FQS FORUM: QUALITATIVE SOCIAL RESEARCH SOZIALFORSCHUNG

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

Small Children—Big Data. Possible Links Between Qualitative and Quantitative Methods in the Analysis of Self-Reports

Christina Krause, Volker Müller-Benedict & Ulrich Wiesmann

Key words: project evaluation, self-image, self-worth, self-reports, health promotion, primary school children, longitudinal study, qualitative content analysis, intercoder reliability, kappa-coefficient

Abstract: The evaluation of a health promotion program of several years for primary school children resulted in two problems. First, qualitative instruments had to be developed for a population (children at the age between 5 and 10), for which standardized procedures would not be suitable. Second, the program was tested in a total of 20 school classes, and longitudinal verbal and pictorial-based data were collected. Thus, over a period of almost four years, an enormous amount of qualitative data were collected. New procedures were developed in order to analyze these qualitative data comprehensibly in a quantitative way.

In addition, it had to be taken into account that the qualitative categories used for data analysis had to be elaborated (had to become more differentiated) in the course of time. In order to ensure the longitudinal comparability, the earlier codings had to be brought into line with the respective elaborations of the coding scheme. Overall, a constant standard of excellence of qualitative analysis could be achieved. Moreover, the coding scheme could be improved by considering the quantitative results.

In this article, these procedures and their efficiency in evaluating the promotion program are presented.

Table of Contents

1. Aim

For the continuous evaluation of our research project in primary schools, which we will briefly describe in Section 2.1, it was necessary to find a meaningful connection between qualitative and quantitative methods. As we will depict in Sections 2.2 and 2.3, qualitative assessment methods are better suited for primary school children. However, our evaluation results should represent a deciding factor for implementing our program in primary schools broadly. Therefore, the qualitative assessments had to be executed at a larger scale, which is normally only achieved by quantitative projects. As a consequence, it was desirable to cover the process of qualitative analysis by using quantitative reliability and validity checks. In Chapter 3 we will demonstrate the utility of quantitative measures for the development of a category scheme. For the assessment of categorizing reliability, a new statistical formula had to be developed, which can determine the so-called "inter-coder-reliability" with respect to our more complex cases. This procedure will be presented in Chapter 4 and can also be obtained via the Internet (http://www.uni-goettingen.de/~vbenedi [Broken link, FQS, December 2004]). [1]

2. Assessment of Self-Relevant Contents in Children

2.1 The project "Promoting Health by Strengthening Self-Worth"

The research project tested a health promotion program which was run from the first through the fourth school class. Its focus was the promotion of mental health, namely the strengthening of health factors by applying a salutogenetic concept (ANTONOVSKY 1993) and the strengthening of coping abilities by teaching how to manage strain in a competent way. ANTONOVSKY (1993) developed his salutogenetical model and his sense of coherence concept while he was searching for "health maintaining factors that help people deal with threats in their lives as successfully as possible" (p.10). In this model, the psycho-social resources and the subjective coping behavior are considered to be crucial pre-conditions for an individual to move towards the healthy pole of the health-disease continuum. If "generalized resistance resources" exist and if they can be utilized in concrete coping behaviors in a stressful situation, a sense of coherence is emerging. The "origins of health" are to be acquired at best in childhood. Moreover, children should be prepared for the risks of life as early as possible. Therefore, we started implementing a promotion program in the first grade. As we don't know—in considering both the individual developmental perspective and the social-cultural and ecological development—which stressors today's six-years-olds will have to face when they will be thirty, the resources to be developed should represent general, cross-situational fundamentals. We consider positive self-worth as one of these foundations and important resistance resources. Its development, maintenance and promotion has a key function in health promotion. [2]

Our promotion program, that was designed on the basis of a salutogenetic concept and that was tested from the first through the fourth grade, was expected to maintain and to promote this resistance resource—positive self-worth. The first school year is especially suitable, because at this age most children show positive self-worth, which is fading in the course of the primary school years for a considerable number of children (cf. KRAUSE 1998). [3]

The first trial was carried out in a total of 20 school classes in the cities of Göttingen and Greifswald. It fulfilled the major aims that were formulated at the beginning of 1995. In each of the schools involved a so-called "health team" carried out the health lessons. The respective contents were assigned to five main foci:

- Strengthening self-worth by self-reflection,
- body experience and body consciousness,
- health promoting interaction and communication,
- leisure time behavior and health, and
- healthy diets. [4]

Each health day (main focus) was worked out by the project team and was discussed with the female teachers before implementation. Furthermore, a joint evaluation was carried out after every health day. Thus, for each class an evaluation protocol afterwards was written. [5]

