FQS FORUM: QUALITATIVE SOZIALFORSCHUNG SOCIAL RESEARCH

Volume 1, No. 2, Art. 16
Juni 2000

Kleine Kinder – große Datenmengen. Möglichkeiten der Verbindung von qualitativen und quantitativen Methoden zur Analyse von Selbstaussagen

Christina Krause, Volker Müller-Benedict & Ulrich Wiesmann

Keywords: Projektevaluation, Selbstbild, Selbstwertgefühl, Selbstaussagen, Gesundheitsförderung, Grundschulkinder, Längsschnittuntersuchung, qualitative Inhaltsanalyse, Interkoderreliabilität, Kappa-Koeffizienten

Zusammenfassung: Die Evaluierung eines mehrjährigen Gesundheitsförderprogramms für Grundschulkinder führte zu zwei Problemen. Erstens mußten qualitative Instrumente für eine Population (5-10jährige Kinder), für die standardisierte Verfahren schlecht geeignet sind, entwickelt werden. Das Programm wurde in insgesamt 20 Schulklassen erprobt und längsschnittlich wurden verbale und bildgestützte Daten erhoben. So entstand zweitens über einen Zeitraum von fast vier Jahren eine riesige qualitative Datenmenge. Deshalb wurden Verfahren entwickelt, um diese qualitativen Daten quantitativ überschaubar auszuwerten.

Zudem mußte berücksichtigt werden, daß die qualitativen Analyse-Kategorien im Laufe der Zeit selbst eine Weiterentwicklung (Differenzierung) erfuhren. Zum Zwecke der längsschnittlichen Vergleichbarkeit mußten frühere Kodierungen dem jeweils überarbeiteten Kategorienschema angepaßt werden. Insgesamt gesehen konnte eine gleichbleibende Güte der qualitativen Analysen sichergestellt werden. Darüber hinaus ergab die quantitative Auswertung Hinweise auf mögliche Verbesserungen des Kategorienschemas.

Der Beitrag stellt diese Verfahren und die Produktivität ihres Einsatzes im Rahmen der Evaluierung des Förderprogramms dar.

Inhaltsverzeichnis

2. Erfassung selbstbezogener Inhalte von Kindern
- 2.1 Das Projekt "Gesundheit fördern durch Selbstwertstärkung"
- 2.2 Probleme bei der Analyse von Selbstaussagen
- 2.3 Der Satzergänzungstest – Merkmale des qualitativen Verfahrens

3. Die Verbindung von qualitativer und quantitativer Forschung
- 3.1 Die Entwicklung des Kategorienschemas
- 3.2 Übereinstimmungsmaße und Kodierleitfaden-Entwicklung
- 3.3 Überprüfung des Kategorienschemas mit Hilfe von Kreuztabellen von Kodierungen

4. Die Messung der Kodierleistung
- 4.1 Neudefinition von Kappa
- 4.2 Kappa-Berechnungen der Kodierleistung

Literatur

Zur Autorin und zu den Autoren

Zitation

1. Zielstellung

Die begleitende Evaluation eines Forschungsprojekts an Grundschulen, das wir in Abschnitt 2.1 kurz beschreiben, machte es notwendig, eine sinnvolle Verbindung von qualitativen und quantitativen Methoden zu finden. Wie wir in Abschnitt 2.2 und 2.3 darlegen werden, sind für Grundschulkinder qualitative Erhebungsverfahren besser geeignet. Die Ergebnisse der Evaluation sollten jedoch so weit verallgemeinerungsfähig sein, daß sie eine Bewertungsgrundlage für den Entschluß bieten können, dieses Programm breit an Grundschulen einzusetzen. Deshalb mußte die qualitative Erhebung in einer Größenordnung, die sonst nur quantitative Projekte erreichen, durchgeführt werden. Daraus ergab sich der Wunsch, den Prozeß der qualitativen Auswertung durch quantitative Maßzahlen zur Reliabilitäts- und Validitätsüberprüfung abzusichern. In Abschnitt 3 werden wir den Nutzen quantitativer Maßzahlen bei der Entwicklung eines Kategorienschemas demonstrieren. Für die Messung der Reliabilität der Kategorisierung mußte ein neues statistisches Verfahren, das die sog. "Interkoderreliabilität" in den hier vorliegenden komplexeren Fällen messen kann, entwickelt werden. Dieses Verfahren wird in Abschnitt 4 vorgestellt und ist auch über das Internet http://www.uni-goettingen.de/~vbenedi [Broken link, FQS, December 2004] zu beziehen. [1]

2. Erfassung selbstbezogener Inhalte von Kindern

2.1 Das Projekt "Gesundheit fördern durch Selbstwertstärkung"

Das Forschungsprojekt erprobte ein Programm zur Gesundheitsförderung, das vom ersten bis zum vierten Schuljahr eingesetzt wurde. Im Mittelpunkt der Förderung stand die psychische Gesundheit, wobei es in Umsetzung des salutogenetischen Konzeptes (ANTONOVSKY 1993) um die Stärkung von Gesundheitsfaktoren ging und um die Befähigung von Grundschüler/innen, Belastungen kompetent zu bewältigen. ANTONOVSKY (1993) hat auf der Suche nach jenen "gesunderhaltenden Faktoren, die Menschen dazu verhelfen, so erfolgreich wie nur möglich mit den Bedrohungen im Leben umzugehen" (S.10f), sein salutogenetisches Modell und das Konzept des Kohärenzsinnes entwickelt. In diesem Modell werden psychosoziale Ressourcen und das subjektive Bewältigungshandeln als entscheidende Bedingungen angesehen, um sich auf dem Gesundheits-Krankheits-Kontinuum mehr im Bereich des gesunden Pols bewegen zu können. Sind die "allgemeinen Widerstandsressourcen" vorhanden und können sie beim konkreten Bewältigungshandeln in einer Streßsituation eingesetzt werden, dann entsteht das Gefühl der Kohärenz. Die "Ursprünge der Gesundheit" sind sicher am besten im Kindesalter zu entwickeln; zudem sollten Kinder so früh als möglich für die Risiken des Lebens gewappnet sein. Deshalb haben wir mit der Implementierung eines Förderprogrammes im ersten Schuljahr begonnen. Da wir aber sowohl mit Blick auf die individuelle Entwicklungsperspektive als auch mit Blick auf die gesellschaftlich-kulturelle und ökologische Entwicklung nicht wissen, welche Stressoren die heute Sechsjährigen als Dreißigjährige zu bewältigen haben, müssen die zu entwickelnden Ressourcen derart sein, daß sie allgemeine,

situationsübergreifende Grundlagen darstellen. Als eine dieser Grundlagen und wichtige Widerstandsressource verstehen wir ein positives Selbstwertgefühl, dessen Entwicklung, Erhaltung und Förderung eine Schlüsselfunktion in Gesundheitsförderprogrammen zukommt. [2]

Das für die vier Grundschuljahre auf der Grundlage des salutogenetischen Konzeptes entwickelte und erprobte Förderprogramm soll diese Widerstandsressource – ein positives Selbstwertgefühl – erhalten und fördern. Der Schulanfang ist dafür besonders geeignet, da zu diesem Zeitpunkt die meisten Kinder ein positives Selbstwertgefühl haben, das aber im Verlaufe der Grundschulzeit bei nicht wenigen Kindern verloren geht (vgl. KRAUSE 1998). [3]

Die Ersterprobung erfolgte in insgesamt 20 Schulklassen der beiden Städte Göttingen und Greifswald und hat im wesentlichen die Ziele, die zu Beginn im Jahre 1995 formuliert worden waren, erfüllt. In jeder der beteiligten Schulen arbeitete ein sog. "Gesundheitsteam", das die Gesundheitsstunden durchführte. Die dazu erarbeiteten Themen ordneten sich folgenden fünf Schwerpunkten zu:

- Selbstwertstärkende Selbstreflexion,
- Körpererfahrung und Körperbewußtsein,
- gesundheitsförderliche Interaktion und Kommunikation,
- Freizeitverhalten und Gesundheit,
- gesunde Ernährung. [4]

Jeder Gesundheitstag wurde von dem Projektteam erarbeitet und vor der Durchführung mit den Lehrerinnen diskutiert. Außerdem erfolgte regelmäßig die gemeinsame Auswertung nach jedem Gesundheitstag, so daß pro Klasse ein Nachbereitungsprotokoll vorliegt. [5]

Um die Effektivität zu überprüfen und das Programm zu evaluieren, wurden umfangreiche qualitative Befragungen von Grundschulkindern durchgeführt. Die Ausgangshypothese war, daß Schulanfänger sich normalerweise subjektiv wohl fühlen, was sich u.a. in einem positiven Selbstwertgefühl, in Selbstvertrauen und in hoher Einschätzung der eigenen Kompetenz (eher Selbstüberschätzung als -unterschätzung) äußert. [6]

Ein Teil der Untersuchung bestand in der Messung des subjektiven Wohlbefindens junger Kinder. Das Problem bestand darin, ein geeignetes Meßinstrument zu finden, um Mitteilungen von Fünf- bis Zehnjährigen über ihre Befindlichkeit zu erhalten. [7]

2.2 Probleme bei der Analyse von Selbstaussagen

Die Erforschung des Selbst ist mit besonderen methodischen Problemen behaftet. Dies liegt darin begründet, daß Subjekt und Objekt der Betrachtung identisch sind. Vor zwanzig Jahren stellte MUMMENDEY (1979) bereits Methodenprobleme dar, die auch heute noch weitgehend ungelöst sind.

