The effect of Training on Interrater Reliability in Dream Content Analysis

Michael Schredl, Ph.D., Natalie Burchert, Ph.D., and Yvonne Gabatin, Ph.D.

Content analysis is an important and a frequently applied tool in dream research. Hall and Van de Castle (1) stressed the importance of interrater reliability in the application of content analytic scales, i.e., how good is the agreement between two judges scoring the same dream material independently? The present study investigated the effect of rater training on the interrater reliability of scales developed by Schredl (2). Three samples of 100 dream reports each have been analyzed by two inexperienced judges who received two training sessions after coding 100 and 200 dreams. The results indicate that the training of raters has a positive effect on interrater reliability and the mean differences of some scales (nominal and interval scales) but not for ordinal scales (e.g., dream emotions). It remains unclear how much training is necessary for different scales and whether it might be necessary to improve the scales themselves if extensive training does not result in a desired improvement in interrater reliability. Thus, more studies investigating rater training for different systems of dream content analysis are needed. **(Sleep and Hypnosis 2004;6(3):139-144)**

Key words: *dream content analysis, interrater reliability, rater training*

INTRODUCTION

Content analysis is an important and a tool in dream research that is applied quite often. The typical procedure is as follows: dream reports of, for example, two groups are randomized and are scored along pre-defined scales by external judges who do not know which dreams stems from which group. An illustrative example is the occurrence of physical aggression in dreams. For the male sample, 15.9% of the dreams examined in this study included physical aggression, whereas only 5.1% of the women's dreams are characterized by this trait (effect size: d=0.36; p=.003; 3). This example elucidates the major purpose of dream content analysis: quantifying specific dream aspects in order to carry out statistical analyses. This methodology matches common scientific criteria such as possible replication by other research groups, minimizing experimenter effects by explicit coding rules and blind judging of the material. The first comprehensive system for dream content analysis was published by Hall and Van de Castle (1); even today it is the rating system that is applied most often (4). Other scales and manuals can be found in Winget and Kramer (5) and Schredl (2).

Hall and Van de Castle (1) stressed the importance of interrater reliability in the application of content analytic scales, i.e., how well do two independent judges scoring the same dream material agree? Within in their

From the Central Institute of Mental Health, Mannheim, 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

reliability study (N=50 to 100 dream reports), the authors reported, for example, the following figures: perfect agreement for all four different aspects (presence, single/group, gender, identity) of dream characters is 74%. For the correct coding of social interactions, the figures are lower: aggression 54%, friendliness 61% and sexuality 64%. These figures show that the raters code the same passage of the text in a similar way. Hall and Van de Castle (1) pointed out that different coefficients can be used to represent interrater reliability (see discussion).

Even when Domhoff's (4) advice to apply the system to 25 to 30 dream reports and read through and code the 10 dream reports provided by Hall and Van de Castle and compare the own codings with their codings is used, follow-up studies failed to achieve the same reliability figures as reported by the authors of the rating system. Sandler et al. (6, 7), for example, reported perfect agreement on all four aspects of dream characters of 62% (H & VC: 74%), activities 50% (H & VC: 85%) and emotions 44% (H & VC: 63%). Schredl et al. (8) also reported lower figures for several categories: all features of dream characters 69% (H & VC: 74%), aggression 43% (H & VC: 54%), friendliness 44% (H & VC:61%) and sexuality 43% (H & VC:64%). This might be explained by the fact that Hall and Van de Castle are more experienced in coding dreams (each of them coded over 10.000 dreams) whereas the judges of the subsequent studies have been rather inexperienced in the methodology of dream content analysis. However, whether training has a positive effect on interrater reliability has not yet studied in a systematic way.

The present study is an attempt to fill in this gap and investigate the effect of rater training on the interrater reliability of scales developed by Schredl (2). It was expected that training does have a positive effect, i.e., yields higher interrater reliability coefficients after the training and, in addition, that absolute differences which might occur in the codings of the rates will decrease with training.

METHODS

Participants

The raters of the study have been two female psychology students (ages: 20 and 21 years). They had no experience in dream content analysis or in other types of quantitative methods.

