Title: Training Effects with the Observer- Rated Cognitive Errors and the Coping Action Patterns Scales

Authors: Kramer U, de Roten Y, Drapeau M

Journal: Swiss Journal of Psychology

Year: 2011

Issue: 70

Volume: 1

Pages: 41-46

DOI: 10.1024/1421-0185/a000037

Running Head : TRAINING IN ASSESSING COGNITIVE ERRORS AND COPING

Training Effects in Using the Observer-Rated Cognitive Errors and Coping Action Patterns Rating Scales

Ueli Kramer[1], Yves de Roten[1], & Martin Drapeau[2]

[1] University of Lausanne, Switzerland

[2] McGill University, Canada

All correspondence concerning this article should be addressed to Dr Ueli Kramer, Institute of Psychotherapy, Department of Psychiatry-CHUV, University of Lausanne, Cèdres-Cery, CH-1008 Prilly-Lausanne, Switzerland. Ph +41-21-643 64 62, Fax +41-21-643 65 93. E-mail: ueli.kramer@chuv.ch

Running Head : TRAINING IN ASSESSING COGNITIVE ERRORS AND COPING

Training Effects in Using the Observer-Rated Cognitive Errors and Coping Action Patterns Rating Scales

Abstract

The use of observer-rated rating scales requires the training of raters until they become reliable in using the scales. However, few studies properly report how training in a given rating scale is conducted or how it should be conducted. This study examined progress in inter-rater reliability over 6 months of training in two observer-rating scales, the Cognitive Errors Rating Scale and the Coping Action Patterns Rating Scale. The evolution of the Intra-Class Correlation Coefficients was modelled using Hierarchical Linear Modelling (HLM). Results showed an overall training effect, as well as an effect of the basic training phase and of the rater calibration phase, with the latter being smaller than the former. The results are discussed in terms of implications for rater training in psychotherapy research.

Key-Words: Observer-Rating, Process Research, Rater Training, Reliability, Cognitive Errors, Coping Action Patterns

Training Effects in Using the Observer-Rated Cognitive Errors and Coping Action Patterns Rating Scales

Observer-rated scales applied to different types of interview formats (e.g., therapy session, assessment interview) are common in psychotherapy research. By means of observer-based methods, relevant clinical and research concepts, such as defense mechanisms (Perry & Henry, 2004) or core conflictual relationship themes (Luborsky, & Crits-Christoph, 1998) may be assessed reliably by trained raters. An important implication of the use of rating scales in psychotherapy research is the necessity of rater training, which is usually highly time- and cost-intensive (e.g., Lorber, 2006; Schanche et al., 2008). Thus, there is a need to define and investigate efficient and cost-effective training formats.

The present report aims at investigating a two-phase training format formulated for the training for assessment of defense mechanisms (Perry & Henry, 2004): (1) basic training and (2) rater calibration phase. During the first phase, the participants learn in a small group the concepts of the rating scales, accommodate to the rating manual that describes the method and learn to apply the rating system to selected training cases. Problems and questions are discussed in the group, in the presence of an expert. At the end of this phase, the raters should be able to rate the training cases reliably. During the second phase, the raters are confronted to new material, not pre-selected, but stemming from various research projects. The participants rate individually and discuss their views with the expert view, and later with another participant's view. Questions and problems are discussed in individual supervision sessions provided by the expert. The calibration phase aims at avoiding "rater drift", *i.e.*, ratings becoming more and more divergent over time without regular supervision sessions (Perry, & Henry, 2004). Such a training format is a time-consuming process; the duration, total time investment and total cost may depend on the complexity of the constructs assessed, the precision of their operationalization and of the rater training, the initial competency of the raters, the size of the training group, among other factors.