Wesentliche Fragen betreffen a) die Gütekriterien (Reliabilität, Validität, Objektivität), b) die "Passung" von theoretischem Konzept (Selbstkonzept, Selbstbild, Selbstwertgefühl) und Erfassungs-Methode, c) die Indikation einer Erfassungs-Methode, d) die Spezifizität/Generalität der erfaßten Merkmale, e) die Entwicklungsangemessenheit der Erfassungs-Methode (Berücksichtigung der Altersstufen) sowie f) die Beurteilung der Veränderungsmessung bei längsschnittlichen Designs. Insbesondere letztere beiden Problemfelder hatten für unser Projekt, das die Selbstwertstärkung von Grundschülern und -schülerinnen zum Ziel hatte, eine besondere Relevanz. [8]

Unser Untersuchungsplan sah eine längsschnittliche Untersuchung der Kinder von der ersten bis zur vierten Klasse vor. Nach Abschluß der fünf Gesundheitstage, am Ende eines jeweiligen Schuljahres, wurden die Kinder über ihre subjektiven Befindlichkeiten mündlich befragt. Bei der Erstellung des Meßinstrumentes bestand die Schwierigkeit darin, eine Methode zu finden, die

- die Erfassung der Entwicklung und Veränderung des Selbstwertgefühls ermöglicht,
- die für fünf- bis zehnjährige Kinder geeignet ist und
- die über mehrere Jahre wiederholt eingesetzt werden kann. [9]

Eine Durchsicht herkömmlicher Verfahren zur Messung von Selbstkonzept oder Selbstwertgefühl zeigte, daß sie für Jugendliche bzw. Erwachsene entwickelt worden sind. Für (angehende) Schulkinder sind adjektivische Selbstbeschreibungsverfahren, Sortierverfahren (sog. Q-Sorts), semantische Differentiale, Rating-Verfahren (z.B. Einschätzung selbstbezogener Aussagen anhand einer Skala) oder Persönlichkeitsfragebogen offensichtlich wenig sinnvoll. Die Problematik dieser Verfahren – unabhängig von der zu untersuchenden Altersgruppe – ist bereits von MUMMENDEY (1979) überzeugend dargestellt worden (siehe auch HAUSSER 1995). Unser Verzicht auf diese herkömmlichen Methoden war entwicklungspsychologisch begründet: Sie weisen eine sehr geringe ökologische Validität für die Erfassung subjektiver Befindlichkeiten im Grundschulalter auf. Es war also notwendig, eine Methodologie zu verwenden, die für die von uns befragten Altersstufen tauglich ist. Immerhin fand die erste Erhebung im Rahmen der Schuleingangsuntersuchung statt, d.h. die Kinder konnten noch gar nicht lesen. [10]

Die Interpretation von Veränderungen des Selbstwertgefühls kann nur durch Selbstreflexionen der Kinder selbst ermittelt werden. Dies ist bei Grundschulkindern aber besonders problematisch. Es entsteht zum Beispiel die Frage, ob eine beobachtete Veränderung über mehrere Zeitpunkte tatsächlich einen Wandel im Selbstwertgefühl darstellt oder ob die Veränderung in den Selbstaussagen der Kinder eine Entwicklung der kognitiven Fähigkeit dokumentiert. Kinder im Schulalter entwickeln zunehmend Kenntnisse über die eigenen kognitiven Prozesse und deren Steuerung. Diese "Konfundierung" läßt sich auflösen, wenngleich nicht vollständig kontrollieren, wenn die Erfassungs-

Methode der Wahl auch etwas über die zunehmende kognitive Differenzierung aussagen kann. [11]

Wie läßt sich die Entwicklungsstufe des Grundschulkinds kurz charakterisieren? Nach Piaget befindet es sich in seiner kognitiven Entwicklung im Stadium konkret-operatorischer Strukturen. Das Denken des Kindes ist in hohem Maße abhängig von gegebenen Informationen, seien sie konkret-anschaulich (z.B. in Bildform) oder sprachlich repräsentiert (MONTADA 1995, S.540). Das Bild und das Spiel sind die Medien der Wahl, um Informationen über die subjektive Sichtweise von Gesundheit und Krankheit zu erfahren. [12]

Untersuchungen zur Gedächtnisentwicklung (SCHNEIDER & BÜTTNER 1995) zeigen, daß Grundschulkinder bereits ein autobiographisches Gedächtnis besitzen, in welchem Erinnerungen an vielschichtige Erlebnisse mit starkem Selbstbezug enthalten sind (FIVUSCH 1993, HOWE & COURAGE 1993, 1997; LEICHTMAN 1999, NELSON 1993, 1997). Diese episodischen Langzeitgedächtnisinhalte sind mit dem semantischen Langzeitgedächtnisteil assoziiert, in welchem konzeptuelle Wissensinhalte gespeichert sind, wie z.B. Sprache, Regeln, Begriffe. Kinder erwerben zwischen dem fünften und achten Lebensjahr eine metalinguistische Bewußtheit für Sprachkategorien und -regularitäten (GRIMM 1995; KARMILOFF-SMITH 1985, 1992). Diese Kompetenzen wirken sich auf die Differenzierung autobiographischer Gedächtnisinhalte mit zunehmendem Alter aus. Verbale Daten könnten demnach aufschlußreich sein, wenn man etwas über subjektive Befindlichkeiten im Grundschulalter erfahren möchte. [13]

Wir entschieden uns für ein kombiniertes Verfahren:

- ein mündliches Interview mit Hilfe eines von dänischen Kooperationspartnern zusammengestellten "Gesundheitsprofils";
- ein von KRAUSE entwickeltes Bilderverfahren "Was ich gern tue", das das subjektive Erleben der Kinder beim Betrachten von Bildern, die alltägliche Tätigkeiten darstellen, erfaßt (vgl. KRAUSE 1998);
- einen Satzergänzungstest, ein bekanntes und häufig verwendetes Verfahren zur Erfassung von Selbstbildinhalten. Entsprechend dem Anliegen des Projekts und unter Berücksichtigung des Alters der Kinder wurden folgende acht Satzanfänge ausgewählt:

1. "Wenn ich etwas nicht schaffe ..."	5. "Ich bin traurig ..."
2. "Ich finde nicht schön ..."	6. "Ich ärgere mich ..."
3. "Die anderen Kinder ..."	7. "Am meisten freue ich mich ..."
4. "Ich habe Angst ..."	8. "In der Schule ..." [14]

2.3 Der Satzergänzungstest – Merkmale des qualitativen Verfahrens

Der Satzergänzungstest ist ein halbstrukturiert-offenes Erhebungsverfahren. Das Kind soll frei antworten und entscheidet über den Inhalt der Aussage. HAUSSER (1982) spricht von "Verbalisierungschancen" des Befragten. Die Satzanfänge lenken die Aufmerksamkeit des Kindes auf erfahrungsbezogene Gedächtnisinhalte. Es werden erwartungsgemäß jene Inhalte häufig genannt, die besonders leicht zugänglich und in der jeweiligen Befragungssituation salient sind. Der Satzergänzungstest ist somit ein individuumzentriertes Verfahren: Die Perspektive des Individuums ist zentral. Diese Orientierung ist für die Selbstkonzeptforschung unbedingt notwendig (vgl. WIECHARDT 1977 und HAUSSER 1995). [15]