Materials and Procedure

The applied rating scales have been developed by Schredl (9) based on a previously published German dream manual (10). The idea to apply global ratings scales dates back to Hauri, Sawyer and Rechtschaffen (11). The following scales have been included: bizarreness (four-point scale: 1=realistic to 4=two or more bizarre elements), positive dream emotions (four-point scale: 0=none, 1=mild, 2=moderate, 3=strong), negative dream emotions (same four-point scale), number of dream persons, verbal interaction (occurrence: 1=Yes, 0=No), physical interaction, four aggression subscales (verbal vs. physical, dreamer as aggressor (outgoing) vs. dream as recipient (receiving)), problems (0=no problems, 1=minor problems, 2=major problems). A cumulative index of aggression was constructed from the four subscales (at least one type of aggression is present in the dream). The explicit coding rules and introductory comments to the scales (in German) can be obtained from the first author.

Three sets a 100 dream reports have been analyzed in the present study. The reports were collected by Schredl et al. (12). The mean lengths of the randomly drawn dream samples were 182.2±157.6 words (1-100), 173.7±149.9 words (101-200) and 138.3±108.7 words (201-300).

After reading the coding rules of the scales and a short introduction into the method of dream content analysis, the raters scored the first 100 dream reports independently. Then, the raters met with the first author and discussed and resolved the inconsistencies within the ratings. After scoring a second sample of 100 dream reports, a similar training session was held. Lastly, the raters scored another 100 dream reports for computed interrater reliabilities.

Interrater reliability coefficients were computed according to the scales' measurement levels: Pearson correlations (number of dream persons), Spearman rank correlations (bizarreness, dream emotions, problems) and exact agreement for the 0/1-coded scales. The difference between the correlation coefficients were tested by a formula given by Bortz (13). The formula for comparing percentages statistically can be found in Domhoff (4). Since an increase in interrater reliability was expected, probabilities have been one-tailed. In addition, the averaged ratings of rater 1 and rater 2 have been compared (t-test for dependent samples (number of dream persons), sign rank test for the ordinal scales and differences in percentages (nominal scales). These tests have been carried out two-tailed.

RESULTS

In Table 1, the ratings of the nominal scales for the three dream samples are depicted. No systematic differences between the two raters

Table 2. Interrater reliabilities of the nominal scales (comparisons between groups)

Scale	1-100	101-200	201-300	Test (1 vs. 2)		Test (1 vs. 3)		Test (2 vs. 3)	
				z=	p=	z=	p=	z=	p=
Verbal interaction	87%	99%	95%	3.8	.0001	2.0	.0213	-1.8	.9618
Physical interaction	91%	96%	95%	1.5	.0719	1.1	.1314	-0.3	.6336
Aggression (verbal, outgoing)	94%	95%	98%	0.3	.3781	1.5	.0677	1.2	.1185
Aggression (verbal, receiving)	97 %	95%	99%	-0.7	.7664	1.0	.1479	1.8	.0381
Aggression (physical, outgoing)	97%	96%	99%	-0.4	.6501	1.0	.1479	1.4	.0762
Aggression (physical, receiving)	93%	93%	98%	0.0	1.000	-1.8	.0375	-1.8	.0375
Aggression (total)	88%	88%	96%	0.0	1.000	2.2	.0155	2.2	.0155

Table 1. Percentages of the nominal scales (comparisons between raters)

Group	Scale	Rater 1	Rater 2	Stat.	test
				z=	p=
1-100	Verbal interaction	62%	65%	-0.4	.6594
	Physical interaction	24%	29%	-0.8	.4226
	Aggression (verbal, outgoing)	8%	10%	-0.5	.6206
	Aggression (verbal, receiving)	10%	10%	0.0	1.000
	Aggression (physical, outgoing)	5%	8%	-0.9	.3864
	Aggression (physical, receiving)	12%	13%	-0.2	.8306
	Aggression (total)	22%	24%	-0.3	.7366
101-200	Verbal interaction	61%	62%	-0.1	.8844
	Physical interaction	16%	16%	0.0	1.000
	Aggression (verbal, outgoing)	9%	8%	0.3	.7996
	Aggression (verbal, receiving)	6%	9%	-0.8	.4182
	Aggression (physical, outgoing)	7%	7%	0.0	1.000
	Aggression (physical, receiving)	9%	10%	-0.2	.8092
	Aggression (total)	21%	21%	0.0	1.000
201-300	Verbal interaction	65%	64%	0.1	.8824
	Physical interaction	15%	16%	-0.2	.8450
	Aggression (verbal, outgoing)	5%	5%	0.0	1.000
	Aggression (verbal, receiving)	4%	3%	0.4	.6996
	Aggression (physical, outgoing)	4%	3%	0.4	.6996
	Aggression (physical, receiving)	3%	3%	0.0	1.000
	Aggression (total)	11%	11%	0.0	1.000

