappa (Scott)       = 0.46292
            Varianz = 0.00624
   z-Wert (kappa=0) = 5.8625

---------------------------------
3.tes Paar von Codierern: C2, C3
--> Hinweis: Anzahl der benutzen
Kategorien stimmt nicht überein!

Code   1   2  Sum
   1   7   0   7
   2   1  62  63
   4   0  10  10
  Sum  8  72  80

Übereinstimmungen:  69
p0 (Cohen)          = 0.98571
            Varianz = 0.00221
Summe Randprodukte  = 3962
pc (Cohen)          = 0.80857
kappa (Cohen)       = 0.92537
            Varianz = 0.01420
   z-Wert (kappa=0) = 7.7638

p0 (Scott)          = 0.8625
            Varianz = 0.00249
pc (Scott)          = 0.72461
kappa (Scott)       = 0.50071
            Varianz = 0.00761
   z-Wert (kappa=0) = 5.7394
```

```
2.tes Paar von Codierern: C1, C3
--> Hinweis: Nichtübereinstimmende
Kategorien in der Hauptdiagonalen!

Code   1   2   3  Sum
   1   6   1   0   7
   2   1  57   5  63
   4   0   5   5  10
  Sum  7  63  10  80

Übereinstimmungen:  63
p0 (Cohen)          = 0.96923
            Varianz = 0.00239
Summe Randprodukte  = 3413
pc (Cohen)          = 0.80781
kappa (Cohen)       = 0.83991
            Varianz = 0.01538
   z-Wert (kappa=0) = 6.7715

p0 (Scott)          = 0.7875
            Varianz = 0.00290
pc (Scott)          = 0.63564
kappa (Scott)       = 0.41680
            Varianz = 0.00570
   z-Wert (kappa=0) = 5.5190
---------------------------------
Alle Codiererpaare:

p0 (Cohen) alle Paare  = 0.97332
                Varianz = 0.00067
pc (Cohen) alle Paare  = 0.81615
kappa(Cohen)alle Paare = 0.85486

Saemtliche Kategorien: 1  2  3   4
Globale Haeufigkeiten:46 434 20 20
Summe aller paarw. Beob.:      520

p0 (Scott) alle Paare  = 0.84231
                Varianz = 0.00080
pc (Scott) alle Paare  = 0.70737
kappa(Scott)alle Paare = 0.46113
                Varianz = 0.00195
      z-Wert (kappa=0) = 10.4486
```

Die Berücksichtigung der Nichtbenutzung der Kategorien 3 und 4 durch jeweils einen der beiden Codierer eines Paares führt hier zu großen Differenzen zwischen Cohens und Scotts κ. Cohens κ wird an Hand der um diese Kategorien reduzierten quadratischen Teilmatrix berechnet. Dadurch ergibt sich eine viel höhere Übereinstimmung, da die Nichtübereinstimmung dieser Fälle nicht berücksichtigt werden kann.

Mit Hilfe der Varianz-Angabe für P_0 kann dieser Wert ebenfalls auf signifikanten Unterschied zu P_c getestet werden; der Test sollte wie der Test zum dazugehörigen κ ausfallen. Hier ergibt sich z.B. für das Paar (C1,C2) ein z-Wert von $(P_0 - P_c)$ / sqrt(Varianz) = 3.4662 bei Cohens Methode.

IV. Messungen der Veränderung der Codierergruppe und der Verbesserungen eines Kategorienschemas

Im oben skizzierten Forschungsprojekt an Grundschulen wurden in der Anfangsphase einige der Satzergänzungen im Rahmen eines Seminars nach Entwicklung eines vorläufigen Kategorienschemas zunächst von 12 StudentInnen codiert. Für alle Paare wurden die obigen Kreuztabellen erstellt. Ein Beispiel:

```
5.tes Paar von Codierern: k, ve
--> Hinweis: Nicht-übereinstimmende Kategorien in der Hauptdiagonalen!