Die Halbstrukturiertheit des Verfahrens gewährleistet eine inter- und intraindividuelle Vergleichbarkeit der Antworten. Für junge Kinder ist es besonders geeignet, da sie einerseits frei antworten können, andererseits aber durch die Satzanfänge zur Selbstaussage angeregt werden. Die verbalen Daten sind sowohl inhaltlich auswertbar (durch Applikation eines inhaltlichen Kategoriensystems, s.u.) als auch qualitativ beurteilbar im Sinne von sprachlicher Komplexität/Variabilität. So ist zum Beispiel zu vermuten, daß die selbstbezogenen Antworten mit zunehmendem Alter an Differenziertheit zunehmen. Der Differenzierungsgrad innerhalb einer Stichprobe gibt Aufschluß über die kognitive Entwicklung. [16]

Beim Satzergänzungstest wurden die Kinder ermuntert, aktuelle selbstbezogene episodische Gedächtnisinhalte, die über die gegenwärtige subjektive Befindlichkeit Aufschluß geben, zu aktivieren. Die Auswahl der Satzanfänge war schwierig, nach mehreren Probedurchläufen entschieden wir uns für die oben genannten acht Sätze. Die Kinder wurden im Kontext Schule befragt. Die Interviewerinnen waren vertraute Personen, die sie über ein oder mehrere Schuljahre in den Gesundheitsstunden begleiteten. Es war trotzdem nicht zu vermeiden, daß die Beziehung zur Versuchsleiterin und die situativen Kontextbedingungen das Antwortverhalten der Kinder wesentlich beeinflußten. Diese Situationsspezifik, die bei Befragungen immer gegeben ist, wurde jedoch dadurch nivelliert, daß die Erhebung jedes Jahr wiederholt und daß der Test von den Kindern insgesamt fünfmal bearbeitet wurde. [17]

3. Die Verbindung von qualitativer und quantitativer Forschung

Die häufig diskutierte Frage, ob qualitative oder quantitative Forschung die bessere sei, ist unserer Meinung nach eine künstliche. Entscheidend ist, inwiefern ein Erhebungsverfahren dem Untersuchungsgegenstand gerecht wird und gleichzeitig die oben genannten Methodenprobleme löst (KRIPPENDORF 1980 hat in diesem Zusammenhang acht inhaltsanalytische Gütekriterien vorgestellt). Eine Möglichkeit, interindividuelle Vergleichbarkeit zu erhöhen, besteht in der Formalisierung der Datenanalyse unter Rückgriff auf mathematische Methoden. Ein geringer Formalisierungsgrad der Datenauswertung bewahrt (zu einem gewissen Grad) die Natur qualitativer

Daten. "Auch in qualitativ orientierten humanwissenschaftlichen Untersuchungen können – mittels qualitativer Analyse – die Voraussetzungen für sinnvolle Quantifizierungen zur Absicherung und Verallgemeinerbarkeit der Ergebnisse geschaffen werden." (MAYRING 1993, S.24). [18]

Der Untersuchungsgegenstand in unserem Falle war die Erfassung des subjektiven Wohlbefindens der Kinder und die Veränderung dieser Befindlichkeit im Verlaufe der Grundschuljahre. In einer westdeutschen und einer ostdeutschen Stadt mittlerer Größe (Göttingen und Greifswald) wurden jeweils an mehreren Schulen, die innerhalb der Orte wiederum unterschiedliche Stadtteil-Lagen repräsentierten, Totalerhebungen durchgeführt. Dieses Design des "Cluster"-Sampling ist zwar nicht repräsentativ, wird aber in Bezug auf Schul-Populationen häufig angewandt, um zu verallgemeinerbaren Aussagen zu kommen. [19]

Um die Subjektivität von Befindlichkeiten nur annähernd erfassen und die individuelle Entwicklung nachvollziehen zu können, sind qualitative Verfahren unerläßlich. Wenn aber gleichzeitig interessiert, ob eine bestimmte Intervention (in unserem Falle das Förderprogramm) sinnvoll ist, sind verallgemeinernde Aussagen notwendig. Deshalb haben wir die Evaluation auf einer breiten Datenbasis durchgeführt und mit quantitativen, formalisierten Verfahren ausgewertet. Dieses Vorgehen könnte gerade für anwendungsorientierte Forschung wie die der Erziehungswissenschaft ein Weg sein, um die bisher häufig praktizierte Dichotomisierung zwischen qualitativen und quantitativen Zugängen zu überwinden und eine Integration zu erreichen. "Quantität für sich ist sinnlos, Qualität für sich genommen bleibt folgenlos" (HUBER 1989). [20]

Für die Erforschung selbstbezogener Inhalte eignet sich ein kombiniertes methodisches Vorgehen (HAUSSER 1995). So kann einerseits die Bedeutung qualitativer Daten erschlossen werden, andererseits können Interpretationsprozesse systematisiert und dokumentiert sowie Befunde quantifizierend geordnet werden (vgl. HUBER 1989). [21]

Durch den Satzergänzungstest haben sich letztlich weit über Tausend auszuwertende Aussagen pro Satz ergeben. Berücksichtigt man außerdem, daß im Verlauf der Projektzeit unterschiedliche Mitarbeiter/innen an der Auswertung beteiligt waren, kann das Problem der Zuverlässigkeit der Auswertung nicht durch die übliche Methode der "Diskussion bis zur Übereinstimmung" über kontroverse Texte[1] gelöst werden. Die Auswertung dieser Daten sollte deshalb durch Maßzahlen für die Reliabilität abgesichert sein, die mit ähnlichen Maßen standardisierter Erhebungsinstrumente vergleichbar sind. [22]

Bei der Methode der Auswertung entschieden wir uns für die qualitative Inhaltsanalyse (MAYRING 1993). Die wichtigste Prämisse dafür war, daß die Kategorienbildung durch prozeßimmanente Auseinandersetzung mit dem empirischen Material reflektierbar und kontrollierbar sein sollte. Hierfür ist aber wichtig, daß Reliabilitäts- und Validitätsbestimmungen bei veränderten

1 So z.B. in den Projekten von HOPF, RIEKER & SANDEN-MARTENS 1995 und HEITMEYER, BUHSE, LIEBE-FREUND, MÖLLER, MÜLLER, RITZ, SILLER & VOSSEN 1992.

Bedingungen erneut aufgenommen und mit früheren Ergebnissen verglichen werden können, vor allem weil es sich ja um ein Längsschnitt-Projekt mehrjähriger Dauer handelt. Dazu eignen sich ebenfalls gut quantitative Maßzahlen. [23]

Die zu bewältigenden Kategorisierungen und die Notwendigkeit der Kontrolle über die Qualität des Kategorienschemas und der Kodierungen auch bei Veränderungen des Schemas, stellten uns vor die Aufgabe, die Kodierleistung im Verlauf des gesamten Projekts quantitativ zu erfassen. Dazu wurden während der Entwicklung und Anwendung des Kategorienschemas mehrfach quantitative Maßzahlen erhoben, die in den Vorgang der Weiterentwicklung eingingen. [24]

Im Folgenden soll diese Methode der Verflechtung von Quantität und Qualität bei der Entwicklung von Kategorienschema und Prüfung der Kodierleistung dargestellt werden. Als Beispiel für das Vorgehen soll die Entwicklung der Kategorien zu den von den Kindern geäußerten Satzergänzungen für den ersten Satzanfang "Wenn ich etwas nicht schaffe" dienen (vgl. KRAUSE & MÜLLER-BENEDICT 1997). [25]

Die Entwicklung des Kodierleitfadens erfolgte in einem zweistufigen Verfahren. In einem ersten Durchlauf mit Material aus nur wenigen Erhebungseinheiten wurden einfache statistische Maßzahlen und Kreuztabellen der Übereinstimmung von je zwei Kodierer/innen bestimmt und der Grad der Interkoderreliabilität zwischen allen beteiligten Kodierer/innen und insgesamt gemessen. Auf dieser Grundlage wurde eine verbesserte Version des Kodierleitfadens erstellt. Vor allem wurden jene Fälle herausgearbeitet, in denen große Streuungen bei der Zuordnung auftraten. Dieses Vorgehen wird in Abschnitt 3.1 und 3.2 geschildert. [26]

Nachdem durch diesen Materialdurchlauf eine verbesserte Version des Kodierleitfadens entstanden war, wurde diese im zweiten Durchlauf mit erheblich mehr Material auf Meßgenauigkeit hin überprüft. Eine Gruppe Kodierer/innen kodierte die früheren Satzergänzungen ein zweites Mal mit der neuen Version des Kodierleitfadens. Zwischen der Bearbeitung mit der ersten und der zweiten Version lagen mindestens drei Monate. Die Meßgenauigkeit dieses zweiten Durchlaufs und die Verbesserung gegenüber dem ersten wurden mit standardisierten Maßzahlen der Interkoderreliabilität validiert. Das Verfahren der Messung der Interkoderreliabilität wird in Abschnitt 3.2 geschildert. [27]

3.1 Die Entwicklung des Kategorienschemas

Auf Grund theoretischer Vorüberlegungen legten wir Kategorien fest, die relativ global waren und nur als vorläufige Orientierung dienten, jedoch der Forderung nach einem einheitlichen Klassifikationsprinzip gerecht wurden (HOLSTI 1969, MERTEN 1983).