In order to check the effectivity and to evaluate the program, extensive qualitative interviews of primary school children were done. The working hypothesis stated that normally school beginners would feel well, that is, they would show positive self-worth, self-confidence and a high evaluation of one's own competence (rather over-estimation than under-estimation of self-relevant aspects). [6]

A part of the investigation was the assessment of subjective well-being of young children. The problem was to find a suitable measurement instrument to obtain statements of five- to ten-year-old children about their feelings. [7]

2.2 Problems in the analysis of self-reports

Self-research is associated with special methodological problems. The reason why is that subject and object of inquiry are identical. Twenty years ago, MUMMENDEY (1979) described methodological problems that are unsolved yet. Essential questions address a) test criteria (reliability, validity, objectivity), b) the "fit" between theoretical concept (self-concept, self-worth) and assessment method, c) the indication of an assessment method, d) the specificity/generality of assessed characteristics, e) the developmental appropriateness of the assessment method (consideration of age levels) and f) the evaluation of the measurement of change in longitudinal designs. Especially the last two problem fields were of special importance for our project, which aimed at the strengthening of self-worth in primary school children. [8]

Our design was concerned with the longitudinal investigation of children from the first through the fourth class. At the end of the five health days, at the end of every school year, our children were interviewed about how they would feel. When putting together the measurement instruments, it was difficult to find a method

- that would assess the development and change of self-worth,
- that would be suitable for children at the age between five and ten, and
- that could be used repeatedly over a period of several years. [9]

A review of traditional assessment procedures for self-concept, or self-worth, showed that they were developed for adolescents or adults. Adjective check lists, Q-sorts, semantic differentials, rating scales (e.g., evaluation of self-referential statements on a scale) or personality questionnaires obviously do not make sense for school-children (or kindergarten children). The impediments of these procedures—excluding age-related matters—has already been described by MUMMENDEY (1979) very convincingly (see also HAUSSER 1995). We decided to do without these traditional methods giving developmental reasons: Their ecological validity for the assessment of mental states in primary school age is very low. It was therefore necessary to use a methodology which is suitable for our population. Most importantly, the first run took place in the context of the pre-school examination, that is, the children could not read yet. [10]

The interpretation of changes in self-worth is only to be assessed by childrens' self-reflections. But this is especially problematic for primary school children: For example, the questions arises, if an observed change over several points in time represents in fact a change in self-worth, or if the change in self-report reveals a development in cognitive skills. School children gain more and more knowledge about their own cognitive processes and about controlling them. This "confounding" can be resolved (although not completely controlled), if the assessment method is able to say something about the increasing cognitive differentiation. [11]

How can the cognitive developmental level of the primary school child be briefly characterized? According to PIAGET, the child is located at the stage of concrete-operational structures. The thinking of the child is highly dependent on the given information, being represented in a concrete-vivid (e.g., pictorial) or verbal mode (MONTADA 1995, p.540). In order to attain information about the subjective perspective on health and illness, the picture and the play should be preferentially used. [12]

Investigations into memory development (SCHNEIDER & BÜTTNER 1995) show that primary school children have already an autobiographical memory which keeps reminiscences of complex and strongly self-referential experiences (FIVUSCH 1993, HOWE & COURAGE 1993, 1997; LEICHTMAN 1999, NELSON 1993, 1997). These episodic long-term-memory contents are associated with the semantic long-term memory, in which conceptual knowledge is stored, for example, language, rules, and terminology. Children acquire a meta-linguistic awareness for verbal categories and regularities at the age between five and eight (GRIMM 1995; KARMILOFF-SMITH 1985, 1992). These competencies have an effect on the differentiation of autobiographical memories with increasing age. Verbal data can be very instructive to learn something about subjective mental states in primary school age. [13]

We decided to use a combined procedure:

- An oral interview, called the "health profile" (developed with the aid of a Danish cooperation partner);
- A pictorial-based method "What I like to do" developed by Krause, which assesses subjective preferences of every-day activities depicted in drawings (cf. KRAUSE 1998)
- A sentence completion test—a well-known and often used method to assess self-relevant contents. With respect to the aim of the project and the age of the children, eight sentence beginnings were chosen:

1. "If I can't do something ..."	5. "I am sad ..."
2. "I don't like ..."	6. "I am angry ..."
3. "The other children ..."	7. "I am delighted about ..."
4. "I am afraid ..."	8. "In the school ..." [14]

2.3 The sentence completion test—Characteristics of the qualitative method

The sentence completion test is a semi-structured open assessment method. The child shall freely answer and shall decide on the contents of his or her statement. HAUSSER (1982) speaks of "verbalization opportunities" of the respondent. The beginnings of the sentence draw the child's attention to experience-based memories. As expected, those contents are frequently mentioned that are especially easy to get and that are salient in the interview situation. Therefore, the sentence completion test is an individual-centered method: The individual's perspective is central. This orientation is a necessary pre-condition for self-concept research (cf. WIECHARDT 1977 and HAUSSER 1995). [15]