A few studies have investigated these factors influencing efficiency and cost-effectiveness of the rater training. For the basic rater training, Schanche et al. (2008) presented data showing that the reliability after a training with university clinical psychology students with minimal clinical experience was satisfying, in particular when the training procedure focused on less concepts at a time. In a controlled study, Lorber (2006) showed that ratings of harsh parental discipline provided by minimally trained raters were as good as the ones provided by raters that were trained according to a very time-consuming “golden standard” procedure, assessed at a one-time characterization of the subjects. However, when looking at the differences between means of the raters, Lorber reported very poor reliability coefficients for the minimally-trained raters. Finally, for the rater calibration phase, a rater training format focusing specifically on the problematic categories for the specific rater may be a cost-reducing strategy, as suggested by Tractenberg, Schafer and Morris (2001) in the context of clinical dementia ratings. In conclusion, there remains a dearth of research into the training effects of psychotherapy research observer-rated scales (Cichetti et al., 2006; Invancevich, 1979). In particular, no study has examined the evolution of training effects – defined as the increase of rater competency in the particular rating scale - as a function of training phase. The present study aims at addressing this question by examining change in inter-rater reliability coefficients over the course of a training in applying two observer-rated methods used in cognitive behavioral psychotherapy process and outcome research : the Cognitive Errors Rating Scale (Drapeau, Perry, & Dunkley, 2008) and the Coping Action Patterns Rating Scales (Perry, Drapeau, Dunkley, & Blake, 2005). We aim at describing the evolution of training effects in these specific rating scales across both training phases.

Method

Participants

In total, $N = 9$ subjects participated in the study; out of these, seven (77%) were female. Their initial training level was heterogeneous: four (44%) were Master's level students in clinical psychology; four (44%) were licensed psychologists (Master level) with a mean of 2 years (SD = 2) of post degree experience; and one (11%) had recently completed a Ph.D. The participants were selected for the training by the supervisor and first author based on an intake interview assessing the participant's motivation in using the rating scales for their own study (see Procedure section). All participants of the training provided written consent for the use of their data for this study.

Instruments

The *Cognitive Errors Rating Scale* (CERS; Drapeau et al., 2008) is an observer-rated system which can be used to rate cognitive errors in any type of transcribed interview, the rating being event-based. It assesses 15 different cognitive errors, based on the works of J. Beck (1995) and A. T. Beck (1976): (1) Fortune-telling, (2) Labeling, (3) Overgeneralizing, (4) All-or-nothing, (5) Discounting the positive/negative, (6) Emotional reasoning, (7) Magnification/minimization of positive/negative, (8) Mental filter, (9) Should and must, (10) Tunnel vision, (11) Jumping to conclusions, (12) Mind-reading, (13) Personalization, (14) Inappropriate blaming of self, (15) Inappropriate blaming of others. Examples of ratings for several categories and a clinical application may be found in Kramer, Bodenmann and Drapeau (2009). The valence (positive or negative) of each error is described. Assessment of cognitive errors profile per interview is done by counting up the errors per category. Based on the works of Lefebvre (1981), the cognitive errors can be classified into four higher-order categories or clusters: fortune-telling, overgeneralizing, selective abstraction, and personalization. Data on the validity of the method have been presented in several studies, demonstrating sufficient internal and external validity (see D'Iuso, Blake and Drapeau, 2007, 2009; Drapeau & Perry, 2005; Perry, Drapeau, Dunkley, Foley, Blake & Banon, 2007, for the original English version; for the French version used in this study, see Kramer & Drapeau, in press, and Kramer, Bodennmann & Drapeau, 2009).

