

Dream content analysis: Basic principles

Michael Schredl

Central Institute of Mental Health, Mannheim, Germany

Summary. Dream content analysis is one of the basic methods applied in psychological dream research. The method can be easily applied but several pitfalls must be avoided in order to get valid results. The present article outlines how to construct new scales and the problems related to reliability and validity of content analytic scales. In addition, statistical issues prominent in the analysis of dream content data are addressed. Despite the various efforts and published findings in dream research, the systematic study of the advantages and problems of the dream content analysis method is still at the beginning.

Keywords: Dream content analysis, reliability, validity

1. Introduction

Dream content analysis is one of the basic methods applied in psychological dream research (Domhoff, 1996; Hall & Van de Castle, 1966; Schredl, 2008d). This method has the advantage that it satisfies the common criteria of science such as replication by another research group, assessment of reliability and validity, and minimizing experimenter bias. First, the basic assumptions and the procedure of dream content analysis will be outlined in Sections 2 and 3. Section 4 includes a brief description of the construction of different types of dream content analytic scales. Dream manuals are scoring systems which are aimed at the comprehensive and, if possible, economic measurement of dream content aspects (section 5). In the last sections of this paper, measurement quality (reliability and validity) and several methodological issues which play a role in data analysis, e. g. the problem of dream length, will be addressed.

2. Basic assumptions

Before the basic assumptions and the goals of dream content analysis are discussed, a fictitious example should illustrate the applicability of content analysis. A research psychologist is asked by a colleague who works as a therapist whether his impression that depressed patients often dream about rejection can be generalized. In order to test this hypothesis, the researcher develops, in corroboration with the clinician, a dream content scale which measures situations that include some form of rejection. The next step is the data collection of dream reports of depressed patients and dream reports of gender- and age-matched healthy controls. Dream reports are typed out and randomly "shuffled". An independent judge rates all dream reports along the rejection scale without knowing to which group each

dream belongs. We will assume that the scale measures the number of rejections per dream. The statistical analysis can reveal, after coding the variable "depression vs. healthy controls" to each dream report, whether there is a significant difference between these two samples, i. e., whether depressed patients dream more often about rejection than healthy persons.

The above mentioned example stresses the crucial aim of content analysis; that particular aspects of the verbal material (in the example: rejection) have to be quantified in order to carry out statistical analyses (cf. Hall & Van de Castle, 1966). Statistical tests are of importance for differentiating between findings which are due to chance and findings which might reflect "real" differences in the population. It is, however, not possible to differentiate exactly but only to a distinct amount of error probability. The application of dream content analysis is manifold: for group comparisons (see example above), for the analysis of dream series and correlational studies between dream content and psychometric measured waking-life variables, e. g. personality dimensions. The advantages of content analysis in contrast to an intuitive comparison are also evident from the above outlined example. First, the researcher makes an effort to formulate explicit rules about which dream sequences will be scored as rejection. This enables a replication of the study with new data material in order to confirm the findings of the first study. Second, the application of such scales permits estimates of the measurement quality. And, third, the researcher specifies explicitly which dream reports were analyzed, the characteristics of the sample and the method of data analysis (most recent dream, diary dreams, and laboratory dreams).

On the other hand, content analysis has some shortcomings. By application of specific scales, a loss of information takes place since only a few aspects of the dream report can be measured. The uniqueness of dreams of a particular person can not be captured. An additional problem which will be discussed in section 6 is the validity of the dream content scales. As shown below, it is not a trivial question how valid dream emotions which were experienced by the dreamer can be measured by analyzing the dream report by an independent judge.

Corresponding address:

Prof. Dr. Michael Schredl, Sleep laboratory, Central Institute of Mental Health, PO Box 122120, 68072 Mannheim, Germany.
Email: Michael.Schredl@zi-mannheim.de

Submitted for publication: December 2009

Accepted for publication: January 2010

3. Procedure

In Table 1, the steps of carrying out a dream content analytic study are listed (cf. the example in section 2). The advantage of formulating a specific hypothesis, e.g., increased rejection rate in dreams of patients with depression, is the reduction of the expenditure and the increase of statistical power because one needs not to correct for multiple testing. If the researcher, on the other hand, applies a dream manual with over 300 different subscales (e.g., Hall & Van de Castle, 1966), then several significant findings at a $p = .05$ would be expected by chance; an application of an alpha-adjusting procedure inevitably reduces statistical power for detecting differences. The next step is to select an existing scale (see section 4) or developing a new scale. This depends on the study's hypotheses. For new scales, it is necessary to compute interrater reliability indices in order to document the quality of the scale.

It should be very carefully documented how dream reports were elicited (cf. Winget & Kramer, 1979) since the type of dream report used for the analysis might have a strong effect on the findings. The advantages and disadvantages of different types, like most recent dreams, diary dreams, laboratory dreams or patients' dreams obtained from therapists are, for example, discussed in Schredl (2008d). For example, laboratory dreams are often less emotional than diary dreams or most recent dreams and they include very often laboratory references (Schredl, 2008b). Each method has pros and cons and it again depends on the goals of the study which dream collection method is best suited.

After collecting the dream reports, they are usually typed in order to facilitate blind rating. All information not reflecting the dream experience like "I dreamed of my uncle whom I have seen the day before" should be removed so the judges are not distracted by irrelevant information. For some research questions, e. g. guessing the dreamer's sex (Schredl, 2008a; Schredl, Schwenger, & Dehe, 2004), it is necessary to edit the dream more thoroughly and remove all explicit references to the dreamer's sex, e.g., "My wife has" changed into "My wife/my husband has....". As stated above, for new scales it is necessary to compute interrater reliability indices to document the measurement properties of the scale. If well-established scales were applied, this is not necessary because one can refer to the indices given by the author or subsequent studies using this particular scale.