The semi-structuredness of this method ensures an inter- and intraindividual comparability of answers. This is especially useful for young children: On the one hand, they can freely answer, on the other hand, they are inspired by the sentence beginnings to give self-reports. The verbal data are both open to content analysis (by applying a content category system, see below) and linguistic analysis (in the sense of evaluating the verbal complexity/variability). For example, it can be assumed that self-reports are getting more and more differentiated with increasing age. The degree of differentiatedness within a sample is instructive for the cognitive development. [16]

In the sentence completion test children are encouraged to activate actual episodic memories with reference to self that are instructive for the current subjective mental state. The selection of the sentences was difficult. After several test periods we decided to use the eight sentences shown above. The children were interviewed in the school setting. The female interviewers were familiar persons who accompanied them over one or more school years in the health lessons. Nevertheless, it could not be avoided that the relationship towards the interviewer and the context of the situation would influence the responses of the children essentially. This situational influence, which is always given in interviews, could be leveled off by the fact that the interview was repeated every year and that the children worked on the test five times, overall. [17]

3. The Connection between Qualitative and Quantitative Research

The often discussed problem, whether qualitative or quantitative research is better, is in our opinion an artificial one. It is crucial, in how far an assessment method is appropriate for the subject under study and, at the same time, is able to solve the methodological problems mentioned above (KRIPPENDORF 1980 introduced eight content analytical test criteria in this context). One possibility of increasing interindividual comparability is the formalization of data analysis by using mathematical methods. A low degree of formalization of data analysis is protecting (to a certain extent) the nature of qualitative data. "Also in qualitative oriented studies in human sciences the pre-conditions for meaningful quantifications for the backup and generalizability of results—via qualitative analysis—can be created." (MAYRING 1993, p.24) [18]

The object under study in our case was the assessment of childrens' subjective mental state and the change of self-worth in the course of the primary school years. In a western and in an eastern German town (the middle-sized Göttingen and Greifswald) total collections were carried out in several schools, which represented within each town different areas of the city. This design of "cluster"-sampling is not representative, but is often used with respect to school populations in order to come to generalizable conclusions. [19]

In order to assess the subjectivity of feelings in a mere approximate way and to recapitulate the individual development, qualitative methods are imperative. But if —at the same time—there is interest in the question whether a special intervention (the promotion program in our case) makes sense, generalizable conclusions are necessary. That is why we carried out the evaluation on a wider scale, using quantitative, formalized methods for analysis. For the applied sciences as pedagogy is, this proceeding could be a way of overcoming the often practiced dichotomizing of qualitative and quantitative approaches and of achieving an integration. "Quantity per se is senseless, quality per se remains without consequences." (HUBER 1989). [20]

For the investigation of self-reports a combined methodological proceeding is suitable (HAUSSER 1995). On the one hand, the meaning of qualitative data can be inferred, on the other hand, interpretation processes can be systematized and documented, and results can be put in order and can be quantified (cf. HUBER 1989). [21]

By use of the sentence completion test, more than thousand statements per sentence were produced ultimately. Considering the fact that in the course of the project different co-workers joined in the evaluation, the problem of evaluation reliability cannot be solved by using the usual method of "discussion until correspondence" on controversial texts.[1] The analysis of our data should be backed up by reliability coefficients that can be compared to similar measures of standardized assessment instruments. [22]

We decided to use qualitative content analysis as our evaluation method (MAYRING 1993). The most important premise was that category formation could be reflected and controlled by dealing with the empirical material inherently. For the attainment of that goal it was important that—in case of changed conditions—renewed reliability and validity determinations could be included and could be compared to former results, as we dealt with a longitudinal project lasting several years. Quantitative measures are suitable for this problem, too. [23]

Imagining the categorizations to be managed and the necessity of controlling the quality of the coding scheme and the codings—even in case of schema changes—it was our task to assess the coding performance in the course of the whole project quantitatively. To attain this, quantitative measures were repeatedly assessed while developing and applying the category scheme. Those measures were included into the ongoing process of development. [24]

In the following this method of connecting quantity and quality with respect to the development of category schema and of checking coding performance will be presented. The category development for the first sentence "If I can't do something ..." will be used as an example for our proceeding in dealing with the self-reports (cf. KRAUSE & MÜLLER-BENEDICT 1997). [25]