"Wenn ich etwas nicht schaffe ..."

Kategorie 1: Mit Hilfe zum Erfolg

Kategorie 2: Ohne Hilfe zum Erfolg

Kategorie 3: Mißerfolg zulassen

Kategorie 4: Mißerfolg mit Selbstbewertung

Kategorie 5: Konkrete Situationsbeschreibung

Kategorie 6: Konsequenzen aus dem Nicht-Schaffen

Kategorie 7: Sonstiges

Kategorie 8: Keine Antwort

Abbildung 1: Vorläufiges Kategorienschema [28]

Zunächst wurden die Ergänzungen von 72 Kindern (Erhebungseinheit einer Schule in Göttingen) aller acht Sätze so aufgeteilt, daß jeweils zwei Mitarbeiter/innen einen Satz bearbeiteten und empirisch begründete Kategorien erstellten. In mehreren Beratungen der Forschungsgruppe wurden diese Vorschläge diskutiert. Anschließend bearbeiteten zwölf Kodierer/innen das Material mit dem folgenden Kategorienschema.

"Wenn ich etwas nicht schaffe ..."

Kategorie 1: Mit Hilfe zum Erfolg

Kategorie 2: Ohne Hilfe zum Erfolg

2a. Unmittelbares Wiederholen, Weitermachen, Anstrengen

2b. Aufschieben

Kategorie 3: Mißerfolg zulassen

3a. Ignorieren

3b. Rückzug, Vermeidung

3c. Ablenkung durch andere Tätigkeiten

Kategorie 4: Frustration / Mißerfolg mit Selbstwertbezug

4a. emotional

4b. kognitiv

Kategorie 5: Konkrete Situationsbeschreibung

Kategorie 6: Sonstiges (Residualklasse)

Kategorie 7: Keine Antwort

Abbildung 2: Überarbeitetes Kategorienschema [29]

3.2 Übereinstimmungsmaße und Kodierleitfaden-Entwicklung

Wie die weiteren Ausführungen zeigen werden, war es ein Problem, ein Kategorienschema zu entwickeln, welches die größtmögliche Inhaltsaufnahme gewährleisten konnte, aber außerdem den Ansprüchen an Reliabilität, Validität und Objektivität genügen mußte. Bei der Auswertung des ersten Materialdurchlaufs wurden folgende Werte berechnet: [30]

1. Modalwerte

Es wurden die Ergebnisse der Kodierungen daraufhin untersucht, ob es bei den Aussagen der Kinder eindeutige Modalwerte gibt, die so herausgehoben sind, daß man von einer überwiegenden Übereinstimmung der Kodierer/innen sprechen kann. [31]

2. Streuungen

Wenn es für bestimmte Antwortsätze keine solchen Modalwerte gab, also eine eindeutige Zuordnung nicht möglich war, wurde die Verteilung der Zuordnungen im Kategorienschema betrachtet, um mögliche Gründe dafür festzustellen. Es war zu überprüfen, ob die Kodierer/innen bei der Einordnung einer bestimmten Aussage unterschiedlicher Auffassungen sind, ob es Polaritäten zwischen zwei oder drei Kategorien bzw. Subkategorien gibt, die auf mangelnde Trennschärfe des Kategorienschemas hinweisen, also auf mangelhafte Erfüllung der Forderung nach Exklusivität und gegenseitiger Abgrenzbarkeit der Kategorien. [32]

Die folgende Matrix bietet die Möglichkeit, diese Fragen zu beantworten. Eine vergleichbare Methode, mit der alle, nicht nur die besonders stark streuenden Aussagen der Kinder, in Bezug auf obige Fragen analysiert werden können, bieten die Kreuztabellen der paarweisen Kodierungen (vgl. Abschnitt 3.3). Beim Satzanfang *"Wenn ich etwas nicht schaffe..."* wurden unter den 72 Ergänzungen der Stichprobe zwölf Sätze gefunden, für die keine Modalwerte bei einzelnen Kategorien existierten. Die Verteilung sah folgendermaßen aus[2]:

2 Bei den 12 schwierigen Satzergänzungen konnten sich einzelne Kodierer/innen nicht entscheiden und ließen die Entscheidung offen, so daß die Reihensumme nicht immer 13 ist.

Prob.Nr.	Kategorie											
	1	**2a**	**2b**	**2c**	**3a**	**3b**	**3c**	**4a**	**4b**	**5**	**6**	**7**
300					7	6						
301	3					1	1				6	1
319						1	1		6		4	
321	7		1		5	4						
324	7	9										
330					5	6		1				
338				2	2	5					4	
349								5		2	6	
353	5						7					
358					5	6			1			
364				1	1	2			3		6	
266					1	9	3					

Tabelle 1: Matrix der Verteilung von zwölf kodierten Aussagen, bei denen keine eindeutige Zuordnung zu Kategorien erfolgte [33]

Als wesentlicher Faktor für die Unstimmigkeiten bei den 12 Aussagen stellte sich die mangelnde Trennschärfe einiger Kategorien heraus. Außerdem wurde deutlich, daß eine längere Schulung der Kodierer/innen notwendig war, um ausreichende Kompetenzen zu erwerben. [34]

Gut erkennbar ist die Polarität zwischen den Subkategorien 3a ("Ignorieren") und 3b ("Rückzug/Vermeidung"), z.B. bei den Aussagen der Probanden Nr.:

300: "... *dann schaff's ich eben nicht*."

321: "... *dann laß ich's, oder frage andere, ob sie mir helfen*."

330: "... *dann lasse ich es*."

358: "... *dann schaff ich das nicht*." [35]

Die Polarität beruht offensichtlich darauf, daß es für die Kodierer/innen schwierig war zu entscheiden, ob z.B. der Satz 300 *"... dann schaff ich's eben nicht"* eine Aussage ist, die Rückzug oder Vermeidung signalisiert oder ob dieser Satz auch als Ignorieren einer Mißerfolgssituation verstanden werden kann. An diesem Beispiel wird deutlich, daß die Forderung nach Exklusivität und gegenseitiger Abgrenzbarkeit der Subkategorien nicht gewährleistet war. Infolge der Diskussion des Problems einigte sich die Gruppe darauf, daß die beiden Unterkategorien 3a und 3b zusammengefaßt werden sollen, da in jedem Fall Ignorieren und Vermeiden bzw. Zurückziehen als ein Verhalten angesehen werden kann, bei dem das Kind weder um eine Lösung des Problems bemüht ist, noch sich weitere

Gedanken darüber macht, also eine Beeinträchtigung des Selbstwertgefühls nicht erkennbar ist. Die Trennschärfe zu den anderen Kategorien bleibt auf jeden Fall erhalten. Die neue Unterkategorie faßt somit alle Möglichkeiten zusammen, die vorher unter 3a und 3b genannt wurden.

Wenn ich etwas nicht schaffe ... **Kategorie 3: "Mißerfolg zulassen"**	
1. Version: Definition: Diese Kategorie beschreibt Aussagen, die erkennen lassen, daß der Mißerfolg keinen Impetus und auch keine erkennbare Valenz für das Kind besitzt. Es wurden innerhalb dieser Kategorie drei verschiedene Vorgehensweisen angesichts des Mißerfolgs beobachtet, die in folgenden drei Unterkategorien ihren Niederschlag finden. 3a: Ignorieren Ankerbeispiel: "... dann laß ich es eben, das meiste schaffe ich sowieso." 3b: Rückzug/Vermeidung Ankerbeispiel: "... dann mach ich das nicht" 3c: Ablenkung durch andere Tätigkeiten Ankerbeispiel: "...dann mache ich ein Spielchen"	*2. Version:* Definition: Aussagen, die erkennen lassen, daß der Mißerfolg (etwas nicht schaffen) keine erkennbare Valenz für das Kind besitzt, werden dieser Kategorie zugeordnet. Dabei können zwei Vorgehensweisen angesichts des Mißerfolgs beobachtet werden. 3a: Ignorieren, Rückzug, Vermeidung Ankerbeispiele: "... dann laß ich das eben, das meiste schaff ich sowieso", "... dann mach ich das nicht." 3b: Ablenkung durch andere Tätigkeiten/ einen anderen Ort aufsuchen Ankerbeispiele: "... dann mache ich ein Spielchen", "... dann gehe ich nach Hause".