Analyzing the data after adding the coding of the dream's origin (depressed patient vs. healthy person) has to be done very carefully to avoid problems like dependent observations (statistical dependency) or multiple testing. For the interpretation of the results, it is necessary to think about the validity of the applied scale. It will be shown that for dream emotions the application of different dream content analytic scales, measuring emotions, can yield very different results, e.g., for the ratio of positive and negative emotions: a preponderance of negative emotions for the Hall and Van de Castle scales and a balance ratio for global self-rating scales (Schredl & Doll, 1998).

To summarize, dream content analysis is a straight-forward method; easily applied even though it can be very time consuming depending on the number of dream reports and number of scales.

Table 1. How to do a dream content analysis

Procedure
• Formulating a hypothesis
• Selecting an existing scale or develop a new one
• Eliciting dream reports
• Blind rating by external judges
• Computing interrater reliability
• Statistics
• Interpreting the findings

4. Scales for dream content analysis

In their book "Dimensions of dreams", Winget and Kramer (1979) have put together 132 scales and rating systems for dream content analysis that have been published in English. In view of the abundance, the question arises whether it is still necessary to develop new scales or will it be sufficient to apply already existing scales, if necessary, in their appropriate translation. Clark, Trinder, Kramer, Roth, & Day (1972), for example, have shown that the masochism scale developed by Beck and Hurvich (1959) can be derived by summing up a few categories of the Hall and Van de Castle (1966) rating system. On the other hand, Hauri (1975), who carried out a factor analysis of 23 scales of different authors for 100 dreams pointed out that scales, which measures similar characteristics, e. g. dreamlike quality, only correlate to about $r = 0.5$ with each other. This means that 75 % of the variance is unexplained and one might assume that each scale measures specific aspects.

A simple recipe how to construct a valid dream content scale seems not to exist. Hall and Van de Castle (1966) have suggested, first, the need to read many dream reports in order to obtain an insight into the variety of dream experiences. However, one has to be careful not to include the material which should be analyzed in respect to the hypothesis since such a scale will measure the aspects which one read into it. For the next step, categories were formulated as exact and comprehensive as possible. The methodological goal is to obtain a high interrater agreement if two or more judges apply the same scale to the same dream sample. The more explicit the coding rules are formulated the simpler are the decisions for the judge. However, one has to keep in mind that there are always "grey areas" and special cases which can not be grasped by the categories of a scale (cf. (Domhoff, 1996). The following example will demonstrate how such coding rules can look (see Table 2). Whereas this scale measures the number of persons in the dream, like in a waking-life experience or theatre play, Hall and Van de Castle (1966) have chosen a broader definition; they include also persons that are mentioned in a conversation or if objects which belong to a person occur in the dream. This example should illustrate the fact that explicit coding rules are essential to comprehend what has been measured. On the other hand, the arbitrariness of the definition, e. g. in respect to groups (see Table 5.3.1) becomes clear. This freedom in the definition of the categories should lead to particularly careful evaluation of dream content analysis studies in respect of the specific scales' definitions.

After developing the scale, two or more judges rate in-

Table 2. Number of dream persons (Schredl, 1998b)

Scale: Number of dream persons
<ul style="list-style-type: none"> • How many single persons occur in the dream? E. g. father, child, friend etc... or unknown persons • Groups which are mentioned solely as group will be coded as single person, despite description with include the group size, e. g. 8 persons. If members of the group appear in the course of the dream as single persons, they were coded. Similar, a division into subgroups will be counted for each new subgroup. • If a person underwent a metamorphosis, i.e. an identity change takes place, this will be coded as one single person. • Each person who appears in the dream will only be coded once. • If persons, seeking contact to persons or objects which belong to other persons (e. g. the car of my uncle) are only mentioned in the dream, these references were not coded. • The dreamer herself/himself will not be coded.

dependently a number of dream reports in order to compute interrater reliability coefficients (see section 6). In cases where the coefficients are low, the subjective experiences of the judges and a discussion of the disagreements often allows an improvement of the scale. However, guidelines or cut-off values classifying "good" and "poor" reliability do not exist in dream research.

For measuring dream characteristics, three kinds of scales are commonly differentiated (see Table 3). Content analytic scales in the narrow sense are defined as nominal scale, e. g., occurrence of persons (mother, child, friend etc) or aggression (physical, verbal etc...) within the rating system of Hall and Van de Castle (1966). These authors differentiate between empirical and theoretical scales. Empirical scales were derived by reading through many dream reports to look for things which occur often or are of special interest. The above mentioned scale for measuring the number of dream persons is an example for an empirical scale. The fact that this scale is also not free from theory becomes clear if one looks at the equivalent coding of persons who appear in the dream and persons who were only mentioned in the dream. Theoretical scales are scales which are derived from a particular theory and which are tested empirically by applying this scale to a sample of dream reports. Hall and Van de Castle (1966) have derived from the psychoanalytic theory (adding a variety of prepositions) the scales 'penis envy',

Table 3. Kinds of scales used for dream content analysis

Type
<ul style="list-style-type: none"> • Content analytic scales • Global rating scales (applied by Judges) • Global rating scales (self-ratings)

Table 4. Global rating scale (Schredl, 1991)

