The Problem of Dream Content Analysis Validity as Shown by a Bizarreness Scale

Michael Schredl, Ph.D. and Daniel Erlacher, Ph.D.

Content analysis is one of the basic methods used in psychological dream research. Whereas reliability issues have been addressed in the literature quite often, the validity of dream content analysis has rarely been studied in a systematic way. The present study investigated the validity of a bizarreness scale by asking whether an external judge estimates the number of bizarre elements per dream in the same way as the dreamer herself or himself. As reported previously for dream emotions, a marked underestimation of bizarreness by the external judges was found. The findings indicate, therefore, that written dream reports do not yield not a complete picture of the original dream experience and hence the validity of dream content analysis which is based on written dream reports is limited at least in several areas. How severely the validity problem affects the results of content analytic studies and which dream characteristics are most susceptible to this kind of error should be investigated in future studies using a methodology similar to that employed in the present study. **(Sleep and Hypnosis 2003;5(3):129-135)**

Key words: *dream recall frequency, dream content, reliability, stability*

INTRODUCTION

Content analysis is one of the basic methods utilized in psychological dream research (1-3). The following example illustrates this approach. A psychotherapist speculates that depressive patients dream about rejection more often than comparable healthy controls. A research colleague elicits 100 dream reports of healthy subjects and 100 dream reports of patients with depression and constructs a scale with explicit coding rules that measures rejection within the dream, i.e., the occurrence of at least one situation in which the dreamer is rejected by other dream characters is coded as 1 (otherwise 0). It is of major importance that the scale is developed without knowing the dream material to be analyzed.

In the next step, the dream reports are sorted randomly and an external judge rates each dream whether rejection as defined by the content analytic scale is present or not. After the judging procedure, the dreams are assigned to the two groups that are compared statistically in order to test whether depressed patients have reported rejection dreams more often according to their reports.

The advantages of this approach are clearly visible: content differences are reflected quantitatively and are thus accessible for statistical testing. In addition, an external judge using a scale with explicit coding rules minimizes the subjectivity of the experimenter and the study is replicable, e.g., when analyzing

From Central Institute of Mental Health, Mannheim (Dr. Schredl) and Institute of Sport and Sports Science, University of Heidelberg (Dr. Erlacher), Germany

Address reprint requests to: Dr. Michael Schredl Sleep laboratory, Central Institute of Mental Health, PO Box 12 21 20, 68072 Mannheim, Germany, e-mail: Schredl@as200.zi-mannheim.de

Accepted June 3, 2003

new dream samples. These are important quality criteria of common scientific practice. A small influence of subjectivity in judging the dream reports is still present if different independent judges rate the same dream material in the same way. These coefficients are termed inter-rater reliability and are given either as exact agreements (3) or correlation coefficients (4). Systematic studies that investigate cut-offs above which the coefficient is sufficiently high have yet to be carried out, however. This kind of reliability, which is related to how the scale is applied, should not be confused with the reliability coefficients that are related to the stable measurement of inter-individual differences and that play an important role in classical test theory (5). With regard to specific dream characteristics, Schredl (6) demonstrated that up to 20 dream reports should be elicited per person due to the large variability of dream content from dream to dream. This number of dreams would yields a sufficient inter-item consistency. Within this context, the measurement of a single dream is analogous to the response to a test item; with an increasing number of items/dreams the reliability coefficient increases as well (5).

In addition to reliability, validity is an important quality criterion. Validity designates the extent to which the measured score is related to the underlying dimension that should be measured by the instrument/scale. Often investigators (1,3) rely on face validity, e.g., an aggression scale like “Are there any aggressive interactions present within the dream?” measures aggression. This seems plausible and appropriate, but another kind of validity problem, outlined in the following using the example of dream emotions, should be considered. Dream content analysis does not aim at the description of the dream report but at the measurement of the subjective dream experience. Several authors (7-9) emphasize that the dream report–even recorded as elaborately as possible–is only a more or less complete recall of the dream experience including emotions, actions, images, thoughts etc. Colors, for example, are rarely reported spontaneously but specific probing has yielded much higher figures (10). This is also valid for other dream characteristics such as emotions (11,12), tactile sensations (8) and bizarre elements (13). On the other hand, Stern, Saayman and Touyz (14) demonstrated that specific instructions for recording the dreams (focusing on settings in nature or urban settings) may alter dream content markedly. I.e., there is a dilemma that arises between eliciting dream reports without any further questioning and giving instructions for recording specific dream characteristics in a more detailed way.