The development of the coding guideline took place in two steps: In a first run, using material from a small sample, simple concordance statistics and cross tables were determined for the respective two coders, and the degree of intercoder reliability between all coders involved and the total concordance was computed. On this basis, an improved version of the coding manual could be accomplished. Particularly, those cases were identified that showed large coding deviations. This proceeding will be described in Sections 3.1 and 3.2. [26]

Having accomplished an improved version of the coding guideline, the reliability of this new version was tested in a second run using a larger sample. A group of coders categorized the former codings a second time using the new version of the guideline. Between the codings with the first and second version at least three months passed. The reliability of this second run and the improvement (compared to the first) were validated using standardized statistics of intercoder reliability. The method of assessing intercoder reliability will be depicted in Section 3.2. [27]

[1] As for example in the projects by HOPF; RIEKER; & SANDEN-MARTENS 1995 and HEITMEYER, BUHSE, LIEBE-FREUND, MÖLLER, MÜLLER, RITZ, SILLER & VOSSEN 1992.

3.1 The development of the coding scheme

On the basis of our theoretical reasoning we determined categories that were relatively global and that served as a preliminary orientation. However, it should meet the requirement of a unitary classification principle (HOLSTI 1969, MERTEN 1983).

"If I can't do something ..."

Category 1: With help to success

Category 2: Without help to success

Category 3: Accepting failure

Category 4: Failure with self-assessment

Category 5: Concrete description of the situation

Category 6: Consequences of non-achieving

Category 7: Other

Category 8: No answer

Figure 1: Preliminary coding scheme [28]

First, 72 children's completions of eight sentences (representing one assessment unit of a school in Göttingen) were shared out in such a way that two co-workers coded one sentence, respectively, and set up empirically based categories. In several meetings of the research group the propositions were discussed. Then 12 coders worked with the following coding scheme.

"If I can't do something ..."

Category 1: With help to success

Category 2: Without help to success

2a. Immediate repeating, carrying on, making an effort

2b. Delaying

Category 3: Accepting failure

3a. Ignoring

3b. Withdrawal, avoidance

3c. Distraction by doing something else

Category 4: Failure with self-assessment

4a. Emotional

4b. Cognitive

Category 5: Concrete description of the situation

Category 6: Other

Category 7: No answer

Figure 2: Revised category scheme [29]

3.2 Measure of correspondence and development of a coding guideline

As will be shown in the further discussion, the development of a category scheme which could provide a maximum in content, and at the same time meet the requirements for reliability, validity and objectivity, was difficult. In the analysis of the first run of the material the following values were calculated: [30]

1. Modal values

The results of the coding were analyzed to determine whether the statements of the children resulted in clear modal values that were sufficiently distinguishable to generally speak of a correspondence between the coders. [31]

2. Frequency distributions

For certain answers where such modal values did not occur, because a clear categorization was not possible, we had to analyze the distribution of the categorizations into the category scheme. It had to be clarified, why different categorizations occurred, for example, if the coders had different understandings of a certain statement, if there were polarities between two or three categories or sub-categories, respectively. This would indicate a poor demarcation of the category scheme and thus the requirement for exclusivity and mutual demarcation of the categories would not be met sufficiently. [32]

The application of the following matrix can help to find an answer to these questions. Cross tables with paired coding (compare Section 3.3) is a comparative method, which in the sense of the questions considered, can be used to analyze all and not just the very strongly scattered statements of children. For the beginning of the sentence "If I can not do something ...", among the 72 completed sentences of the sample modal values of the categories could not be determined for 12 of them. The distribution was as follows:[2]

[2] In the case of the 12 problematic completed sentences some coders were not able to make a decision and left it open, thus n (sum in one row) does not always add up to 13.

Child No.	**Category**											
	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					

Table 1: Matrix of the distribution of twelve coded statements for which no clear categorization was found. [33]

The variation concerning the 12 statements was mainly due to the poor demarcation of some categories. Besides this, it became obvious that a longer training time was necessary to acquire sound skills as a coder. [34]

The polarity between the sub-categories 3a ("ignoring") and 3b ("retreat/avoidance) is clearly noticeable as can be seen in the statements by the test persons No:

> 300: *"... then I can't do it."*
>
> *321: "... then I will leave it, or I ask others for help."*
>
> *330: "... then I will leave it."*
>
> *358: "... then I can not* do this." [35]

Obviously the polarity results from the difficulty to decide, whether for example sentence 300 "... then I can't do it" is a statement which signals retreat or avoidance, or whether it can be understood as ignoring a situation of failure. This example shows clearly that the requirement for exclusivity and mutual demarcation of the sub-categories against each other was not met. After discussing this problem in the group, it was agreed that the two sub-categories 3a and 3b can be combined, as ignoring and avoidance or retreat in any case can be considered as a behavior showing that a child does not attempt to solve the problem, nor gives it any further thought. Thus, an impairment of the child's sense of self-worth is not assumed. The demarcation accuracy against the other categories is still maintained. The new sub-category combines all possible variations, which were mentioned under 3a and 3b.