Scale: Realism/Bizarreness
<p>In four steps, the closeness of the dream action to everyday reality should be estimated. Are the dream events parts of everyday life of a person of a Western culture or are they uncommon or impossible?</p> <p>(1) Possible in waking life and dream events are part of the normal everyday life</p> <p>(2) Many elements of the waking life, but the dream action is uncommon but impossible for the dreamer in real life</p> <p>(3) Occurrence of one or two fantasy objects, bizarre connections or actions that are not possible in the real world. Global rating scales (applied by Judges)</p> <p>(4) Occurrence of several fantasy objects, bizarre connections or actions that are not possible in the real world.</p>

'castration anxiety' and 'castration wish'. If one looks at the coding rules and the given examples of the castration anxiety scale, e. g., "My finger was cut off.", "I wasn't able to get an erection.", "I couldn't get my key in the lock.", it becomes clear that the validity of these scales are doubtful (Domhoff, 1996). Similar external validation studies did not yield the expected results; men tend to have more dreams which suggest penis envy ($N = 25$) than women ($N = 13$ dreams; Domhoff, 1996).

In addition to the content analytic scales (see above), global rating scales were commonly applied. The following scale is an example (Table 4).

Whereas the content analytic scales are dependent from summing up, for example, bizarre elements (e. g. Domhoff, 1996), the global rating scale allows the assessment of the whole dream. Global rating scale offers the opportunity to measure intensity, e.g. the intensity of negative emotions by using a four-point scale (none, mild, moderate, strong) applied by an external judge. The coding system of Hall and Van de Castle (1966) offers only the opportunity to sum up the frequency of mentioned emotions. It seems evident that two occurrences of mild negative emotions do not correspond with the occurrence of one stronger emotion. Similarly, other attempts (e. g. Saul, Sheppard, Selby, Lhamon, & Sachs, 1954) to define intensity levels of hostility (1 = minor discomfort of an object to 6 = death of a person) can yield bizarre statements (if an interval level of the scale is assumed as done by the authors) that 6 minor discomforts equal one death of a person. Since such approaches are unsatisfactory, Gaillard and Phelipeau (1977) have suggested that the most intense occurrence of a category, e. g. negative emotion, should be assessed in a sense of a global rating scale and that this scale has an ordinal measurement level. In that way, dreams can be classified into more or less aggressive or emotional intense dreams. Very simple global rating scale assesses the presence or absence of a particular theme (e. g. health related topics, respiratory-related topics). Such scales and rating scales with a limited number of categories often show sufficient interrater reliability (see section 6).

Global ratings by the dreamer herself/himself were especially used for measuring aspects of the dream experience which might not be fully reflected in the dream report. For

example, one may assume that the dreamer did not explicitly report all emotions which she/he has experienced during the dream, thus, a self-rating scale is more valid than a rating scale applied by an external judge to the dream report (see section 6). Another application of self-rating is the measurement of the temporal reference of dream elements. In a study carried out by Strauch and Meier (1996), the participants were asked whether dream elements occurred within the one's waking thoughts.

To summarize, all three kinds of scales have their pros and cons which, however, have rarely been investigated systematically. For research practice, the question of which type of scale is the most appropriate for the study's rationale can be applied with justifiable expenditure.

5. Dream manuals

The term dream manual is used for collections or systems of content analytic scales which are aimed at the comprehensive measurement of the dream experiences or at least the major dream characteristics. Essentially, two systems have been developed and applied in multiple studies: Hall and Van de Castle (1966) and Schredl (1999a). The most often applied coding system are the Hall and Van de Castle scales (overview: Domhoff, 1996) which, thus, will be presented in more detailed. As further development, the manual by Dippel (1988), the dream manual of Schredl (1999a), will be outlined since it encompasses, in contrast to the Hall and Van de Castle system, global rating scales and is conceptualized within a modular-design principle. In addition to the comprehensive systems, systems for measuring special aspects of the dream, e. g. anxiety and aggression (Gottschalk-Gleser scales; Gottschalk, Gleser, & Springer, 1963), latent structures (Foulkes, 1978), relationship patterns and emotions (Enke, Ohlmeier, & Nast, 1968) or systems for special population groups, e. g. children (Foulkes & Sheperd, 1971) have been designed.

The coding system of Hall and Van de Castle (1966) has been developed over a period of about 20 years and includes the authors' experiences with over 10.000 dream reports. Aside from the three theoretical scales 'penis envy', 'castration anxiety' and 'castration wish', the manual attempts to assess the empirical material in a detailed way. The system includes eight main categories (see Table 5) which are divided into subscales with over 300 different coding options.

If a person occurs in the dream, her/his gender, age, familiarity, identity will be coded. For interactions, it will be coded whether the dream ego is initiating the interaction or if it is recipient or if it is a reciprocal interaction (e. g. fight = mutual aggression). The exact coding rules which are partly quite complex are outlined and illustrated with many examples both in Hall and Van de Castle (1966) and Domhoff (1996).