Abbildung 3: Ausschnitt aus dem Kodierleitfaden für Satz 1 vor und nach der letzten Überarbeitung [36]

Ein anderes Problem, welches ebenfalls auf mangelnde Trennschärfe zurückzuführen ist, trat bei vier Aussagen in der Kategorie 1 ("Mit Hilfe zum Erfolg") auf:

> 301: "... *irgendwas sagen*."
>
> 321: "... *dann lass ich's oder frage andere, ob sie mir helfen*"
>
> 324: "... *dann versuch ich es nochmal und dann frag ich Papa*."
>
> 353: "... *dann gehe ich zu einer Freundin*." [37]

Die Aussagen 321 und 324 beinhalten jeweils zwei Sinnzusammenhänge. In beiden Sätzen werden zwei Möglichkeiten zur Bewältigung des eigenen Unvermögens vom Kind in Betracht gezogen. Die Einordnung in zwei Kategorien ist somit kein Problem der Trennschärfe des Kodierleitfadens, da in den Aussagen zwei verschiedene Inhalte genannt werden. [38]

Die Aussage 301 ist offensichtlich so unverständlich, daß sie nicht eindeutig zuzuordnen ist. Die meisten Kodierer/innen haben sie deshalb auch in die dafür vorgesehene Kategorie 6 ("Sonstiges") eingeordnet. Drei Kodierer/innen aber haben diese Aussage der Kategorie 1 zugeordnet und sind damit der Meinung, daß die Ergänzung des Satzanfanges "*Wenn ich etwas nicht schaffe ...*" durch "*... irgendwas sagen*" ein Hilfeersuchen des Kindes zum Ausdruck bringt. Es wären sicherlich auch noch andere Interpretationen möglich. In der Besprechung, die nach dem ersten Materialdurchlauf erfolgte, wurde deshalb vor allem das Problem des Interpretierens oder des Suchens bzw. Vermutens von Sinn diskutiert. [39]

Der Satz 353 wurde trotz seiner klaren Formulierung nicht eindeutig zugeordnet. Hier besteht Polarität zwischen den Kategorien 1 ('Mit Hilfe zum Erfolg') und 3c ('Ablenkung durch andere Tätigkeiten'). Fünfmal wurde die Aussage "*... dann gehe ich zu einer Freundin*" der Kategorie 1 und siebenmal der Kategorie 3c zugeordnet. Hier wird nun deutlich, wie der Spielraum unterschiedlicher Interpretationsmuster der Kodier/innen beim Überprüfen der Interkoderreliabilität zum Tragen kommt. Die Aussage des Kindes "*... dann gehe ich zu einer Freundin*" ist eigentlich ein klar formulierter Satz, der aber trotz seiner konkreten Aussage nicht eindeutig erschlossen werden kann. Sicherlich kann ein Kind zu seiner Freundin gehen, um Hilfe zu bekommen. Es wäre aber auch möglich, daß es zu seiner Freundin geht, um dort Ablenkung von der nicht zu bewältigenden Aufgabe zu erfahren. Das aber sind Interpretationen, die zu subjektiven Entscheidungen bei der Kategorienwahl führen. Dieses Problem mußte durch eine eindeutigere Formulierung der Kategorie gelöst werden. In diesem Falle erhielt die Unterkategorie 3c den Zusatz "einen anderen Ort aufsuchen". Diese Entscheidung erscheint auch gerechtfertigt, da mehr Kodierer/innen die Kategorie 3c (sie wurde in der Endfassung zu 3b) als 1 gewählt haben. [40]

Nach der Überarbeitung wurde der Kodierleitfaden erstellt, der nunmehr auch Definitionen und Ankerbeispiele enthielt (vgl. Beispiel in Abb.3). [41]

Danach wurden die Texte nochmals kodiert, diesmal von der sog. "Kodierergruppe", bestehend aus vier Mitarbeiterinnen und einem Mitarbeiter der Forschergruppe. Wichtig ist in diesem Zusammenhang, daß diese fünf Kodiererinnen und Kodierer von Anfang an in das Forschungsvorhaben einbezogen waren (im Rahmen eines Forschungspraktikums und später als studentische bzw. wissenschaftliche Hilfskräfte). Die Ergebnisse der Erstkodierung dieser Kodierergruppe und die der Zweitkodierung wurden einer Reliabilitätsprüfung unterzogen, was im nächsten Abschnitt näher beschrieben wird. In einem dritten Durchgang setzte sich die Kodierergruppe zusammen, diskutierte auf der Grundlage der einzelnen Kodierergebnisse jede Aussage und legte eine endgültige Zuordnung fest. Wir gingen dabei immer so vor, daß pro Sitzung lediglich ein Satz diskutiert wurde, so daß zunächst eine Beeinflussung der Kodierleistung durch den Gesamteindruck der acht Satzergänzungen eines Kindes vermieden wurde. Als Ergebnis dieser Arbeit lag eine vollständige Zuordnung aller Zitate zu den in den Kodierleitfäden vorgegebenen Kategorien vor. [42]

Erst bei spezifischen Analysen wurden alle Aussagen des Kindes – die acht Satzergänzungen zu allen Meßzeitpunkten – herangezogen. Eine solche Analyse zum Beispiel ergab, daß einige Kategorien innerhalb des ersten Satzes ("Wenn ich etwas nicht schaffe ..."), des dritten Satzes ("Die anderen Kinder ...") und des achten Satzes ("In der Schule ...") besonders aufschlußreich waren, um Kinder heraus zu finden, deren Selbstwertgefühl gefährdet war. Bei Schüler Nr. 162 war die Ergänzung des Satzanfanges "Wenn ich etwas nicht schaffe ..." mehrmals der Kategorie 4a "Mißerfolg mit Selbstbewertung (emotional)" zuzuordnen. Er führte den Satzanfang folgendermaßen fort:

> "Wenn ich etwas nicht schaffe ...
>
> ... dann schimpft Mama ein bißchen." (Kindergarten)
>
> ... dann mag ich's nicht gern und hab ein komisches Gefühl." (1. Schuljahr)
>
> ... kommt darauf an was, z.B. Hausaufgaben: fühl' ich mich erst mal nicht so gut." (2. Schuljahr)
>
> ... dann fühle ich mich nicht gut." (3. Schuljahr)
>
> ... dann mach ich es eigentlich nicht zu Ende." (4. Schuljahr) [43]

3.3 Überprüfung des Kategorienschemas mit Hilfe von Kreuztabellen von Kodierungen

Grundlage für die Berechnung von Maßzahlen für die Güte von Kodierungen sind Kreuztabellen der Kodierungen von Kodierer/innen-Paaren. Sie werden vom hier entwickelten Programm zur Berechnung der Interkoderreliabilität (s. nächsten Abschnitt) erstellt. Jede Zelle einer solchen Tabelle ist durch eine Kombination der von beiden Kodierer/innen jeweils angewandten Kategorie lokalisiert. Für jedes Zitat (eine Aussage) steht ein Punkt in derjenigen Zelle, die durch die von beiden Kodierer/innen für dieses Zitat gewählten Kategorie repräsentiert wird. Auch anhand dieser Tabellen lassen sich aufschlußreiche Hinweise für das Kategorienschema geben. Ein Beispiel: 52 Texte, von den beiden Kodiererinnen mit den Kürzeln "k" und "ve" kategorisiert (Kategorie "31" bedeutet "3a", "23" = "2c" in Kasten 3).