Schredl and Doll (15) analyzed the validity of two content analytic scales measuring emotions based on a sample of 133 home dream reports. The self-rated dream emotions (positive and negative) measured on two four-point scales (none, mild, moderate, strong) have been chosen as criteria. The same two scales were used for the external judgment procedure (9). The judges were instructed to also code emotions that they can infer from the dream action. The second method used in that study was the emotion scale of Hall and Van de Castle (1) which measures only explicit mentioned emotions (5 classes: anger, apprehension, happiness, sadness, confusion). In the fictive dream example “I saw a monster and ran away.” an external judge is not allowed to code any emotions according the Hall and Van de Castle (1) rules; using the Schredl (9) rating scales a coding of negative emotions is possible and probable. Whereas 0.8% of the dreamers did not report any dream emotions (self-rated), external judgment applying the rating scales yielded 13.5% dream reports that were without emotions and–according to the Hall and Van de Castle scale–57.9% dreams were void of emotions. I.e., the external judges relying solely on the dream report without any further information were not able to extract all the emotions experienced by the dreamer (assuming self-rating of emotions upon awakening are predicative). Interestingly, the correlation coefficients between external judgment and self-rating war high (positive emotions: r=.557; negative emotions: r=.669). With respect to reliability coefficients of about r=.80 (9) these correlations which represents criteria validity can be considered as sufficiently high since validity coefficients can not exceed reliability coefficients. Another study, however, reported much smaller values of criteria validity, e.g., r=.31 (anxiety; 16).

To summarize, the validity of dream content analysis, based on the dream reports, might vary considerably with the investigated dream characteristics. One can imagine that the number of dram characters, for example, might be measured validly, but the measurement of dream emotions is much more complicated and less valid (underestimation by external judges despite the relatively high correlation).

Continuing the work of Schredl and Doll (15), the present study investigated the validity of a bizarreness scale by asking the question whether an external judge estimates the number of bizarre elements per dream in the same way as the dreamer herself or himself was studied. For measuring bizarreness, a relatively narrow definition when compared to that of Hobson et al. (7) was chosen since the authors themselves stated that the coding of improbable events/features are much more difficult than coding events/features which are impossible in waking reality. Since the scope of the present paper did not encompass the development of a bizarreness scale but to investigate validity issues, a scale as simple as possible was chosen. A detailed definition is presented in the method section. In order to avoid the possible effect of instructions on dream content, the participants were instructed to evaluate the bizarre elements directly after recording the dream. It was expected–parallel to the measurement of dream emotions–that a marked underestimation of the bizarre elements by the external judges would be found despite a relatively high inter-correlation of the two measures.

METHODS

Measurement instruments

The questionnaire was comprised of questions regarding socio-demographic variables (age, gender, study subjects) and a seven-point scale measuring dream recall frequency (17). The scale's re-test reliability is high (r=.83; average of 70 days; 17). In addition, the participants were given a form with the following written instruction on the top: "Please record your dream in as much detail and elaborately as possible. Allow enough time to avoid forgetting something." The bizarreness scale (see below) and a dream evaluation form were given to the participant in a sealed envelope (see procedure section). The dream evaluation form includes the instruction that dream content must not be altered during the evaluation process.

Bizarreness Scale

The bizarreness scale applied in the present study was based on the research of Hobson et al. (7) and Revonsuo and Salmivalli (13). In order to facilitate the judgment process, only elements that are impossible or extremely improbable in waking life were to be coded. In addition, a definition of non-bizarre elements was included.

Definition of "Bizarreness"

General: Objects/actions/persons, etc. are defined as bizarre if they do not exist or are impossible in waking-life reality.

1. Incongruity
- An element inconsistent with waking life (e.g., a dog talking, dreamer has three arms)
- Discrepancy from physical laws (e.g., flying, time travels for example into the Middle

Ages)
- Mismatching features (e.g., standing in a burning house and freezing, failing an important examination and feel joy)

2. Discontinuity
- Changes of features: elements disappear, appear suddenly or change shape (e.g., dreamer talks to a friend who changes into an animal)
- Impossible or very improbable alterations in familiar settings (e.g., being in the living room and snake heads jut out of the wall)

3. Uncertainty
- Obscure or undetermined elements (e.g., unknown monster, plant-like object emits indefinable noises)

Notice: The examples given above in parentheses serve only as illustrations for better comprehension of the bizarreness definition.