Schredl (1991, 1999a) has developed the manual of Dippel (1988) which comprised in the last version 206 items (cf. Majer-Trendel, 1991) further with a new concept. The basic idea was a kind of modular-design principle, i.e., the manual includes "important", general applicable scales and a few specific scales which will be developed by the researcher for testing the hypotheses of the study. The idea to extract the most important scales dates back to Hauri, Sawyer and Rechtschaffen (1967) who carried out a factor analysis of 20 ordinal scales and obtained 6 factors: "vivid fantasy" (including realism), "active control", "pleasantness", "verbal aggression", "physical aggression" and "heterosexuality". Based on these findings and own experiences, the manual

Table 5. Coding system of Hall and Van de Castle (1966)

Main categories
Physical surroundings: Settings and Objects indoor, outdoor, uncertain, distorted settings Objects such as architecture, body parts, household, clothing, food
Characters Persons (Number (individual vs. group), sex, identity, age) Animal, creature, mythic creatures
Social interactions Aggressive (8 subclasses, e. g. causing a death, physical attack, verbal threat) Friendly (7 subclasses, e. g. long-term close relationship, offering a gift) Sexual (5 subclasses, e. g. sexual intercourse, petting, sexual overtures)
Activities Expressive communication, thinking, physical activities
Achievement outcomes Success, failure
Environmental press Misfortune, good fortune
Emotions Anger, apprehension, happiness, sadness, confusion
Descriptive elements Modifiers, e. g. size, velocity, age, color Temporal scale, Negative scale (e. g. no, never, not, unsure)
Theoretical scales castration anxiety, castration wish, penis envy, oral incorporations, regression

of Schredl (1991) and the subsequent versions (Schredl, 1998b; Schredl, Schäfer, Hofmann, & Jacob, 1999; Schredl, Schäfer, Weber, & Heuser, 1998; Schredl, Schröder, & Löw, 1996) enclosed the basic scales which are depicted in Table 6.

The advantage of this rating system lies within the clearness and the simple applicability of the basic scales and the requirement to think about every new scale which must be developed for a study measuring aspects which are not assessed by the basic scales. For example, Schredl (1991) developed a three-point scale for measuring the occurrence of problems within in a dream (none, minor and major problems) in order to test the hypothesis whether patients with insomnia tend to have more problematic dreams than healthy controls. Another example was provided by Montasser (1996) who investigated dreams of anorectic and bulimic patients. She constructed scales measuring the occurrence of food themes and food rejection in dreams. The global rating scales allow an assessment of emotional intensity and can be easily applied to short dream reports. The interrater reliability coefficients are mostly sufficient (see section 5.5). The experiences of the various studies applying the system, e. g. dreaming in the elderly (Schredl, et al., 1996), dreaming and eating disorders (Schredl & Montasser, 1999), gender differences in dreams (Schredl, Sahin, & Schäfer, 1998), dreaming in psychiatric inpatients (Schredl & Engelhardt, 2001), patients with sleep disorders (Schredl, 1998b; Schredl, Schäfer, et al., 1998) and the relationship of dreaming and personality, (Schredl, et al., 1999), confirm the usefulness of the scales since the major findings have been replicated and extended as well.

The brief overview illustrates the efforts which have been done in the field of dream content analysis to assess dream

Table 6. Dream manual by Schredl (1998b)

Basic scales
<ul style="list-style-type: none"> • Dream length (three-point) • Realism/bizarreness (four-point) • Number of dream persons • Verbal interaction (presence vs. absence) • Physical interaction (presence vs. absence) • Aggression (4 Subscales, physical vs. verbal, directed to the dreamer vs. directed by the dreamer to others; presence vs. absence) • positive dream emotions (four-point) • negative dream emotions (four-point)

content comprehensively but economically. It has to be mentioned that, on the one hand, the standardization of the scales is desirable but, on the other hand, the results should not depend on the kind of scale (if sufficient scale quality is achieved) used in the study. For example, the comparison of physical aggression in women's dreams and in men's dreams should be independent from the utilization of a content analytic scale or a global rating scale.

6. Quality criteria: Reliability and validity

In this section, the quality criteria of reliability and validity will be discussed. Whereas the procedure to obtain reliability coefficients in dream content analysis is very simple and straightforward, the assessment of the scales' validity is much more complex. Similarly, the analysis of the data which follows the dream collection and the rating process presents some difficulties.

The reliability coefficient is determined as measure of correspondence between different judges, i.e., two or more judges rate the same dreams and – according to the scales' measurement levels – coefficients of exact agreement, Spearman rank correlations or Pearson correlations can be computed. By "exact agreement" the proportion of codings exactly the same between the two judges to all codings is meant. These coefficients indicate how consistent the ratings of two different judges, who rate the same dreams, are for a distinct scale. Low values mean that the subjectivity of the judges plays a major role in the rating process; this can be found especially for scales with few and imprecise

coding rules. High values indicate a sufficient intersubjective comparability. Hall and Van de Castle reported the following reliability coefficients for their rating system, e. g., occurrence of a person (93% exact agreement), and all features of a person (76%). The coefficients for social interaction are markedly smaller, e. g. 54% for aggression, 61% for friendliness and 63% for emotions. These coefficients are derived by comparison of the ratings done by the authors themselves and, thus, indicate that even a comprehensive, explicit formulation of coding rules and extensive practice yield a vagueness in some ratings.

In subsequent studies (e. g. Kramer & Roth, 1979) which applied parts of the Hall and Van de Castle system, the exact agreement for the rating of single aspects of dream characters exceeded 90%. The classification of emotions corresponded in 93% to 98% (Schredl & Doll, 1998). A drawback of the coefficient of exact agreement is the fact that high coefficients will be obtained if the measured characteristic occurs rarely. In that case, the application of Cohen's kappa is recommended (Cohen, 1960); a coefficient which is seldom used in dream research.

For ordinal scales' reliability often Spearman rank correlations are computed. The coefficients for 17 different scales which were studied by Hauri, Sawyer and Rechtschaffen (1967) varied between .59 and .69. These values have been improved by Gaillard and Phelipeau (1977) according to their account by reducing the categories of the scales from seven to five. However, they only reported exact agreements ranging between 88% and 98% if the difference of one point was allowed, but a direct comparison to the Hauri et al. (1967) study was not carried out. In Table 7, Spearman rank correlations stemming from three studies applying the dream manual of Schredl are depicted.