Kat.	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

Tabelle 2: Kreuztabelle der gemeinsamen Kodierungen der Kodiererinnen k, ve Übereinstimmungen: 37 --> Hinweis: In der Hauptdiagonalen befinden sich auch nicht-überstimmende Kategorien! [44]

Die Hauptdiagonale besteht – bis auf die (ohne die Summenzeilen) letzte Zelle unten rechts – aus den übereinstimmend gewählten Kategorien. Die Zeile und Spalte, die diese letzte Zelle bilden, in der die Kategorien nicht übereinstimmen, zeigen, daß zwei Kategorien jeweils nur von einer der Kodiererinnen gewählt wurde: Kodiererin "k" hat zweimal die "50" gewählt, "ve" dagegen gar nicht, und andererseits "ve" einmal die "23", "k" dagegen die 23 gar nicht. An der starken Besetzung der Hauptdiagonalen sieht man, daß hier eine gute Übereinstimmung besteht. Weiter erkennt man auf einen Blick, daß die Kategorie "10" ca. 50% aller Antworten ausmacht. Ob das im Hinblick auf die Aussagekraft der Kategorie gerechtfertigt ist, muß inhaltlich entschieden werden. Bei insgesamt ca. 10 Kategorien könnte man hier überlegen, die Kategorie "10" noch zu differenzieren. [45]

Auf Grund der Kreuztabellen lassen sich relativ einfach weitere Mängel des Kategoriensystems diagnostizieren, wenn man die Ergebnisse an mehreren Tabellen überprüft:

- Kategorien, die gar nicht oder nur sehr selten benutzt wurden (hier 23), sollten daraufhin überprüft werden, ob sie theoretisch notwendig sind.
- Bei gehäuften Nichtübereinstimmungen, bei denen immer die zwei selben unterschiedlichen Kategorien gewählt wurden (hier 31 und 32), sollte die Trennschärfe dieser Kategorien verbessert werden.
- Kategorien, die mit fast allen anderen kombiniert wurden (hier 33), sind ein Hinweis darauf, daß die Qualität dieser Kategorie fast allen Texten anhaften könnte, also nicht "wechselseitig exklusiv" genug ist. [46]

Die Kombination der obigen Matrix der Kodierungen und der Kreuztabellen der Kodier-Paare liefert nach den Projekterfahrungen reichhaltige Anhaltspunkte für fruchtbare Diskussionen zur qualitativen Verbesserung des getesteten Kategorienschemas und des dazugehörigen Kodierleitfadens. [47]

4. Die Messung der Kodierleistung

"Interkoderreliabilität" bezeichnet den "Grad an Übereinstimmung" zwischen Kodierer/innen. Hierfür wird i.A. ein sog. "kappa"-Koeffizient berechnet (KRIPPENDORF 1970, COHEN 1960). Es ergaben sich jedoch in diesem Projekt einige Besonderheiten bei der Messung eines "kappa", die sicher auch typisch für ähnliche sozialwissenschaftliche Projekte sind:

- Wie soll die Übereinstimmung zwischen mehr als zwei Kodierer/innen generell gemessen werden?
- Wie soll sie speziell gemessen werden, wenn nicht alle die gleiche Anzahl Texte (hier: Kinder) verkodet haben?
- Wie soll Übereinstimmung gemessen werden, wenn die Kodierer/innen nicht alle dieselben Kategorien verwendet haben?
- Wie lassen sich Änderungen am Kategorienschema im Hinblick auf Verbesserung/Verschlechterung messen? [48]

Für diese Probleme gibt es keine Verfahren, die soweit erprobt sind, daß sie in Lehrbüchern über Inhaltsanalyse und in der Standard-Software zu finden wären. Deshalb wurden ein eigener "kappa"-Koeffizient, der die obigen Probleme behandeln kann, und ein Computer-Programm für seine Berechnung entwickelt[3] (MÜLLER-BENEDICT 1998). [49]

4.1 Neudefinition von Kappa

Ein kappa-Koeffizient soll die Übereinstimmung zwischen zwei Kodierer/innen in einer Maßzahl zwischen 0 und 1 messen. Dabei ist 1 definiert als völlige Übereinstimmung, 0 definiert als die Übereinstimmung, die erwartet werden kann, wenn die Kodierer/innen die Kategorien zufällig auswählen. Je nachdem, was an Kodiermöglichkeiten zugelassen wird und was unter "zufällig" verstanden wird, sind hier verschiedene Berechnungsmöglichkeiten vorhanden. [50]

Es war zu klären, wie mit Satzergänzungen zu verfahren ist, die zwei Aussagen enthalten, z.B. "Wenn ich etwas nicht schaffe, gehe ich zu meiner Mama oder versuche es nochmal". Zum ersten Untersuchungszeitpunkt war dies nur dreimal der Fall und wir entschieden uns zunächst dafür, nur die erste Aussage auszuwählen. Diese Entscheidung mußten wir aber schon nach der zweiten Untersuchung revidieren, da dieser Fall häufiger auftrat. Es wurden beide

3 Das Programm und eine ausführlichere Beschreibung des hier verwendeten "kappas" kann man im Internet erhalten. Die Adresse lautet: http://www.uni-goettingen.de/~vbenedi [Broken link, FQS, December 2004]

Aussagen ausgewertet und für die Berechnung von "kappa" eine Zusatzkategorie ("Mehrfach-Aussage") erstellt. [51]

Eine weitere Entscheidung betraf die Wahrscheinlichkeit, mit der eine bestimmte Kategorie "zufällig" übereinstimmend gewählt wird. Für die Berechnung des verbreitetsten Koeffizienten, COHENS kappa (BOS & TARNAI 1989, S.183 u. 203), wird angenommen, daß diese Wahrscheinlichkeit von den Kodierer/innen abhängt. Sie wird als die Produktwahrscheinlichkeit aus den Häufigkeiten, mit denen diese Kategorie von jedem Kodierer und jeder Kodiererin gewählt wurde, berechnet[4]. Damit ergibt sich, daß eine Kategorie, die nur von einer Person und von den anderen nicht benutzt worden ist, auch eine Wahrscheinlichkeit der Übereinstimmung von 0 hat, also mit Sicherheit nie Übereinstimmung erzielt. [52]

Das ist in unseren Augen eine bei der Verkodung von Texten nicht vertretbare Wahl. Auch wenn die Kategorie nur von einer Person benutzt wurde, wird deutlich, daß dem Text Eigenschaften dieser Kategorie durchaus anhaften, also auch Übereinstimmung mit einer zwar kleinen, aber positiven Wahrscheinlichkeit möglich gewesen wäre. Bei dieser Argumentation haben wir die Herkunft der Wahrscheinlichkeit für zufällige Übereinstimmung anders verortet. Sie liegt im Text begründet und nicht in den Eigenarten der Kodierer/innen. Sie ist deshalb nach SCOTT (1957) anders zu berechnen, und zwar als Produkt der gemittelten Häufigkeit der Benutzung dieser Kategorie durch beide Kodierer/innen[5]. Das scheint uns eine generalisierbare Überlegung für die Verkodung jeder Art von sozialwissenschaftlichen Texten in der Inhaltsanalyse im Unterschied zur Verkodung von z.B. Handlungssequenzen (bei Beobachtungen), Patientenäußerungen (in der Psychologie), Echtzeit-Interviews (ohne Transkription während des Interviews) zu sein. In diesen Fällen nämlich ist es durchaus vorstellbar, daß eine Kodiererin bestimmte Kategorien nicht anwenden kann, da sie z.B. für sie nicht sichtbar oder hörbar sind, weil sie zu kurz auftreten, oder nicht benutzbar sind, weil sie sie auf Grund der psychologischen Konstellation verdrängt. Das sollte aber gerade bei der Text-Verkodung ausgeschlossen sein; die Kategorien sollten allen Kodierer/innen gleichermaßen offen stehen, so daß die Wahrscheinlichkeit der Wahl einer Kategorie nur vom Text abhängt[6]. [53]

4.2 Kappa-Berechnungen der Kodierleistung

Mit diesen Annahmen ist die Berechnungsmethode des Übereinstimmungsmaßes für zwei Kodierer/innen – das von SCOTT vorgeschlagene kappa – festgelegt. Damit ist gleichzeitig die Frage geklärt, wie zu verfahren ist, wenn zwei Kodierer/innen nicht dieselben Kategorien angewendet haben. Die nur von einem

4 Ein Beispiel: Haben Kodierer A und B je 100 Texte verkodet, und hat A die Kategorie i 20 mal, B sie 30 mal angewendet, so ist die Wahrscheinlichkeit p_i, daß sie "zufällig" übereinstimmend angewendet wird $p_i = (20/100)\times(30/100) = 3/50$

5 Dann gilt (s. Anmerkung 3): $p_i = ((20 + 30)/(100+100)) \times ((20+30)/(100+100)) = 1/16$ (= 3/48 im Vergleich zu $p_i = 3/50$ in Anmerkung 4)