Definition of "Non-Bizarreness"

General: Objects/actions/persons, etc. that are possible in waking life are regarded as non-bizarre. These might be extraordinary or improbable but nevertheless possible.

The following examples that are grouped like the categories presented above depict non-bizarre elements.

1. Incongruity
- Madonna is singing in your living room.
- Wax-coated dumplings are on the lunch table.

2. Discontinuity
- The living room wallpaper is green within the dream and not white, as it really is.

3. Uncertainty
- You are in an unfamiliar setting like an isolated island.

Procedure and Participants

The participants received the questionnaire, the recording form and a sealed envelope which was to be opened after recording the dream. Subsequent to reading the bizarreness and non-bizarreness definitions, the participants evaluated and recorded the bizarre elements that occurred in their dream. In addition, a short explanation as to why the element was rated as bizarre was requested. The number of bizarre elements per dream was the variable included in subsequent analysis. Dream reports then were typed and coded separately by two independent judges. These were trained using 40 dream reports stemming from a different study (18). Within this training period discrepancies were resolved by discussion. The inter-rater reliability was determined as the Pearson correlation for the number of bizarre elements per dream.

Overall, 46 psychology students (38 women, 8 men) participated. Their mean age was 22.0±3.7 years. Each participant contributed only one dream.

RESULTS

On average, the participants recalled dreams on about three mornings per week (2.79±2.17). The mean word count of the dream reports amounted to 176.2±144.6 words. The inter-rater reliability between Rater 1 and Rater 2 was r=.910 (Pearson correlation for the number of bizarre elements per dreams). The exact agreement of recognizing the bizarre element in the same position of the dream text was, however, relatively low (42%) between the two judges. The means of the number of bizarre elements variable are depicted in Table 1. Whereas the correlations between judges and self-rating were very high (Self-rating - Rater 1: r=.868; Self-rating - Rater 2:

Table 1. Means and standard deviations (SD) of "Number of bizarre elements per dream" and the comparison with the self-rated value

Variable	Mean±SD	Statistical test[1]
Self-rating	2.54±3.24	
Rater 1	1.24±2.65	t=5.5 p<.0001
Rater 2	0.65±1.18	t=5.2 p<.0001
Mean of Rater 1 and Rater 2	0.95±1.88	t=5.6 p<.0001

[1] t-test for dependent samples (N = 46)

r=.750; Self-rating - Rater 1/2: r=.848; all p < .001), a marked underestimation of bizarre elements by the judges was found. In addition, the means of Rater 1 and Rater 2 also differed significantly (t=2.4, p=.0203).

Dream example

"My two sons and I are on the wooden terrace of a stadium. The two are quarreling. They stand opposite each other. I do not understand what they are saying. Pierre lay his hand onto Marcel's shoulders (who is much smaller). He is shaking him, Marcel resists the pressure. I shout: "Stop." In that moment Marcel falls backwards down the steep wooden seat rows and interspaces that lay behind him (several meters). I go down. Marcel is lying there, calm, apparently uninjured, with a sleeping face. I am very excited and consider whether he is still alive. I address him but he does not respond."

The dreamer reported two bizarre elements within this dream sequence. First, she stated that no one would be unharmed after such a fall. The two external judges also rated this element as bizarre. The second bizarre element is related to the fall itself. The dreamer made two drawings (see Figure 1) in order to illustrate that it is impossible to fall down this way (because of the interspaces). The first drawing made during recording the dream was not as illustrative as the second drawing. The judges who were given only the text did not code this element as bizarre.