Coping Action Patterns Rating Scale (CAPRS; Perry, Drapeau, Dunkley, & Blake, 2005). The CAPRS is an observer-rated system which can be used to assess coping processes based on interview-transcripts (Drapeau & Perry, 2005), the rating being event-based. The rating scale encompasses 12 categories of coping, based on the works of Skinner, Edge, Altman and Sherwood (2003). Three general domains are identified (relatedness, competence, autonomy), each one encompassing four categories of coping. Six of the coping categories are presumed to be coping with stress appraised as challenge (problem-solving, information-seeking, self-reliance, support-seeking, accommodation, negotiation) and the other six as coping with stress appraised as threat (helplessness, escape, delegation, isolation, submission, opposition). Each coping category is further broken down into three action levels (affective, behavioral and cognitive). Examples of ratings for several categories and a clinical application may be found in Kramer, Drapeau, Khazaal and Bodenmann (2009). In total, 36 coping processes are assessed by this instrument by counting up the coping action patterns per category. Data on the validation of the method have been presented by D'Iuso, Blake and Drapeau (2007; 2009), Drapeau and Perry (2005), Drapeau, Perry, Blake and D'Iuso (2007), Perry, Drapeau, Dunkley, Foley, Blake and Banon (2007) for the original English version, and by Kramer and Drapeau (in press), Kramer, Drapeau, Khazaal and Bodenmann (2009), Kramer, de Roten, Michel and Despland (2009) and by Kramer, Despland, Michel, Drapeau and de Roten (in press) for the French version used in this study.

Procedure

The training took place from October 2008 until april 2009, all 9 raters were trained in both (CERS and CAPRS) scales together, by the same trainer. The training procedure followed Perry and Henry (2004) and was structured into a basic training phase and a rater calibration phase. The basic training phase was based on ten 90-minutes training sessions, held once weekly, with one supervisor, in which nine selected transcripts (transcripts 1-9) as training stimulus were used. The selected transcripts were psychotherapy sessions rated previously by at least two trained raters and stemmed from the supervisor's data pool. The ratings were done individually by the participants between the group sessions. The calibration phase started six weeks after the end of the group sessions – a break due to vacation - comprised four transcripts (transcripts 10-13) per rater stemming from an ongoing research project and lasted three months in total. These transcripts were not previously rated ; they were rated independently by the supervisor and the trainee ; the training stimulus was therefore different for each trainee and each supervision session. Each rater received four individual 1-hour supervision sessions (60 minutes per transcript) with the supervisor. In these sessions, the supervisor provided structured feed-back based on the reliability check related to the specific transcript and chose a set of maximum five problematic categories to discuss with the rater. This focused procedure is in line with recommendations emitted by Tractenberg et al. (2001) and enabled to provide an individualized procedure based on the categories objectively identified as problematic; it has proven to be cost-effective. Finally, after this two-phase training, reliabilities should be sufficient and the participants may rate their own material. Regular supervision sessions were provided after the end of the training, in order to avoid rater drift in the ratings of the participants (not included in the present study).

Inter-rater reliability on the rating scales was established using the Intra-Class Correlation Coefficient, according to Shrout and Fleiss (1979). The model ICC (2, 1) was used (Wirtz, & Caspar, 2002); the reliability was computed for each rater and transcript separately, across all the categories of the two rating scales (15 cognitive errors and 12 coping categories). The Intra-Class Correlation Coefficients were computed for transcripts six through nine of the basic training phase, as well as for the remaining four cases of the calibration phase (transcripts 10-13). The reliability results of transcripts one through five were not taken into account for this study, as they represented the beginning of the rater training and thus, no specific evolution of the reliability was expected in this initial familiarization with the scales. Therefore, all transcripts from training stimulus 6 until 13 entered the study; no selection was made.

Data Analytic Strategy

Hierarchical Linear Modelling (HLM; Bryk, & Raudenbush, 1987; using MIXREG developed by Hedeker and Gibbons, 1996) was applied for the computation of change coefficients over the course of training. These models take optimally into account the dependency in the data and deal ideally with missing values. The following model was computed: Level 1: $\gamma_{ij} = \beta_{0j} + \beta_{1j} + \varepsilon$; Level 2: $\beta_{0j} = \gamma_{00} + \mu_{0j}$; $\beta_{1j} = \gamma_{10} + u_{1j}$. This computation was done for the training phase and the calibration phase separately. In order to respect the metric of weeks across the entire training, the statistical program treated training session 10 through 13 as sessions 16 through 19, due to the 6-weeks-break. Growth curve analysis including a quadratic term was implemented as post-hoc analysis (Kenny, Kashy, & Bolger, 1998). The slopes yielded from the HLM model were compared between the basic training phase and the rater calibration phase, by using Paired *t*-test statistics. In order to save space, we will only present the essential results for each model.