6 HUBERT (1977, S.295) behandelt diesen Fall als "Levenes Modell" und bemerkt dazu: "Levenes notion may be generally more popular in the social sciences than either of the two matching concepts presented earlier." (s.a. KRIPPENDORF 1970)

angewandten Kategorien gehen mit einer kleinen, aber positiven Wahrscheinlichkeit in die Berechnung der zufälligen Übereinstimmung ein[7]. [54]

Zu bestimmen bleibt, wie die Übereinstimmung zwischen mehreren Kodierer/innen gemessen werden soll. Die Überlegung war hier, daß eine Maßzahl dafür die Eigenschaft haben müßte, gleich zu bleiben, wenn zu einer Gruppe von Kodierer/innen ein weiterer Kodierer, der eine vergleichbare Kodierleistung wie die anderen aufweist, hinzustößt. Dann kann z.B. festgestellt werden, ob sich die Gruppenkodierleistung durch Ersatz eines Kodierers durch einen anderen erhöht, oder ob sich z.B. durch die Einstellung weiterer Kodierer/innen die Gesamtkodierleistung verschlechtert. Damit ist vorgezeichnet, daß das kappa für mehrere Kodierer/innen ein "Durchschnitt" aus allen paarweisen Kodierungen sein sollte. Die Berechnung dieses Durchschnitts muß so gestaltet sein, daß die "0" wieder die erwartete zufällige Übereinstimmung nach SCOTT darstellt. [55]

Mit diesem Koeffizienten lassen sich Maßzahlen von Übereinstimmungen für die meisten inhaltsanalytischen Verkodungen bestimmen und über die Texte hinweg vergleichen. Im Allgemeinen gelten Maße von über 0.7 als akzeptabel oder sogar als gut (BAKEMAN & GOTTMAN 1986, S.82), da z.B. auch bei standardisierten Befragungen in Retests u.ä. die Reliabilität im Durchschnitt bei 70% liegt (KÖNIG 1973, S.175). [56]

Über die Messung und Sicherung der Kodierqualität hinaus lassen sich damit aber auch die Verbesserungen eines Kategorienschemas messen. Die Änderungen am Kategorienschema repräsentieren im Kern den Fortgang der qualitativen Auswertung der empirischen Ergebnisse. Sie ziehen sich deshalb oft über eine beträchtliche Zeit des Forschungsprojekts hin. Deshalb ist es möglich, sogar dieselben Texte, die am Anfang des Projekts mit dem zu diesem Zeitpunkt vorliegenden Kodierleitfaden verkodet wurden, mit dem endgültigen Kodierleitfaden noch einmal zu verkoden und jeweils den Grad an Übereinstimmung der Kodierer/innen zu messen. In unserem Projekt war es notwendig, den Kodierleitfaden zu erweitern, da die Aussagen der Kinder sich inhaltlich veränderten (z.B. kamen Aussagen zur Schule und zum Lernen erst später hinzu) und an Differenziertheit zunahmen. Eine substantielle Erhöhung der kappa-Koeffizienten deutet dann auf eine Verbesserung des Kategorienschemas hin, soweit man sie nicht ausschließlich dem "Training" und "Lernerfolg" der Kodierer/innen zuschreiben will. [57]

Nach dem ersten Durchgang der Verkodung wurden in diesem Projekt die kappa-Koeffizienten für alle Paare und alle Kodierer/innen gemeinsam jeweils für alle acht zu verkodenden Sätze bestimmt. Es ergaben sich Werte von 0.6 bis 0.8, also schon ein recht zufriedenstellendes Ergebnis. Nach der Weiterentwicklung wurden alle Sätze mit dem endgültigen Kategorienschema und neuem Kodierleitfaden erneut verkodet. Das geschah in unabhängigen Einzelsitzungen. Auch von diesem Durchgang wurden die obigen kappa-Werte errechnet. Es

7 Habe z.B. B die Kategorie i nicht angewandt, so gilt (s. Anmerkung 4): $p_i = ((20 + 0)/(100 + 100))2 = 1/100$.

zeigte sich bei allen Sätzen, sowohl für die Paarungen als auch für die Gesamtwerte, im Durchschnitt eine mehr als 10%-ige Erhöhung von kappa, so daß die Verbesserung des Kategorienschemas zufriedenstellend war. Damit wurde eine Übereinstimmung, die sich durchaus mit standardisierten Befragungen vergleichen läßt, erreicht. Im einzelnen ergaben sich z.B. für die ersten beiden Sätze folgende Werte[8]:

1. Satz											
Paar	s,k	s,ve	s,ch	s,v	k,ve	k,ch	k,v	ve,ch	ve,v	ch,v	alle
1.Kd	.6431	.6605	.7613	.7883	.5928	.6391	.6692	.7872	.6362	.7154	.6913
2.Kd	.7901	.7869	.8115	.8327	.7824	.8550	.7570	.8780	.8025	.8032	.8103
2. Satz											
Paar	s,k	s,ve	s,ch	s,v	k,ve	k,ch	k,v	ve,ch	ve,v	ch,v	alle
1.Kd	.8070	.7900	.8482	.7800	.8090	.8007	.8243	.7833	.7273	.7154	.7881
2.Kd	.8569	.8736	.8733	.8413	.8855	.9045	.8564	.8862	.8732	.8888	.8683

Tabelle 3: Interkoderreliabilität der ersten und zweiten Kodierung des ersten und zweiten Satzes für alle Kodierer/innen-Paare und insgesamt [58]

Der Gewinn dieser umfangreichen Prüfung der Kodierer/innen-Leistung für das Projekt liegt nicht nur in der gewonnenen Sicherheit, die Kodierqualität zu messen, dauerhaft zu gewährleisten und mit anderen Reliabilitätswerten vergleichen zu können. Von Vorteil war auch, daß diese Prüfung über die inhaltlichen Kritiken hinausgehende Hinweise auf Mängel im Kategoriensystem, die erst bei der Quantifizierung sichtbar werden, gibt, und daß sie den Fortschritt bei der Weiterentwicklung des Kategorienschemas meßbar machen konnte. [59]

5. Zusammenfassung

Die Verbindung qualitativer und quantitativer Verfahren hat sich speziell für die Belange der Auswertung großer Mengen qualitativen Materials, wie sie bei der Evaluation des Gesundheitsförderprogramms an Grundschulen anfielen, als sinnvoll erwiesen. Der Einsatz qualitativer Erhebungsmethoden bei hohen Fallzahlen stellt spezielle Anforderungen an Flexibilität und Offenheit des Kategorienschemas und führt zu praktischen Schwierigkeiten bei der Sicherung des Qualitätsstandards der Kodierungen. Deshalb war es notwendig, standardisierende Verfahren zu entwickeln, die eine im Zeitverlauf gleichmäßige Forschungsleistung gewährleisten können. Sowohl der heuristische Einsatz von quantitativen Auszählungen zur Entdeckung von Unstimmigkeiten und Lücken im Kategorienschema als auch die Möglichkeit, Reliabilität in einer Maßzahl prüfen zu können, zahlten sich für die Weiterentwicklung der qualitativen Methodologie des Projekts aus. So war die wechselseitige Ergänzung qualitativer und quantitativer Forschung in diesem Fall besonders hilfreich, da quantitative

8 s, k, ve, ch, v sind die Kürzel für die Kodierer/innen.

Auffächerung explorativen Materials zur Verbesserung der Reliabilität der qualitativen Untersuchung führte. [60]

Literatur

Antonovsky, Aaron (1993). Gesundheitsforschung versus Krankheitsforschung. In Alexa Franke & Michael Broda (Hrsg.), *Psychosomatische Gesundheit: Versuch einer Abkehr vom Pathogenese-Konzept* (S.3-14). Tübingen: DGVT-Verlag.

Bakeman, Roger & Gottman, John Mordechai (1986). *Observing interaction. An introduction to sequential analysis*. Cambridge: University Press.

Bos, Wilfried & Tarnai, Christian (1989). Entwicklung und Verfahren der Inhaltsanalyse in der empirischen Sozialforschung. In Wilfrid Bos & Christian Tarnai (Hrsg.), *Angewandte Inhaltsanalyse in Empirischer Pädagogik und Psychologie* (S.1-13). Münster, New York: Waxmann.

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

Fivush, Robyn (1993). Developmental perspectives on autobiographical recall. In Gale S. Goodman & Bette L. Bottoms (Hrsg.), *Child victims, child witnesses: Understanding and improving testimony* (S.1-24). London: Guilford Press.