<table>
<tr><th colspan="2">If I can't do something ...
Category 3: "accepting failure"</th></tr>
<tr><td>*1st Version:*
Definition: This category describes statements which show that failure is not an impetus and has no obvious valence for the child. Within this category three different reactions in situations of failure were observed which resulted in the following three sub-categories.
3a: Ignoring
Model sentence: "... then I will leave it, most of it I can do anyway."
3b: Retreat/avoidance
Model sentence: "... then I will do not do this"
3c: Distraction by doing something else
Model sentence: "... then I will play a little game"</td><td>*2nd Version:*
Definition: Statements, which show that failure (to be not capable of doing something) has no obvious valence for the child, are allocated to this category. Two different responses in situations of failure were observed.
3a: Ignoring, retreat, avoidance
Model sentence: "... then I will leave it (I don't mind) most of it I can do anyway", " ... then I don't do this."
3b: Distraction by doing something else/going somewhere else
Model sentence: "... then I will play a little game", ... then I will go home".</td></tr>
</table>

Figure 3: Section of the coding guideline for sentence 1 before and after the last revision [36]

Another problem due to poor demarcation occurred in four statements of category 1 ("with help to success)

301: *"... to say something"*

321: *"... then I will leave it, or ask others for help"*

324: *"... then I will try again and then ask my dad"*

353: *"... then I will go to my girl friend"* [37]

Each of the statements 321 and 324 implies two meanings. Both contain two possible considerations of the child how to handle the situation of failure. As the statements have two different contents, the allocation to two categories is not due to poor demarcation of the coding guideline. [38]

Statement 301 obviously is so ambivalent that a clear categorization is not possible. As a result most coders allocated it to the provisional category 6 (others). Three of the coders however allocated this statement to category 1 which shows their understanding of the completion *"... say something"* of the sentence *"If I can not do something ..."* as an attempt to ask for help. Obviously, other interpretations are also possible. As a consequence, in the meeting after the first run of the material, the problem of interpretation and making sense or making assumptions was intensively discussed. [39]

Sentence 353 was not clearly categorized despite its clear formulation. Here the polarity occurs between category 1 (with help to success) and 3c (distraction by other activities). The statement *"... then I go to a girlfriend"* was allocated five times to category 1 and seven times to category 3c. Testing the inter-coder reliability the effect of the range of different interpretation patterns of the coders is apparent. The statement of a child "... then I will go to a girl friend" is actually a clearly formulated sentence which despite of its definite statement can not be interpreted only in one way. Surely, a child can ask a friend for help. But it is also possible that a child goes to see a friend to find distraction from a task he/she cannot cope with. These, however, are interpretations which lead to a subjective decision regarding the selection of a category. Thus, a more unambiguous formulation of the category had to be found. In this case the sub-category 3c was additionally defined as "to go to a different place". This change seems also justified because the coders chose category 3c (in the final draft this is 3b) more often than 1. [40]

After this revision a coding guideline was compiled which also provided definitions and model sentences (compare example given in Figure 3). [41]

After this, the texts were coded again, this time by a coding group consisting of four coders who had been taking part in the project work from the beginning (as part of a research practical and then as student or research assistants). The results of the first coding by this coding group and the results of the second coding were tested for reliability. This is described in more detail in the subsequent section. As a third step the coders discussed each statement with reference to the individual coding results and determined a final categorization. The procedure was thus, that in order to avoid an influence on the coding process, per meeting only one sentence was discussed. There is the risk that, if a coder is provided with 8 completed sentences of the same child, he/she is provided with sufficient information to have a general conception of this child. As a result of this work all statements were allocated to the given categories of the coding guideline. [42]

Only for specific analyses were all statements of any one child (8 completed sentences at all measuring times) included. Such an analysis for example showed, that some categories within the 1st sentence ("If I can not do something ..."), the 3rd sentence ("The other children ...") and the 8th sentence ("At school ...") were particularly useful to identify those children whose sense of self-worth was impaired. The completion of the sentence "If I can not do something ..." by pupil No 162 had to be categorized several times under 4a "failure with self-assessment (emotional). The pupil continued the start of the sentence as follows:

> "If I can not do something ...
>
> ... then mum is angry with me". (Kindergarten)
>
> ... then I don't like it and feel strange". (1st grade)
>
> ... it depends on what, for example homework: First of all I don't feel very good". (2nd grade)
>
> ... then I don't feel good". (3rd grade)
>
> ... then I actually don't finish it". (4th grade) [43]

3.3 Testing of the category scheme with crosstabulations of coding

The calculation of nominal numbers for the quality of the coding was based on cross-tables of the paired codings. They were compiled using a program specifically developed for the calculation of inter-coder reliability (see following section). Each cell of the table is localized by a combination of the category used by both coders. For each statement a dot is placed in the cell which represents the category selected by both coders for this statement. From this table it is possible to draw conclusions about the category scheme. For example: 52 texts categorized by the coders with the abbreviations "k" and "ve" (category "31" equals "3a", "23" = "2c" in Table 3).