Results

In total, $N = 72$ transcripts were to be handed in for the study, both training phases taken together; $n = 10$ missing values were found, due to participants who did not hand in a rated transcript. Of these, four (6%) were in the basic training phase and six (8%) were in the rater calibration phase. Thus, the computation was done on $N = 62$ rated transcripts.

The graphic representation of raw scores of each evolution of the reliability of the whole process (basic training transcripts 6-9; rater calibration transcripts 10-13) is shown in figure 1. Table 1 reports the mean of the last and second-to-last session, which may account for the final result of the reliability of the scales after each training phase. They are satisfactory. Similar validation results were found on these instruments in a different rater group (Kramer & Drapeau, in press). In our sample of trainees, no difference was found between the four university psychology students and the trained psychologists.

Exploratory analyses yielded the following coefficients for the basic training phase, compared to the rater calibration phase: in the former, the overall mean of the four training sessions was ICC (2, 1) = .51 (SD = .09), whereas in the latter, the overall mean of the four calibration/supervision sessions was ICC (2, 1) = .62 (SD = .09). A Paired *t*-test comparison on the two training phases yielded a significant difference between training effects in favour of the calibration phase ($t(8) = -3.51$; $p = .01$).

For the linear model computed by HLM, an overall training effect was confirmed, taken both phases together, with a slope of .05 ($p = .00$; SE = .00; $Z = 5.14$). For the basic training phase, we found a significant slope of .18 ($p = .00$; SE = .00; $Z = 5.29$), whereas for the rater calibration phase, we found a signification slope of .09 ($p = .00$; SE = .02; $Z = 5.92$). We performed as post-hoc analysis a curvilinear growth model, adding to the linear model a quadratic term (Kenny, Kashy, & Bolger, 1998), for both training phases separately. For the basic training phase, we found a quadratic term of .00 ($p = .45$; SE = .65; $Z = .83$) and for the rater calibration phase .01 ($p = .32$; SE = .88; $Z = .62$), both non-significant. A Paired t-test between the slopes of the basic training phase and the rater calibration phase yielded a significant differences with regard to the slopes of the linear model ($t(8) = 8.01$; $p = .00$). No further analysis was conducted, in particular including a cubic term, as both training phases needed to be analyzed separately.

Discussion

The present study examined the effects of structured rater training on the reliability of ratings. Overall, as expected, the training is efficient, yielding sufficient mean reliability per sub-scale of the Cognitive Errors and Coping Action Patterns Rating Scales. The results reported indicate that the two-phase model in training is a helpful heuristic, as the basic rater training and the rater calibration phases yielded some differential evolutionary patterns in the reliability. First, as expected, the overall mean in the basic training was lower than in the calibration phase. Second, as expected, the training effect is higher in the basic training phase than in the rater calibration phase; a linear model is sufficient in order to describe this effect, the quadratic term being non-significant in both training phases. Finally, on a more qualitative descriptive level, we have found the following evolutionary pattern: whereas at the end of basic training, the participants' mean reliability was satisfying (ICC > .60), it drops at the first transcripts to be rated on their own (transcripts 10 and 11, figure 1). Later, at the end of the calibration phase, all participants perform reliable ratings. We hypothesize that the large decrease in ICC between the two training phases may be due to the absence of training during six weeks. The learning process may be described in terms of stages when the participants learn to accommodate to the clinical concepts of the CE-CAP rating scales: (1) initial grasp of the concepts, (2) fundamental questioning of the concepts when confronted to new clinical material, including a trial and errors strategy, (3) reconstruction of the know-how based on the previous experience of the initial grasp and the experience of fundamental questioning through trial and error. In our study (figure 1), stage (1) encompasses transcripts 6 through 9, stage (2) encompasses transcripts 10 and 11 and stage (3) represents transcripts 12 and 13.