Code  31  10  21  42  33  32  41  60  22  23  Sum
  31   0   0   0   0   1   3   0   0   0   1   5
  10   0  26   0   0   1   0   0   0   0   0  27
  21   0   0   3   0   0   0   0   0   2   0   5
  42   0   0   0   1   1   0   0   0   0   0   2
  33   0   0   0   0   3   0   0   0   0   0   3
  32   2   0   0   0   1   0   0   0   0   0   3
  41   0   0   0   0   0   0   2   0   0   0   2
  60   0   0   0   0   0   1   0   1   0   0   2
  22   0   0   0   0   0   0   0   0   1   0   1
  50   0   0   0   0   1   0   0   1   0   0   2
 Sum   2  26   3   1   8   4   2   2   3   1  52
Übereinstimmungen:  37
```

Auf Grund der Ergebnisse der Kreuztabellen ließen sich relativ einfach Mängel am Kategorienschema diagnostizieren: z.B. Kategorien, die gar nicht oder nur sehr selten benutzt wurden (hier Code 23), gehäufte Nichtübereinstimmungen, bei denen immer die zwei selben unterschiedlichen Kategorien gewählt wurden (Code hier 31 und 32, Hinweis auf nicht trennscharfe Kategorien), Kategorien, die mit fast allen anderen kombiniert wurden (hier Code 33, Hinweis darauf, daß die Qualität dieser Kategorie fast allen Texten anhaften könnte) etc. Diese Mängel wurden in einer weiteren Fassung des Kategoreinschemas und einer verfeinerten Beschreibung seiner Anwendung behoben. Außerdem wurde deutlich, daß so gut wie nie die Bedingungen für Cohens κ vorlagen: fast immer wurden auch Kategorien nicht gemeinsam verwendet.

Aus den anfangs 12 CodiererInnen wurden dann unter Mitbeachtung der von ihnen erzielten κ - Werte 5 CodiereInnen für das Projekt dauerhaft ausgewählt. Danach wurden dieselben Texte noch einmal von denselben CodiererInnen, jetzt anhand des neuen Schemas, codiert. Zwischen diesen beiden Codierungen lagen ca. 5 Monate, so daß die Erinnerungen an die frühere Codierung (immerhin insg. ca 500 Sätze) wohl eher blaß waren. Die Analyse erbrachte für den ersten zu ergänzenden Satz folgende Ergebnisse[28]:

Paar	s,k	s,ve	s,ch	s,v	k,ve	k,ch	k,v	ve,ch	ve,v	ch,v	alle
1.Cd	.6431	.6605	.7613	.7883	.5928	.6391	.6692	.7872	.6362	.7154	.6913
2.Cd	.7901	.7869	.8115	.8327	.7824	.8550	.7570	.8780	.8025	.8032	.8103

[28]s, k, ve, ch, v sind die Kürzel für die CodiererInnen

Ersichtlich hat für alle Paarungen und für das Gesamt - κ die Änderung des Kategorienschemas zu einer im Durchschnitt 10%-igen Verbesserung der Übereinstimmung geführt (Zeile „1.Codierung“ im Vergleich zu „2.Codierung“), die hier im übrigen bei Berücksichtigung der Anzahl der Kategorien - je mehr Kategorien, desto geringer die Wahrscheinlichkeit der Übereinstimmung - als sehr zufriedenstellend bezeichnet werden kann, da i.A.Werte über 0.7 als akzeptabel angesehen werden.

Das zweite Projekt arbeitet mit nur 3 CodiererInnen und vercodet im Gegensatz zum obigen jeden Text mit mehreren Kategorienschemata. Hier stellt sich u.a. das Problem, ob es sich bei schlechten Einzelergebnissen, d.h. κ - Werten für ein Kategorienschema und ein Codiererpaar, eher um einen vom allgemeinen Standard abweichenden Codierer oder um ein verbesserungswürdiges Kategorienschema handelt. Diese Frage kann dadurch entschieden werden, daß berechnet wird, ob eher durch Herausnahme des Codierers oder durch Herausnahme der Kategorie des Kategorienschemas die jeweiligen Gesamt-κ-Werte sinken. Ein Ergebniss lautete hier z.B. im Einzelnen[29]:

	aw	af	sw	sf	ow	of	pw	pf	alle	o. pw
C1,C2	0.496	0.237	0.481	0.719	0.417	0.478	-0.04	0.697	0.464	0.513
C1,C3	0.191	0.273	0.259	0.420	0.373	0.339	0.08	0.349	0.291	0.321
C2,C3	0.167	0.189	0.281	0.258	0.334	0.319	-0.07	0.299	0.229	0.279
alle	0.302	0.244	0.338	0.463	0.376	0.380	-0-024	0.445		

Es ergab sich, daß bei Herausnahme des Kategrorienschemas „pw“ die κ - Werte für das Gesamtmaß über alle anderen Kategorienschemata für alle Paarungen stiegen (Spalte „ohne pw“ im Vergleich zu „alle“); aber auch die Werte des Codierers 3[30] lagen systematisch unter dem Gesamtdurchschnitt. Hier überlagern sich also zwei Probleme. Im übrigen zeigt die Tabelle, die allerdings den ersten Pretest des Projekts darstellt, daß an der Vercodung noch gearbeitet werden muß, bevor die Hauptmasse der Texte in Angriff genommen werden sollte.