The correlation coefficients commonly varied between .70 and .90. An exception was the correlation of the realism/bizarreness scale in the study of Schredl, Schröder and Löw (1996). Interestingly, the subsequent discussion revealed that one judge who had grown up in another culture (Egypt) rated the realism of dream events in a different way. The interrater agreement of the general nominal scales, e. g. aggression, verbal interaction, was satisfactory with values ranging from 88% to 100% (Schredl, Sahin, et al., 1998). One exception was a scale for rating the ability to cope with a problem within the dream (exact agreement: 58%).

Experience with different coding systems in our research group indicates that a rater training with subsequent discussion of the disagreements is needed to obtain high interrater reliability. The only systematic study on this topic, (Schredl, Burchert, & Grabatin, 2004) using three sets of 100 dreams, clearly showed that training improves interrater reliability in

Table 7. Interrater reliability (Spearman rank correlations)

Scale	Schredl (1991)	Schredl, Schröder & Löw (1996)	Schredl (1998b)
• Realism/bizarreness	r = .745	r = .379	r = .669
• Positive emotions	r = .717	r = .714	r = .824
• Negative emotions	r = .788	r = .765	r = .888
• Number of dream persons	r = .838	r = .937	r = .841

Table 8. Dream emotions (N = 133 dreams; Schredl & Doll, 1998)

Category	Self-rating	Rating by judge	Hall & Van de Castle
No emotions	0.8 %	13.5 %	57.9 %
Balanced emotions	12.0 %	9.0 %	6.8 %
predominantly negative emotions	50.4 %	56.4 %	26.3 %
predominantly positive emotions	36.8 %	21.1 %	9.0 %

several scales, e.g., verbal interaction, aggression. More studies on the effect of rater training on the reliability of the scales are desirable.

As mentioned above, the problem of validity can not be handled as easily as the reliability issue. Whereas reliability designs the exactness of measurement, validity means the extent to which the scale's measurement value is related to the dimension which one would like to measure. Many rating systems (e. g. Hall & Van de Castle, 1966) rely on the so-called, face validity, i.e., one "sees" that the scale is measuring for what it has been constructed. If the coding rules, for example, describe the coding of explicitly mentioned emotions such as guilt, anxiety, fear, embarrassment as "apprehension", it seems obvious that this is given if a dreamer reports that she/he has experienced anxiety during the dream. This coding of explicitly mentioned emotions, however, has shortcomings if applied to the following fictive dream report: "I see a monster and run away as fast as possible." It can be hypothesized that the dream ego experiences fear but did not mention it explicitly in her/his dream report. The emotion score of zero did not reflect the emotion which really occurred. The major validity problem arises because analyzed dream reports are more or less detailed in reflecting the actual dream experience. Astonishingly, very few studies that have compared the dreamers' self-ratings to the ratings of external judges, have been published. For the scales "Participation" ($r = .31$), "Emotional tone" ($r = .65$), "Anxiety" ($r = .31$) and "Aggression" ($r = .31$), Riemann et al. (1985) obtained different correlation coefficients between self-ratings and the ratings made by judges. The total score of aggression (Gottschalk-Gleser rating system) correlated with the global estimate of aggression made by the dreamer to $r = .53$ (Stegie, 1986). Validity coefficients of $r = .53$ or $r = .65$ can be viewed as good since the scale's validity cannot exceed its reliability.

Using the example of measuring dream emotions, some issues which are important for the assessment of validity will be demonstrated in the following. First, the problem of earlier studies (e. g. Hauri, et al., 1967) which used a bipolar scale for measuring dream emotions (strongly negative to strongly positive emotions) has to be addressed since, for example, Gaillard and Philippeau (1977) have shown that 13 % of all dreams included positive and negative emotions. In a sample of diary dreams, the proportion (34 %) was still higher (Schredl & Doll, 1998). In contrast to two separate scales for measuring positive and negative emotions, this bipolar scale will result in an underestimation of the less pronounced emotion. The second problem was the question whether the dream ego experiences the same emotions in specific situations in the dream which she/he also would experience in waking-life (appropriateness of dream emotions). Foulkes, Sullivan, Kerr, & Brown (1988) and Merritt,

Stickgold, Pace-Schott, Williams, & Hobson (1994), however, have shown that such errors due to inappropriateness are present in less than 5 %, i.e., the experience within the dream is very similar to waking-life experience. Analyzing 133 dream reports, Schredl and Doll (1998) have investigated the relationship between three types of scales, the emotion scale of Hall and Van de Castle, the rating scales for external judges constructed by Schredl (1991) and two similar self-rating scales for measuring positive and negative dream emotions. The classification of the dream reports into four groups is depicted in Table 8.

The findings of this study clearly demonstrate an underestimation of dream emotions if the Hall and Van de Castle emotion scales are applied, i.e., if only explicitly mentioned emotions are measured, a good deal of information will be lost. Similarly, the rating scales designed for judges with the similar format of the self-rating scales (four-point: none, mild, moderate, strong) did show a marked underestimation of dream emotions. This underestimation was even stronger for positive emotions than for negative emotions (Schredl & Doll, 1998). Whereas the underestimation revealed a serious drawback of ratings made by external judges, the correlations between external ratings and self-ratings were satisfactory (negative emotions: $r = .669$; positive emotions: $r = .557$) since these values were almost as large as the reliability coefficients of these scales. A second study (Schredl & Erlacher, 2003) found a systematic underestimation of the number of bizarre elements per dream for estimates by external judges compared to the dreamer himself/herself. This clearly indicates that the dream report did not contain the complete information about all different characteristics of the dream. Strauch and Meier (1996) have reported examples in which the experience of touching, e. g., touching a piece of cloth, was not explicitly mentioned by the dreamer in her dream report but was revealed after intense questioning. Similar results were obtained if colors in dreams were analyzed: in 25 % of diary dream reports colors were reported spontaneously (Schredl, 2008c), whereas 100% of diary dreams included colored objects when the dreamer was probed to report all colored objects of the dream (Schredl, Fuchedzhieva, Häming, & Schindel, 2008).