Table 3. Means and standard deviations of the ordinal and interval scales (comparisons between raters)

Group	Scale	Rater 1	Rater 2	Stat.	test[1]
				z=	p=
1-100	Bizarreness	2.51 ± 1.09	2.50 ± 1.09	12	.8628
	Positive emotions	0.47 ± 0.87	0.41 ± 0.79	27	.3899
	Negative emotions	1.22 ± 1.19	1.06 ± 1.21	123.5	.0284
	Dream persons	3.20 ± 2.65	4.01 ± 3.43	-5.8	.0001
	Problems	1.03 ± 0.87	0.85 ± 0.77	139.5	.0004
101-200	Bizarreness	2.77 ± 1.16	2.38 ± 1.09	537	.0001
	Positive emotions	0.37 ± 0.77	0.27 ± 0.74	40	.1414
	Negative emotions	1.44 ± 1.14	1.04 ± 1.13	375.5	.0001
	Dream persons	3.62 ± 2.87	3.55 ± 2.89	0.9	.3564
	Problems	1.05 ± 0.81	0.95 ± 0.82	63	.0755
201-300	Bizarreness	1.97 ± 0.96	2.60 ± 1.11	-809	.0001
	Positive emotions	0.37 ± 0.79	0.43 ± 0.89	-24	.3303
	Negative emotions	0.86 ± 0.93	0.91 ± 1.11	-43.5	.5110
	Dream persons	3.33 ± 2.83	3.32 ± 2.98	0.1	.9005
	Problems	0.71 ± 0.81	0.82 ± 0.70	-95	.0989

[1]Statistical tests: Sign rank test, except for "Dream persons" t-test

Table 4. Interrater reliabilities of the ordinal and interval scales (comparisons between groups)

Scale	1-100	101-200	201-300	Test (1 vs. 2)		Test (1 vs. 3)		Test (2 vs. 3)	
				z=	p=	z=	p=	z=	p=
Bizarreness	.765	.689	.779	-1.1	.8705	0.0	.4046	1.4	.0852
Positive emotions	.642	.512	.682	-1.4	.9140	0.5	.3098	1.9	.0312
Negative emotions	.825	.811	.711	-0.3	.6159	-2.0	.9756	-1.7	.9531
Dream persons	.926	.966	.964	2.8	.0027	2.6	.0049	-0.2	.5802
Problems	.816	.764	.634	-1.0	.8334	-2.8	.9971	-1.8	.9636

occurred. Significant increases in interrater reliability (exact agreement) have been found for verbal interaction, physical aggression (received), verbal aggression (received) and total aggression. Marginally significant increases have been detected for the other scales. Overall, the coefficients exceeded 87%.

The means of the ordinal and interval scales of rater 1 and rater 2 are depicted in Table 3. Only for the purpose of clear presentation, means were also computed for the ordinal scales. For the bizarreness scale, the means of sample 2 and 3 differed significantly, though in opposite directions. The mean estimates of the positive dream emotions between raters have been comparable. At first, problems in dreams and negative dream emotions were coded more often by rater 1, but in the third dream sample the means did not differ significantly. A different coding regarding the number of dream persons occurred only in the first rating period. A significant increase in interrater reliability was only found for the person scale. For the other scales a marked decrease in coefficients occurred sometimes, although most coefficients exceeded r=.70.

DISCUSSION

The results of the present study indicate that the training of raters has a positive effect on interrater reliability and mean differences of some scales that have been developed for the purpose of dream content analysis. Marked improvements have been found for the nominal scales and the persons scale, even though the coefficients have been quite high at the beginning. On the other hand, the interrater reliability coefficients of the ordinal scales, like positive and negative dream emotions, bizarreness and problems, did not change in the expected direction; solely the mean differences for negative emotions and problems diminished.

The question arises whether a more extensive training is necessary for the ordinal scales that are more sophisticated in the applications since, in addition to explicitly mentioned emotions, emotion can also be inferred by the dreamer's actions (e.g., I see a monster and ran away; 14). It will be interesting to carry out a similar study for the Hall and Van de Castle system of dream content analysis in order to determine how much training is necessary for previously inexperienced judges to achieve the coefficients reported by the authors since subsequent studies (e.g., 6-8) partly reported much lower figures.