Cat.	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

Table 2: Crosstabulation of the paired coding k, ve
Correspondences: 37 --> Note: Along the main diagonal also non-corresponding categories are found [44]

Except for the last cell at the bottom right (not including the sum rows), the main diagonal shows the categories where the selection corresponded. The row and column forming this last cell with the categories which do not correspond shows, that two categories were selected by one coder only: Coder "k" selected "50" twice, but "ve" not at all. However, "ve" selected "23" once while "k" did not. The strong accumulation along the main diagonal shows a good correspondence. It is also obvious at a glance that category "10" accounts for 50% of all answers. Whether this is justified based on the expressiveness of the category still has to be decided on the basis of the content. With a total number of categories of about 10 it is worth considering a further differentiation of category "10". [45]

The crosstabulations enable us to diagnose relatively easily further shortcomings of the category system, if the results are tested in several tables.

- Categories which are only rarely used or not at all (here 23) should be tested as to whether they are theoretically necessary.
- If non-correspondence occurs frequently, where always the same two different categories are selected (here 31 and 32), the demarcation of these categories should be increased.
- Categories which were used in combination with almost all others (here 33) indicate, that the quality of this category could be immanent in almost all texts and thus is not sufficiently "mutually exclusive". [46]

A combination of the matrix for the coding and the crosstabulations of paired coding described above, after the experience gathered in the project, raises a number of points for a fruitful discussion about the qualitative improvement of the tested category scheme and its coding guideline. [47]

4. Measurement of the Coding Performance

"Inter-coder reliability" describes the "degree of correspondence between the coders. For this in general a so-called "kappa" coefficient is calculated (KRIPPENDORF 1970, COHEN 1960). However, in this project several specific problems occurred, when the "kappa" was determined. These are also very likely to occur in similar projects carried out in social sciences:

- How should correspondence between two coders in general be measured?
- How should it be measured specifically if not all have coded the same number of texts (here children)?
- How should correspondence be measured, if the coders have not used the same categories?
- How can changes in the category scheme with respect to an improvement/worsening be measured? [48]

There are no techniques available that are tested well enough to be mentioned in text books about content analysis or in the standard software. As a consequence we had to develop a kappa coefficient, which covers the problems described above and also a computer program for its calculation[3] (MÜLLER-BENEDICT 1998). [49]

[3] The program for a detailed description of the "kappa" as described in this work is available in the internet. The address is: http://www.uni-goettingen.de/~vbenedi [Broken link, FQS, December 2004].

4.1 A new definition of Kappa

A kappa coefficient must measure the correspondence between two coders on the basis of a nominal value between 0 and 1. Then 1 is defined as a total correspondence, 0 is defined as the correspondence expected if two coders select the categories by chance. Dependent on what possible codings are acceptable and what is understood as "by chance" a variety of calculations are available. [50]

A clarification about the processing of completed sentences which contained two statements such as "If I can't do something, I will go to my mum or try again" was necessary. This occurred three times in the first investigation and we decided only to include the first statement in the analysis. However, after the second investigation we had to reconsider this, because this type of answer was used more frequently. In consequence both statements were analyzed and for the calculation of the kappa an additional category was established (multiple statement). [51]

Another decision had to be made regarding the probability that a specific category was selected correspondingly "by chance" by both coders. The calculation of the most frequently occurring coefficient, COHEN'S kappa (BOS & TARNAI 1989, pp.183; 203), is based on the assumption that this probability is related to the coders as persons. It is calculated as the product probability of how often this category was selected by each coder.[4] In consequence, a category which is only used by one person and not by the others has a probability of correspondence of 0 and correspondence will certainly never be reached. [52]

In our view this concept is not acceptable for the coding of texts. Even if a category is only used by one person, this shows that indeed the text implies a meaning contained in this category and thus correspondence with a small but positive probability could have occurred. In accordance with this argument, we related the probability of random correspondence to a different source other than the person. It is assumed that the text itself and not the individual inclinations of the coders is the object of analysis, and according to SCOTT (1957) must be calculated as a product of the mean of how frequently the coders have used this category.[5] This seems to be a concept which could be generalized for the coding of any text in social sciences, in content analysis as opposed to the coding of action sequences (in observations), client statements (in Psychology) and real-time interviews (without transcriptions during the interviews). Especially in situations like these, it must be taken into account that a coder might not use a certain category, because it is not visible or audible. Another reason why a category is not used by a coder is, that she/he suppresses it due to a psychological constellation. These restrictions concerning the possibilities of

[4] For example: If the coders A and B have coded 100 texts, and if A applied category i 20 times, B applied it 30 times, then the probability p_i that it was applied randomly is p_i = (20/100)×(30/100) = 3/50.