To conclude, whereas researches had handled the problems of reliability sufficiently, the issue of validity has to be investigated in a more detailed way. For some aspects of the dream like emotions, bizarreness, colors, sensory impression, it seems to be necessary to develop self-rating scales in order to obtain valid findings.

7. Data analysis

At the end of this paper, some considerations about data analysis which was handled in different ways by research-

ers will be presented. First, the problem of statistical independence has to be addressed. A basic preposition for most statistical tests (e.g., t-test, Mann-Whitney-U test) is the statistical independence of the observations. If repeated measurements were made (several dream reports per participant), it is commonly assumed that these values are dependent. In the study of Brink and Allen (1992), 139 dreams of 12 patients with eating disorders and 136 dreams of 11 healthy controls were compared. The difference regarding degrees of freedom ($N - 2$) is considerable whether number of participants ($df = 21$) or number of dreams ($df = 175$) are the computational basis and, thus, the significant findings based on the number of dreams which were reported by Brink and Allen (1992) can be questioned. Although Schredl (Schredl, 1998a) demonstrated that the intercorrelations between the scores of distinct scales such as realism, positive and negative emotions are relatively small in a sample of healthy persons ($r < .10$), this problem has to be addressed by choosing an appropriate statistical procedure. A possible solution to this problem which results in the above mentioned reduction in the degrees of freedom is the computation of means per participant. This will be a loss of statistical power. Another solution to the problem is the application of analyses of variance for repeated measures (cf. Heather-Greener et al., 1996). This method can only be applied if participants contribute equal numbers of dream reports, i.e., often this will also result in an information loss. A more sophisticated approach is the use of mixed models (Schredl, 2006; Schredl & Reinhard, 2008) accounting for the with-subject variance.

Another issue which was investigated by Kramer and Roth (1979) is the stability respectively the variability of dream content, i.e., how content varies from dream to dream. It may be possible that one person has very vivid and bizarre dreams one night and the dreams of the following night are bland and realistic. This issue is especially of importance when dream content measures are related to trait aspects of waking-life such as personality dimensions. The study of Kramer and Roth (1979) has shown that the sum scores (number of dream persons, physical activities, verbal activities) based on three to five REM dreams correlated on average to $r = .40$ between subsequent nights. The data of 14 subjects who were awakened out of REM sleep over 20 consecutive nights were included in their analyses. The mean correlation coefficient of $r = .40$ indicate that about 16 % of the total variance is explained by stability but more than 80 % is unexplained. Schredl (1998a) who has analyzed diary dreams of 98 participants (to a maximum of five dreams per participant) has found for dream length ($r = .48$) and number of dream persons ($r = .30$) significant correlations but not for the scales "realism/bizarreness" ($r = .08$), "positive emotions" ($r = .08$), "negative emotions" ($r = .18$) and "verbal interaction" ($r = -.01$), i.e., for single dreams the variability of dream contents is very large. Similarly, low stability coefficients ($r = .07$ to $.37$) were reported by Bernstein and Belicki (1995) who correlated the scores of two dream series of five dreams elicited at an interval of two to three months. Schredl (1998a) has drawn on the analogy to psychometric tests whose reliability can be increased by including more items into the questionnaire (assuming homogeneity of the items, i.e., the items measure the same dimension). Sufficient reliability coefficients ($r > .80$) computed as internal consistency were obtained for all of the above mentioned scales for 20 dreams per participant (Schredl, 1998a). This

kind of reliability is consistent with the notion of measurement stability and should not be confused with the interrater reliability (scales' exactness of measurement). The effect of stability of dream content measures on the results of studies investigating the relationship between dream content and personality dimensions has not been investigated systematically though one might hypothesize that it does play an important role.

The last problem regarding data analysis which will be addressed in this section is the problem of dream length. Many scales, especially sum scores of the content analytic scales, correlate strongly with dream length which is often measured by counting the words of the dream report. This is plausible since more things can happen in a long dream. The same is valid for global rating scales; the bizarreness/realism scale, for example, correlated to $r = .466$ with dream length (Schredl, 1998b). The simplest correction procedure is the division of the scores by word count, but Trinder, Kramer, Riechers, Fishbein, & Roth (1970) have shown that the relationship between number of dream persons and dream length is non-linear. On the other hand, Schredl (1999b) who analyzed 537 dream reports, have demonstrated that the linear correlation coefficients by far exceed the coefficients of quadratic and cubic correlations and, thus, the division by word count is a good approximation. Urbina (1981) has suggested controlling dream length statistically using analyses of covariance or logistic regressions. Domhoff (1996) has suggested using ratios (e.g., aggressive interaction per dream characters or ratio between aggressive and friendly interactions) instead of absolute values (e.g., aggressions per dream). This approach also has disadvantages if scores for single dreams have to be obtained, e.g. male/female percent cannot be computed for each dream and might vary considerably (0 to 100%) depending on the number of male and female dream characters. Using the number of male or female characters would avoid these problems.