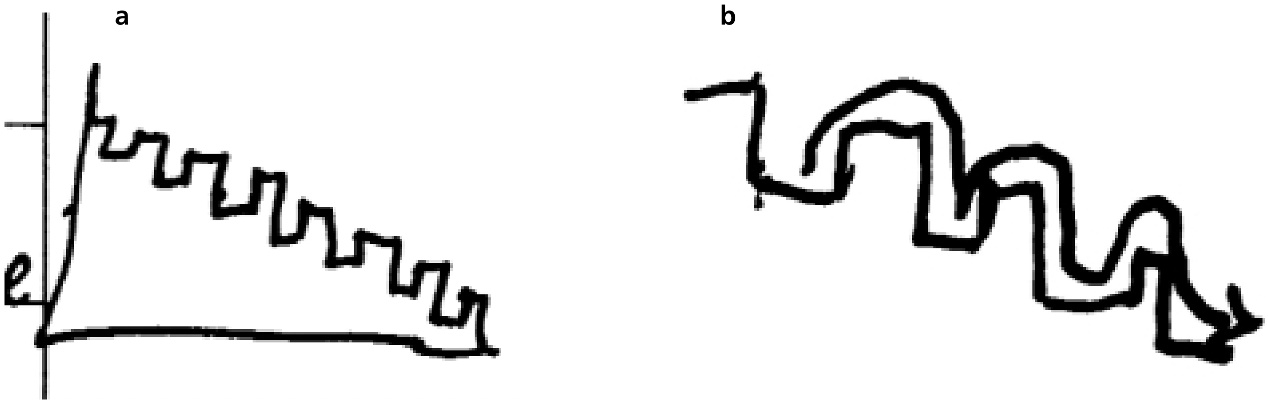

Figure 1. Drawings by the dreamer illustrating bizarre elements (see dream example section)
Figure 1a was drawn with the dream report showing the seating rows.
Figure 1b was drawn along with the rating of the bizarre element: fall of the Marcel dream character.

DISCUSSION

The results support the hypothesis formulated at the beginning that external judges code fewer bizarre elements than the dreamer herself or himself. In the following, the implications of this finding are discussed.

First, the issue of inter-rater reliability should be examined carefully. Despite the high correlation, a marked and significant mean difference between the judges was found, i.e., Rater 1 coded more bizarre elements than Rater 2. Regarding the application of the scale, one should consider a more intensive training period in order to minimize discrepancies. In general, it is desirable to specify inter-rater reliability not only by the correlation coefficient but also by mean comparisons (if exact agreement is not computed). This has been done very rarely in the literature (9). It is also important that the reliability of the variable used in the analysis, namely "number of bizarre elements per dream", was sufficiently high despite the relatively low exact agreement of the two judges. Again generally speaking, is seems adequate to compute the inter-rater reliabilities for the indices presented, e.g., male/female percent, aggression percent (3) in addition to the percentages of exact agreement of the content analytic procedure itself.

In what way should the differences between self-ratings and external judging be interpreted? First, it seems difficult to estimate the reliability of self-ratings. It might be reduced by misinterpretations of the bizarreness scale or evaluation errors. The method of computing inter-rater reliability coefficients is not applicable but maybe the problem can be minimized by training the participants in the same way that the judges have been trained (coding independent dream material and discussing discrepancies). One should keep in mind, however, that focusing on particular dream characteristics before recording (and actually having) the dream might affect dream content considerably (see discussion below).

The aim of dream content analysis is the measurement of the original dream that has been dreamt and Hobson and Stickgold (19) summarized their experience: "One lesson that we have learned is that the open ended inquiry into dream mentation that has been typical of most past work (including our own) is grossly inadequate (p. 10)." The present results and the findings of Schredl and Doll (15) clearly indicate that dream content analysis based on the external judgment of written dream reports yields marked underestimation with regard to the number of bizarre elements and dream emotions. I. e., a dream report that is recorded by a participant in response to an open ended question (see method section of the present study) and the subsequent carried out content analysis is insufficient for eliciting all subjective experiences of the dream. On the other hand, it seems plausible that for other dream characteristics, e.g., occurrence of sexual interaction, aggression, death themes, this problem plays only a minor role. I.e., these kinds of studies (comparison of self-ratings with external judgments) should be carried out in a systematic way for a variety of different dream characteristics.