Another question is about defining a possible cut-off value, i.e., above which value coefficients can be classified as acceptable and sufficient. Hartmann, Rosen and Rand (15), for example, defined a threshold of r=.60; several scales with lower interrater reliability coefficients have been dropped from further analyses. The reliability coefficients in the Hauri, Sawyer and Rechtschaffen (11) ranged between r=.59 and r=.69 and in three reliability studies (9,16,17) the coefficients mostly exceeded r=.70 (like in the present study). Reviewing the literature, however, reveals that recommendations or critical values have not yet been published.

Especially difficult is the interpretation of exact agreements as have been published for the Hall and Van de Castle system (1), for example. This will be illustrated by the following example: A content analysis regarding bizarre elements within the dream yielded an exact agreement between two raters of 42% (presence and subclass correct), whereas the interrater reliability for the number of bizarre elements per dream amounted to r=.910 (18). Since the second variable (number of bizarre elements) was used for subsequent analyses, the high correlation represents the appropriate reliability coefficient. This means that reliability coefficients should be computed for the variable or indices (cf. male/female percent in 4) used in the statistical analyses. Hall and Van de Castle (1) pointed out this approach already, e.g., correlating the number of elements within 5 dreams.

In case individual dream scores are computed in a study (19), it should be taken into consideration that the number of dreams per person does affect the reliability of these scores. As in classical test theory, the larger the number of items (here: dreams), the higher is the reliability coefficient of the total score. Schredl (20) was able to demonstrate that for some scales up to 20 dreams per person are necessary to achieve stable and reliable measurement which is necessary, for example, if dream characteristics are correlated with personality variables.

To summarize, rater training does improve interrater reliability in dream content analysis. It remains unclear, however, how much training is necessary for different scales and whether it might be necessary to improve the scales themselves if extensive training does not result in a desired improvement in interrater reliability. Thus, more studies investigating rater training for different systems of dream content analysis are needed.

REFERENCES

1. *Hall CS, Van de Castle RL. The content analysis of dreams, New York: Appleton-Century-Crofts, 1966.*

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

3. *Schredl M, Sahin V, Schäfer G. Gender differences in dreams: do they reflect gender differences in waking life? Personality and Individual Differences 1998;25:433-442.*

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

5. *Winget C, Kramer M. Dimensions of dreams, Gainesville: University of Florida Press, 1979.*

6. *Sandler L, Kramer M, Fishbein H, Trinder J. Interlaboratory reliability of the Hall-Van de Castle scale. Psychophysiology 1969;6:248-249.*

7. *Sandler L, Kramer M, Trinder J, Fishbein H. Interlaboratory reliability of the Hall-Van de Castle characters, social interaction, activities and emotional scales. Psychophysiology 1970;7:333-334.*

8. *Schredl M, Ciric P, Bishop A, Gölitz E, Buschtöns D. Content analysis of German students' dreams: comparison to American findings. Dreaming 2003;13:237-243.*

9. *Schredl M. Traumerinnerungshäufigkeit und Trauminhalt bei Schlafgestörten, psychiatrischen Patienten und Gesunden. Universität Mannheim: unveröffentlichte Diplomarbeit, 1991.*

10. *Dippel B, Riemann D, Majer-Trendel K, Berger M. Untersuchungen zum manifesten Trauminhalt bei Anorexie- und Bulimiepatienten-ein Zwischenbericht. Psychotherapie, Psychosomatik, Medizinische Psychologie 1988;38:199-204.*

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

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

13. *Bortz J. Statistik für Sozialwissenschaftler, Berlin: Springer, 1999.*

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

15. *Hartmann E, Rosen R, Rand W. Personality and dreaming: boundary structure and dream content. Dreaming 1998;8:31-39.*

16. *Schredl M. Träume und Schlafstörungen: Empirische Studie zur Traumerinnerungshäufigkeit und zum Trauminhalt schlafgestörter PatientInnen, Marburg: Tectum, 1998.*

17. *Schredl M, Schröder A, Löw H. Traumerleben von älteren Menschen - Teil 2: Empirische Studie und Diskussion. Zeitschrift für Gerontopsychologie und–psychiatrie 1996;9:43-53.*

18. *Schredl M, Erlacher D. The Problem of Dream Content Analysis Validity as Shown by a Bizarreness Scale. Sleep and Hypnosis 2003;5:129-135.*

19. *Schredl M, Hofmann F. Continuity between waking activities and dream activities. Consciousness and Cognition 2003;12:298-308.*

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