Grimm, Hannelore (1995). Sprachentwicklung – allgemeintheoretisch und differentiell betrachtet. In Rolf Oerter & Leo Montada (Hrsg.), *Entwicklungspsychologie* (S.705-757). Weinheim: Psychologie Verlags Union.

Haußer, Karl (1982). Forschungsinteraktion und Forschungskonzeption. In Günter L. Huber (Hrsg.), *Verbale Daten: Eine Einführung in die Grundlagen und Methoden der Erhebung und Auswertung* (S.61-78). Weinheim: Beltz.

Haußer, Karl (1995). *Identitätspsychologie.* Berlin: Springer.

Heitmeyer, Wilhelm; Buhse, Heike; Liebe-Freund, Joachim; Möller, Kurt; Müller, Joachim; Ritz, Helmut; Siller, Gertrud & Vossen, Johannes *(*1992). *Die Bielefelder Rechtsextremismus-Studie. Erste Langzeituntersuchung zur politischen Sozialisation männlicher Jugendlicher.* Weinheim: Juventa.

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

Hopf, Christel; Rieker, Peter & Sanden-Martens, Martina (1995). *Familie und Rechtsextremismus: Familiale Sozialisation und rechtsextremistische Orientierung junger Männer*. Weinheim: Juventa.

Howe, Mark L. & Courage, Mary L. (1993). On resolving the enigma of infantile amnesia. *Psychological Bulletin, 113,* 305-326.

Howe, Mark L. & Courage, Mary L. (1997). The emergence and early development of autobiographical memory. *Psychological Review, 104,* 499-523.

Huber, Günter L. (1989). Qualität versus Quantität in der Inhaltsanalyse. In Wilfrid Bos & Christian Tarnai (Hrsg.), *Angewandte Inhaltsanalyse in Empirischer Pädagogik und Psychologie (S.1-13).* Münster, New York: Waxmann.

Hubert, Lawrence (1977). Kappa revisited. *Psychological Bulletin, 84*, 289-297.

Karmiloff-Smith, Annette (1985). Language and cognitive processes from a developmental perspective. *Language and Cognitive Processes*, *1*, 61-85.

Karmiloff-Smith, Annette (1992). *Beyond modularity. A developmental perspective on cognitive science.*Cambridge, MA: MIT Press.

König, Rene (Hrsg.) (1973). *Handbuch der empirischen Sozialforschung*. Bd. 1: Geschichte und Grundprobleme der empirischen Sozialforschung. Stuttgart: Enke

Krause, Christina & Müller-Benedict, Volker (1997). Ergebnisse und Probleme qualitativer Datenanalysen im Kontext eines Programmes zur Gesundheitsförderung. *Empirische Pädagogik, 11*(1), 31-61.

Krause, Christina (1998). Ich bin Ich. Gesundheitsförderung durch Selbstwertstärkung. Bericht über ein Projekt zur Gesundheitsförderung in Grundschulen. *Göttinger Beiträge zur erziehungswissenschaftlichen Forschung, Nr. 15*, Pädagogisches Seminar der Georg-August-Universität Göttingen.

Krippendorf, Klaus (1970). Bivariate agreement coefficients for reliability of data. In Edgar F. Bortatta (Hrsg.). *Sociological Methodology* (S.139-150). San Francisco: Jossey-Bass.

Krippendorff, Klaus (1980). *Content analysis. An introduction to its methodology.* Beverly Hills: Sage.

Leichtman, Michelle D. (1999). Cultural, social, and maturational influences on childhood amnesia. In Lawrence Balter, Catherine S. Tamis-LeMonda et al. (Hrsg.), *Child psychology: A handbook of contemporary issues* (S.447-466). Philadelphia, PA: Psychology Press/Taylor & Francis.

Lisch, Ralf & Kriz, Jürgen (1978). *Grundlagen und Modelle der Inhaltsanalyse. Bestandsaufnahme und Kritik*. Frankfurt/M: rororo.

Mayring, Philipp (1993). *Einführung in die qualitative Sozialforschung.* Weinheim: Psychologie Verlags Union.

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

Montada, Leo (1995). Die geistige Entwicklung aus der Sicht Jean Piagets. In Ralf Oerter & Leo Montada (Hrsg.), *Entwicklungspsychologie* (S.518-560). Weinheim: Psychologie Verlags Union.

Müller-Benedict, Volker (1998). Neue Berechnungsmethode der Interkoderreliabilität. *ZSE - Zeitschrift für Sozialisationsforschungs und Erziehungssoziologie*, *1*, 105

Mummendey, Hans-Dieter (1979). Methoden und Probleme der Messung von Selbstkonzepten. In Sigrun-Heide Filipp (Hrsg.), *Selbstkonzept-Forschung: Probleme, Befunde, Perspektiven* (S.171-189). Stuttgart: Klett.

Nelson, Katherine (1993). The psychological and social origins of autobiographical memory. *Psychological Science, 4,* 7-14.

Nelson, Katherine (1997). Finding one's self in time. In Joan Gay Snodgrass & Robert L. Thompson (Hrsg.), *The self across psychology: Self-recognition, self-awareness, and the self concept. Annals of the New York Academy of Sciences, Vol. 818* (S.103-116). New York, NY: New York Academy of Sciences.

Schneider, Wolfgang & Büttner, Gerhard (1995). Entwicklung des Gedächtnisses. In Rolf Oerter & Leo Montada (Hrsg.), *Entwicklungspsychologie* (S.654-704). Weinheim: Psychologie Verlags Union.

Scott, William A. (1955). Reliability of content analysis: The case of nominal scaling. *Public Opinion Quarterly, 19*, 321-325.

Wiechardt, Dörte (1977). Zur Erfassung des Selbstkonzepts. *Psychologische Rundschau*, *28*, 294-304.

Zur Autorin und zu den Autoren

Dr. Christina KRAUSE, Dipl.-Päd., Professorin für Pädagogische Psychologie am Pädagogischen Seminar der Georg-August-Universität Göttingen, Schwerpunkt "Diagnose und Beratung"

Forschungsschwerpunkte: Entwicklung des Selbst im Kindes- und Jugendalter, Gesundheitsförderung in der Schule, Lebens- und Berufsorientierung von Jugendlichen. Zu letzteren läuft ein vom DAAD gefördertes vierjähriges Kooperationsprojekt mit der Universität Monterrey (Mexiko)

Kontakt:

Christina Krause

Pädagogisches Seminar
Baurat-Gerber-Straße 4-6
D - 37073 Göttingen

Tel.: +49 / 0551 / 399 455

E-Mail: ckrause@gwdg.de oder
Dr.ChristinaKrause@t-online.de

Dr. disc. pol. Volker MÜLLER-BENEDICT, Dipl.-Math., Privatdozent, Assistent am Soziologischen Seminar der Universität Göttingen

Forschungsgebiete: Bildungsforschung, quantitative Methodologie, formale Modellbildung

Neuere Veröffentlichungen:

- Strukturelle Grenzen sozialer Mobilität. Ein Modell des Mikro-Makro-Übergangs nach Boudon, Kölner Zeitschrift für Soziologie und Sozialpsychologie 51(1999), 313-338
- Bedingungen selbstorganisatorischer Prozesse, ZUMA-Nachrichten 41(1997), 44-72

Kontakt:

Volker Müller-Benedict

E-Mail: vbenedi@uni-goettingen.de

Dr. phil. Ulrich WIESMANN, Dipl.-Psych., Wissenschaftlicher Assistent am Institut für Medizinische Psychologie der Universität Greifswald

Forschungsschwerpunkte: Salutogenese, Selbstwertstärkung im Grundschulalter, Körperwahrnehmung und Gesundheit im Alter, Gesundheitsbewußtsein und Gesundheitsverhalten junger Erwachsener, Multiple Sklerose und Arbeitsmotivation, Multiple Sklerose und kognitive Anpassung

Kontakt:

Ulrich Wiesmann

E-Mail: wiesmann@mail.uni-greifswald.de

Zitation

Krause, Christina; Müller-Benedict, Volker & Wiesmann, Ulrich (2000). Kleine Kinder – große Datenmengen. Möglichkeiten der Verbindung von qualitativen und quantitativen Methoden zur Analyse von Selbstaussagen [60 Absätze]. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research*, *1*(2), Art. 16, http://nbn-resolving.de/urn:nbn:de:0114-fqs0002165.

Revised 7/2008