In order to solve the problem of validity, Hobson and Stickgold (19) suggested using affirmative phenomenological probes, i.e., to ask the dreamer explicitly about the characteristic studied. A problem not addressed by the authors is the possible effect of the instructions, e.g., to record dream emotions explicitly (12), on subsequent dreams. Stern, Saayman and Touyz (14) were able to demonstrate that instructions did affect dream reports. The question whether these instructions result in a more detailed description of the dream experiences (desirable) or whether a bias regarding dream contents is the result should be studied in a more detailed. way. One might design a study, for example, which used the methodology of the present study (instructions after recording the dream) and a second sample of participants who receive the instructions right on the beginning of the study. Another idea that, for example, was applied by Leuschner et al. (20) and is discussed by Hobson and Hoffman (21): using drawings made by the dreamer in addition to the written dream report. It is possible to use coding systems with high inter-rater reliability coefficients (20).

To summarize, the written dream report did not yield a complete picture of the original dream experience and hence the validity of dream content analysis based on written dream reports is limited at least in several areas. How severely the validity problem affects the results of content analytic studies and which dream characteristics are most susceptible to this kind of error should be investigated in future studies using a methodology similar to that of the present study.

REFERENCES

1. *Hall CS, Van de Castle RL. The content analysis of dreams. New York: Appleton-Century-Crofts, 1966.*
2. *Winget C, Kramer M. Dimensions of dreams. Gainesville: University of Florida Press, 1979.*

3. *Domhoff GW. Finding meaning in dreams: a quantitative approach. New York: Plenum Press, 1996.*

4. *Hauri P, Sawyer J, Rechtschaffen A. Dimensions of dreaming: a functional scale for rating dream reports. Journal of Abnormal Psychology 1967;72:16-22.*

5. *Rost J. Lehrbuch Testtheorie: Testkonstruktion (Textbook test theory: Test Construction). Bern: Huber, 1996.*

6. *Schredl M. The stability and variability of dream content. Perceptual and Motor Skills 1998;86:733-734.*

7. *Hobson JA, Hoffman SA, Helfand R, Kostner D. Dream bizarreness and the activation-synthesis hypothesis. Human Neurobiology 1987;6:157-164.*

8. *Strauch I, Meier B. In search of dreams: results of experimental dream research. Albany: State University of New York Press, 1996.*

9. *Schredl M. Die nächtliche Traumwelt: Eine Einführung in die psychologische Traumforschung. Stuttgart: Kohlhammer, 1999.*

10. *Kahn E, Dement W, Fisher C, Barmack JE. Incidence of color in immediately recalled dreams. Science 1962;137:1054-1055.*

11. *Nielsen TA, Deslauriers D, Baylor GW. Emotions in dream and waking event reports. Dreaming 1991;1:287-300.*

12. *Merritt JM, Stickgold R, Pace-Schott E, Williams J, Hobson JA. Emotion profiles in the dreams of men and women. Consciousness and Cognition 1994;3:46-60.*

13. *Revonsuo A, Salmivalli C. A content analysis of bizarre elements in dreams. Dreaming 1995;5:169-187.*

14. *Stern DA, Saayman GS, Touyz SW. A methodological study of the effect of experimentally induced demand characterictics in research of nocturnal dreams. Journal of Abnormal Psychology 1978;87:459-462.*

15. *Schredl M, Doll E. Emotions in diary dreams. Consciousness and Cognition 1998;7:634-646.*

16. *Riemann D, Beyer J, Wiegand M, Berger M. A comprehensive manual for scoring manifest dream content. In: Koella WP, Rüther E, Schulz H, eds. Sleep 1984. Stuttgart: Gustav Fischer Verlag, 1985;355-357.*

17. *Schredl M. Messung der Traumerinnerung: siebenstufige Skala und Daten gesunder Personen. Somnologie 2002;6:34-38.*

18. *Schredl M, Wittmann L, Ciric P, Götz S. Factors of home dream recall: a structural equation mode. Journal of Sleep Research 2003;12:133-141.*

19. *Hobson JA, Stickgold R. Dreaming: a neurocognitive approach. Consciousness and Cognition 1994;3:1-15.*

20. *Leuschner W, Hau S, Brech E, Volk S. Disassociation and reassociation of subliminally induced stimulus material in drawings of dreams and drawings of waking free imagery. Dreaming 1994;4:1-27.*

21. *Hobson JA, Hoffman S. Picturing dreaming: Some features of the drawings in a dream journal. In: Bosinelli M, Cicogna P, eds. Psychology of dreaming. Bologna: CLUEB, 1984;11-30.*