The effect of dream length on the content analytic findings will be especially pronounced if two groups are compared whose dream length differ considerably, e. g. patients with insomnia vs. healthy controls (Schredl, 1998b).

8. Conclusions

Although the method of dream content analysis can be easily applied, several problems have to be considered. The most often encountered topics are depicted in Table 9.

If no appropriate rating scale is available in the literature, it is necessary to develop a new scale. This has to be done before the person constructing the scale has read any of the dreams to be analyzed because this will bias the definitions of the scale's categories by adapting it to the material. Of course, the person can read dreams stemming from other studies in order to get a feeling for how to measure specific topics within a dream. This precaution is also necessary if the scale has to be redefined because of low interrater reliability. The editing of the scale must also be done by a person who is not familiar with the specific dreams of the study. Otherwise, the interrater reliability would improve for the given dream sample but not in general when applied to a new sample of dreams.

In section 6, the problem of validity of dream content analytic scales was discussed. This is not a problem of constructing a scale in a proper way but is based on the simple fact that the dream report did not include all details about the dream experience. For several characteristics,

Table 9. Possible problems in dream content analysis

Problem	What to do
• No appropriate content analytic scale available	• Develop a new scale (before looking at the material to be analyzed)
• Interrater reliability of the scale is low	• Refine the definition of the scale's categories in a more detailed way
• Validity of the scale	• Maybe self-rating scales (e.g. for measuring emotions) or probing (e.g. for measuring colors) are more suitable
• Multiple testing	• Formulate specific hypotheses
• Multiple dreams per participant	• Use repeated measurement analysis or a mixed model approach to account for statistical dependency
• Correlating dream content with trait aspects	• Use a sufficient number of dreams per participant ($N > 20$)

like dream emotions and colors, it seems obvious that specific questions have to be constructed which the dreamer himself/herself answers after reporting or writing down the dream.

Statistical issues are also of importance. The formulation of specific hypotheses will decrease the number of statistical tests and, therefore, increase statistical power since correcting the alpha-level of the tests is not as substantial as if many tests were carried out. For most dream studies, which elicit more than one dream per participant, the number of dreams per person differ and a simple repeated measurement analysis can not be carried out. The application of mixed models for this data set is therefore necessary in order to obtain a maximum of statistical power, for example, by avoiding the use of averaged scores per participant. Correlating dream content characters with trait measures, like personality dimension, will require a reasonable number of dreams per subject in order to obtain meaningful correlation coefficients.

To summarize, despite the various efforts and published findings in dream research, the systematic study of the advantages and problems of the dream content analysis method is still at the beginning. Issues like validity or applying the appropriate statistical procedures have to be studied more thoroughly in the future.

References

- Beck, A. T., & Hurvich, M. S. (1959). Psychological correlates of depression: I. frequency of "masochistic" dream content in a private practice sample. *Psychosomatic Medicine*, 1, 50-55.
- Bernstein, D. M., & Belicki, K. (1995). On the psychometric properties of retrospective dream content questionnaires. *Imagination, Cognition and Personality*, 15, 351-364.
- Brink, S. G., & Allan, J. A. B. (1992). Dreams of anorexic and bulimic women: a research study. *Journal of Analytical Psychology*, 37, 275-297.
- Clark, J., Trinder, J., Kramer, M., Roth, T., & Day, N. (1972). An approach to the content analysis of dream content analysis scales. *Sleep Research*, 1, 118.
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, 37-46.
- Dippel, B. (1988). Manual zur Erfassung des manifesten Trauminhaltes (3. Version, 204 Items). München: Unveröffentlichtes Manuskript.
- Domhoff, G. W. (1996). Finding meaning in dreams: a quantita-
- tive approach. Plenum Press. New York.
- Enke, H., Ohlmeier, D., & Nast, J. (1968). Eine formale Affekt- und Beziehungsanalyse in Traumserien von Patienten mit psychosomatischen Krankheitsbildern. *Zeitschrift für Psychosomatische Medizin*, 15-33.
- Foulkes, D. (1978). A grammar of dreams. New York: Harvester Press.
- Foulkes, D., & Sheperd, J. (1971). Manual for a scoring system for children's dreams. Unpublished manuscript. University of Wyoming.
- Foulkes, D., Sullivan, B., Kerr, N. H., & Brown, L. (1988). Appropriateness of dream feelings to dreamed situations. *Cognition and Emotion*, 2, 29-39.
- Gaillard, J.-M., & Phelippeau, M. (1977). Analysis of dream contents by scaled and rated measurements. *Psychological Medicine*, 7, 275-282.
- Gottschalk, L. A., Gleser, G., & Springer, K. J. (1963). Three hostility scales applicable to verbal samples. *Archives of General Psychiatry*, 9, 254-279.
- Hall, C. S., & Van de Castle, R. L. (1966). The content analysis of dreams. New York: Appleton-Century-Crofts.
- Hauri, P. (1975). Categorization of sleep mental activity for psychophysiological studies. In G. C. Lairy & P. Salzorulo (Eds.), *The experimental study of human sleep: methodological problems*. Elsevier (pp. 271-281). Amsterdam.
- Hauri, P., Sawyer, J., & Rechtschaffen, A. (1967). Dimensions of dreaming: a functional scale for rating dream reports. *Journal of Abnormal Psychology*, 72, 16-22.
- Kramer, M., & Roth, T. (1979). The stability and variability of dreaming. *Sleep*, 1, 319-325.
- Majer-Trendel, K. (1991). Manifester Trauminhalt bei neurotischen Eßstörungen. Medizinische Dissertation. Universität München.
- Merritt, J. M., Stickgold, R., Pace-Schott, E., Williams, J., & Hobson, J. A. (1994). Emotion profiles in the dreams of men and women. *Consciousness and Cognition*, 3, 46-60.
- Montasser, A. (1996). Dreaming in the eating disordered. Universität Mannheim: Unveröffentlichte Diplomarbeit.
- Riemann, D., Beyer, J., Wiegand, M., & Berger, M. (1985). A comprehensive manual for scoring manifest dream content. In W. P. Koella, E. Rüther & H. Schulz (Eds.), *Sleep 1984* (pp. 355-357). Stuttgart: Gustav Fischer Verlag.
- Saul, L., Sheppard, E., Selby, D., Lhamon, W., & Sachs, D. (1954). The quantification of hostility in dreams with reference to essential hypertension. *Science*, 119, 382-383.
- Schredl, M. (1991). Traumerinnerungshäufigkeit und Trauminhalt bei Schlafgestörten, psychiatrischen Patienten und