[5] Then (see annotation 3): p_i = ((20 + 30)/(100+100)) × ((20+30)/(100+100)) = 1/16 (= 3/48 compared to p_i = 3/50 in Annotation 4).

perception cannot occur in the codings of texts. In content analysis all categories are open to each coder, which makes the probability of selecting a certain category solely dependent on the text.[6] [53]

4.2 Kappa calculations of the coding performance

On the basis of these assumptions the calculation method for the determination of the degree of correspondence between two coders, in accordance with SCOTT expressed as kappa, has been outlined. At the same time it determines the procedure, if two coders have not used the same categories. Those categories which were used by only one coder are included with a small but positive probability in the calculation of random correspondence.[7] [54]

Finally, the method of measuring correspondence between several coders needs to be determined. The nominal number for measuring would have to remain constant even if an additional coder with a comparable coding performance joined the coding group. Then it is possible to determine, whether the coding performance of the group increases as a result of replacing one coder, or whether for example the total performance is poorer as a result of additional coders. Thus it follows, that the kappa for a number of coders should be an "average" of all paired codings. The calculation of this average must be thus, that "0" represents the expected random correspondence according to SCOTT. [55]

Using this coefficient the nominal number of correspondence for most codings concerning content analysis can be determined and compared beyond the texts. In general, values above 0.7 are considered as acceptable or even good (BAKEMAN & GOTTMAN 1986, p.82), as for example the reliability also in standardized interviews of re-tests on average is about 70% (KÖNIG 1973, p.175). [56]

In addition to the measuring and testing of the coding quality also the improvements of the category scheme can be measured. The changes of the category scheme in its core can be considered as a continuation of the qualitative analysis of empirical results. For this reason they continue over a considerable time of the research project. This has the advantage, that we can code the same texts again, which were coded on the basis of the first version of the coding guideline at the beginning of the project, with the final version, and determine the degree of correspondence between the coders. In the course of our project it became necessary to extend the coding guideline, because the contents of the children's statements diversified (for example statements where school and learning were implied, did not occur until later) and became more complex. A substantial increase of the kappa coefficient thus indicates an improvement of the

[6] HUBERT (1977, p. 295) considers this case as "Levene's model" and notes: "Levenes notion may be generally more popular in the social sciences than either of the two matching concepts presented earlier." (see also KRIPPENDORF 1970)

[7] If for example category i is not applied, then (see Annotation 4): p_i = ((20 + 0)/(100 + 100))2 = 1/100.

category scheme, unless it is solely interpreted as an effect of "training" and "learning" of the coders. [57]

After the first run of the coding procedure, in this project, for each of the eight sentences that had to be coded, the kappa coefficient was determined for all pairs and all coders together. Values of 0.6 to 0.8 were obtained, which is a satisfactory result. After a further modification of the category scheme all sentences were coded again according to the final version and the new guidelines. This was carried out in separate meetings. Again, after this run the kappa values were calculated. For all sentences the values for the pairs as well as the average total amount, showed a more than 10% increase of kappa, which means that the improvement of the category scheme was satisfactory. The correspondence which was achieved here, is comparable to results obtained with standardized interviews. To show this in more detail for example for the first two sentences the following values were calculated:[8]

1st sentence											
Pair	s,k	s,ve	s,ch	s,v	k,ve	k,ch	k,v	ve,ch	ve,v	ch,v	alle
C1	.6431	.6605	.7613	.7883	.5928	.6391	.6692	.7872	.6362	.7154	.6913
C2	.7901	.7869	.8115	.8327	.7824	.8550	.7570	.8780	.8025	.8032	.8103
2nd sentence											
Pair	s,k	s,ve	s,ch	s,v	k,ve	k,ch	k,v	ve,ch	ve,v	ch,v	alle
C1	.8070	.7900	.8482	.7800	.8090	.8007	.8243	.7833	.7273	.7154	.7881
C2	.8569	.8736	.8733	.8413	.8855	.9045	.8564	.8862	.8732	.8888	.8683