V. Zusammenfassung

Aus den verschiedenen Möglichkeiten, eine Maßzahl der Intercoder-Reliabilität zu berechnen, wurde hier eine Verallgemeinerung der von Scott (1955) vorgeschlagenen Methode als diejenige ausgewählt, die für die Codierung von Textmaterialien im Rahmen von Inhaltsanalysen am geeignetsten erscheint. Die vorgestellte Methode läßt sich nicht nur am ehesten mit elementaren Anforderungen an ein Kategorienschema vereinbaren, sondern bietet zudem den

[29] aw etc. sind die Kürzel für die Kategorienschemata, C1, C2, C3 für die Codierer.

[30] Interessanterweise sind sowohl in diesem Projekt der Codierer 3, als auch im obigen die Codiererin k, die - wie der Tabelle leicht zu entnehmen ist - im ersten Durchgang ebenfalls generell unter dem Durchschnitt liegt, die Projektleiter. Gründe mögen sein, daß sie die Katgeorie „weiß nicht“ zu vermeiden versuchen, oder daß sie zu viel in das Kategorienschema „hineininterpretieren“. Darauf sollte bei der Entwicklung auch geachtet werden.

praktischen Vorteil, auch in den in der Forschungspraxis auftretenden Fällen von ungleicher Anzahl vercodeter Texte, mehr als zwei Codierern und nicht gewählten Kategorien berechenbar zu sein. An Beispielen für die Berechnungsmethode konnte auf die statistischen Eigenschaften eingegangen und die Verwendung der Ergebnisse für die Verbesserung der Codierpraxis und die Entwicklung eines Kategorienschemas demonstriert werden. Die konkrete Anwendung des Programms zur Berechnung der Koeffizienten in zwei Projekten führte zu erheblichen Änderungen an den Kategorienschemata, zur Auswahl von geeigneten CodiererInnen aus einer größeren Gruppe und zur Vereinheitlichung der Codierung. Dies zeigt, daß die Berechung der Intercoder-Reliabilität zu einer fruchtbaren Verschränkung von quantitativen und qualitativen Methoden führen kann, indem sie Veränderungen und Verschiedenheiten bei Codiervorgängen und Kategorienschema-Entwicklung quantitativ erfaßbar und vergleichbar macht und damit Hinweise auf Verbesserungen der qualitativen Analyse geben kann.

Literatur:

Bakeman, R., Gottman, J.M. 1986: Observing interaction. An introduction to sequential analysis. Cambridge (Universtity Press)

Bos, W., Tarnai, C. 1989 (Hg.): Angewandte Inhaltsanalyse in Empirischer Pädagogik und Psychologie. Münster (Waxmann)

Cohen, J. 1960: A coefficient for agreement of nominal scales. In: Educational and Psychological Measurement 20 (1960):37-46

Craig, R. 1981: Generalization of Scott's Index of Intercoder Agreement. In: Public Opinion Quarterly 45 (1981): 260-264

Fleiss, J., Cohen, J., Everitt, B. 1969: Large sample standard errors of kappa and weighted kappa. In: Psychological Bulletin 72(1969):323-327

Holsti, O.R. 1969: Content Analysis fpr the Social Sciences and Humanities. Reading (Addison-Wesley)

Hubert, L. 1977: Kappa revisited. In: Psychological Bulletin 84(1977):289-297.

König, R. (Hg.) 1962: Handbuch der empirischen Sozialforschung. Bd. I. Stuttgart (Enke)

Krippendorf. K.1970: Bivariate agreement coefficients for reliability of data. In: Bortatta, E. (Hg.) 1970: Sociological Methodology 1970. San Francisco

Lange, B., Willenberg, H. 1989: Inhaltsanalyse in der literaturdidaktischen Unterrichtsforschung. In: Bos, W., Tarnai, C. 1989:173-190

Lisch, R. Kriz, J. 1978: Grundlagen und Modelle der Inhaltsanalyse. Bestandsaufnahme und Kritik. Frankfurt/M (rororo)

MacLennan, R.N. 1993: Interrater Reliability With SPSS for Windows 5.0. In: The American Statistician 47(1993):292-296

Merten, K. 1983: Inhaltsanalyse. Einführung in Theorie, Methode und Praxis.Opladen (Westdeutscher Verlag)

Schnell, R., Hill, P., Esser, H. 1988: Methoden der empirischen Sozialforschung. München

Scott, W.A. 1955: Reliability of content analysis: the case of nominal scaling. In: Public Opinion Quarterly 19(1955):321-325

Uebersax, J.S. 1982: A generalized kappa coefficient. In: Educational and Psychological Measurement 42(1982):181-183

Anschrift des Autors: Prof. Dr. Volker Müller-Benedict, Zentrum für Methodenlehre, Auf dem Campus 1, 24943 Flensburg, T.: 0461/ 805-2355, Mail: vbenedi@uni-flensburg.de, www.zml.uni-flensburg.de.