- Gesunden: Unveröffentlichte Diplomarbeit. Universität Mannheim.
- Schredl, M. (1998a). The stability and variability of dream content. *Perceptual and Motor Skills*, 86, 733-734.
- Schredl, M. (1998b). Träume und Schlafstörungen: Empirische Studie zur Traumerinnerungshäufigkeit und zum Trauminhalt schlafgestörter PatientInnen. *Tectum*. Marburg.
- Schredl, M. (1999a). Die nächtliche Traumwelt: Eine Einführung in die psychologische Traumforschung. Stuttgart: Kohlhammer.
- Schredl, M. (1999b). The problem of dream length in analysis of content. *Perceptual and Motor Skills*, 88, 434-436.
- Schredl, M. (2006). Factors affecting the continuity between waking and dreaming: emotional intensity and emotional tone of the waking-life event. *Sleep and Hypnosis*, 8, 1-5.
- Schredl, M. (2008a). Determining the dreamer's gender from a single dream report: a matching study. *International Journal of Dream Research*, 1, 23-26.
- Schredl, M. (2008b). Laboratory references in dreams: Methodological problem and/or evidence for the continuity hypothesis of dreaming? *International Journal of Dream Research*, 1, 3-6.
- Schredl, M. (2008c). Spontaneously reported colors in dreams: correlations with attitude towards creativity, personality and memory. *Sleep and Hypnosis*, 10, 54-60.
- Schredl, M. (2008d). Traum. München: Reinhardt/UTB.
- Schredl, M., Burchert, N., & Grabatin, Y. (2004). The effect of training on interrater reliability in dream content analysis. *Sleep and Hypnosis*, 6, 139-144.
- Schredl, M., & Doll, E. (1998). Emotions in diary dreams. *Consciousness and Cognition*, 7, 634-646.
- Schredl, M., & Engelhardt, H. (2001). Dreaming and psychopathology: dream recall and dream content of psychiatric inpatients. *Sleep and Hypnosis*, 3, 44-54.
- Schredl, M., & Erlacher, D. (2003). The problem of dream content analysis validity as shown by a bizarreness scale. *Sleep and Hypnosis*, 5, 129-135.
- Schredl, M., Fuchedzhieva, A., Hämig, H., & Schindele, V. (2008). Do we think dreams are in black and white due to memory problems? *Dreaming*, 18, 175-180.
- Schredl, M., & Montasser, A. (1999). Dreaming and eating disorder. *Sleep and Hypnosis*, 1, 225-231.
- Schredl, M., & Reinhard, I. (2008). Dream recall, dream length and sleep duration: state or trait factor. *Perceptual and Motor Skills*, 106, 633-636.
- Schredl, M., Sahin, V., & Schäfer, G. (1998). Gender differences in dreams: do they reflect gender differences in waking life? *Personality and Individual Differences*, 25, 433-442.
- Schredl, M., Schäfer, G., Hofmann, F., & Jacob, S. (1999). Dream content and personality: thick vs. thin boundaries. *Dreaming*, 9, 257-263.
- Schredl, M., Schäfer, G., Weber, B., & Heuser, I. (1998). Dreaming and insomnia: dream recall and dream content of patients with insomnia. *Journal of Sleep Research*, 7, 191-198.
- Schredl, M., Schröder, A., & Löw, H. (1996). Traumerleben von älteren Menschen - Teil 2: Empirische Studie und Diskussion. *Zeitschrift für Gerontopsychologie und -psychiatrie*, 9, 43-53.
- Schredl, M., Schwenger, C., & Dehe, A. (2004). Gender differences in dreams: a matching study. *Counseling and Clinical Psychology Journal*, 1, 61-67.
- Stegie, R. (1986). Zur Anwendung des Gottschalk-Gleser-Verfahrens in einer psychophysiologicalen Traumstudie. In U. Koch & G. Schoefer (Eds.), *Sprachinhaltsanalyse* in der psychiatrischen und psychosomatischen Forschung. Psychologie Verlags Union (pp. 424-437). Weinheim.
- Strauch, I., & Meier, B. (1996). In search of dreams: results of experimental dream research. Albany: State University of New York Press.
- Trinder, J., Kramer, M., Riechers, M., Fishbein, H., & Roth, T. (1970). The effect of dream length on dream content. *Psychophysiology*, 7, 333.
- Urbina, S. P. (1981). Methodological issues in the quantitative analysis of dream content. *Journal of Personality Assessment*, 45, 71-78.
- Winget, C., & Kramer, M. (1979). Dimensions of dreams. Gainesville: University of Florida Press.