Table 3: Inter-coder-reliability of the first and second coding of the first and second sentence for all coding pairs and in total. [58]

One of the achievements for the project resulting from this extensive testing of the coding performance was, that it could be shown, that measuring of the coding quality and its stability as well as comparability with other reliability values, is possible. Besides this, the testing generated information about shortcomings in the category system, which go beyond content related critiques, and which only become obvious with quantification. Thus the test provided a tool to measure the progressive changes of the category scheme. [59]

5. Summary

The combination of qualitative and quantitative methods has proved to be sensible, especially with regard to problems arising in the analysis of large amounts of qualitative material, such as in the evaluation of the program for health promotion at primary schools. The application of qualitative methods for the investigation of a large number of cases puts high demands on the flexibility

[8] s, k, ve, ch, v are the abbreviations for the coders.

and openness of the category system. Also problems in its practical implementation arise, if the quality standards of coding are to be met. For this reason it is necessary to develop standardizing methods, which guarantee a continuously regular research performance. Both, the heuristic application of quantitative analysis allowing to detect inaccuracies and loop holes in the category scheme as well as the possibility to test reliability expressed in a nominal number, were valuable tools for the further development of the qualitative methodology of the project. In conclusion, the complementary application of qualitative and quantitative research was very useful for our specific project. By the quantitative division of the exploratory material an improved reliability of the qualitative investigation was achieved. [60]

References

Antonovsky, Aaron (1993). Gesundheitsforschung versus Krankheitsforschung. In Alexa Franke & Michael Broda (Eds.), *Psychosomatische Gesundheit: Versuch einer Abkehr vom Pathogenese-Konzept* (pp.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 (Eds.), *Angewandte Inhaltsanalyse in Empirischer Pädagogik und Psychologie* (pp.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 (Eds.), *Child victims, child witnesses: Understanding and improving testimony* (pp.1-24). London: Guilford Press.

Grimm, Hannelore (1995). Sprachentwicklung – allgemeintheoretisch und differentiell betrachtet. In Rolf Oerter & Leo Montada (Eds.), *Entwicklungspsychologie* (pp.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* (pp.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 (Eds.), *Angewandte Inhaltsanalyse in Empirischer Pädagogik und Psychologie* (pp.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 (Ed.) (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 (Ed.). *Sociological Methodology* (pp.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. (Eds.), *Child psychology: A handbook of contemporary issues* (pp.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 (Eds.), *Entwicklungspsychologie* (pp.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 (Ed.), *Selbstkonzept-Forschung: Probleme, Befunde, Perspektiven* (pp.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 (Eds.), *The self across psychology: Self-recognition, self-awareness, and the self concept. Annals of the New York Academy of Sciences, Vol. 818* (pp.103-116). New York, NY: New York Academy of Sciences.

Schneider, Wolfgang & Büttner, Gerhard (1995). Entwicklung des Gedächtnisses. In Rolf Oerter & Leo Montada (Eds.), *Entwicklungspsychologie* (pp.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.

Authors

Dr. *Christina KRAUSE*, Dipl.-Päd., Professor of Pedagogic Psychology at the Pedagogic Seminar of the Georg-August-University Göttingen, Main Focus "Diagnosis and Councelling"

Main focus in research: Development of self in childhood and adolescence, health promotion in the school, life and occupational orientation of adolescents. With respect to the latter topic, a cooperation project with the University of Monterrey (Mexico) lasting four years is going on.

Contact:

Christina Krause

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

Phone: +49 / 0551 / 399 455

E-mail: ckrause@gwdg.de or Dr.ChristinaKrause@t-online.de

Dr. disc. pol. *Volker MÜLLER-BENEDICT*, Dipl.-Math., Privatdozent, assistant at the Sociological Seminar of the University of Göttingen

Research topics: Education research, quantitative methodology, formal modeling

Contact:

Volker Müller-Benedict

E-mail: vbenedi@uni-goettingen.de

Dr. phil. *Ulrich WIESMANN,* Dipl.-Psych., scientific assistant at the Institute of Medical Psychology of the University of Greifswald

Main foci in research: Salutogenesis, strengthening of self-worth in primary school age, body perception and health, health awareness and health behavior of young adults, multiple sclerosis and work motivation, multiple sclerosis and cognitive adaptation

Contact:

Ulrich Wiesmann

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

Citation

Krause, Christina; Müller-Benedict, Volker & Wiesmann, Ulrich (2000). Small Children—Big Data. Possible Links Between Qualitative and Quantitative Methods in the Analysis of Self-Reports [60 paragraphs]. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research*, *1*(2), Art. 16, http://nbn-resolving.de/urn:nbn:de:0114-fqs0002165.

Revised 7/2008