We need to acknowledge the limitations of the present study. This is an observational study on the training process of specific rating scales. No direct generalizing should be made to any other rating scales in psychotherapy research, nor the other training groups using the same instruments. The number of participants is quite low, but we adapted our data analytic strategy to the ensuing power limitations. Finally, the degree of difficulty of the clinical material might influence the evolution of reliability; we did not take into account this variable.

Nevertheless, two tentative implications for rater training in psychotherapy research are to be mentioned. First, the relevance of the two-phase model of rater training (basic training, followed by rater calibration) suggests that after the initial training, the participants may be reliant on the training material, but their calibration when applying the rating scale on different material remains necessary, in order to guarantee satisfactory reliability for the specific research project. Second, shortenend and more focused training formats might be useful and feasible (Lorber, 2006; Tractenberg et al., 2001), also on University students (Schanche et al., 2008), if the number of training sessions is adapted to the complexity of the clinical phenomena to be rated and the complexity of the rating scales.

References

Bryk, A. S., & Raudenbush, S. W. (1987). Application of Hierarchical Linear Models to assessing Change. *Psychological Bulletin, 101*, 147-158.

Cicchetti, D., Bronen, R., Spencer, S., Haut, S., Berg, A., Oliver, P., & Tyrer, P. (2006). Rating Scales, Scales of Measurement, Issues of Reliability. Resolving some Critical Issues for Clinicians and Researchers. *The Journal of Nervous and Mental Disease, 194*(8), 557-564.

D'Iuso, D., Blake, E., & Drapeau, M. (2007). *An Investigation of the Relationship of Cognitive Errors and Coping Action Patterns with the Therapeutic Alliance*. Canadian Psychological Association Annual Convention. Ottawa, June.

D'Iuso, D., Blake, E., & Drapeau, M. (2009). Cognitive Errors, Coping Patterns, and the Therapeutic Alliance: A Pilot Study of In-Session Process. *Counselling and Psychotherapy Research, 9*(2), 108-114.

Drapeau, M., & Perry, J. C. (2005). *Observer-Ratings of Coping and Cognitive Errors : The Cognitive Errors and Coping Action Patterns (CECAP) Method*. Paper presented at the 36th Annual Meeting of the Society for Psychotherapy Research. Montréal: June.

Drapeau, M., Perry, J. C., Blake, E., & D'Iuso, D. (2007). *A pilot study of the validity and reliability of the CECAP method*. Ottawa: Canadian Psychological Association Annual Convention.

Drapeau, M., Perry, J. C., & Dunkley, D. (2008). *The Cognitive Errors Rating System*. Montréal: McGill University.

Hedeker, D., & Gibbons, R. D. (1996). MIXREG: A Computer Program for Mixed-Effects Regression Analysis with Autocorrelated Errors. *Computer Methods and Programs in Biomedecine, 49*, 229-252.

Ivancevich, J. M. (1979). Longitudinal Study of the Effects of Rater Training on Psychometric Error in Ratings. *Journal of Applied Psychology, 64*(5), 502-508.

Kenny, D. A., Kashy, D. £A., & Bolger, N. (1998). Data Analysis in Social Psychology. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *The Handbook of Social Psychology, 4th Edition* (pp. 233-265). Boston: McGraw-Hill.

Kramer, U., de Roten, Y., Michel, L., & Despland, J.-N. (2009). Early Change of Defense Mechanisms and Coping in Short-Term Dynamic Psychotherapy: Relations with Symptoms and Alliance. *Clinical Psychology and Psychotherapy, 16*, 408-417.

Kramer, U., Despland, J.-N., Michel, L., Drapeau, M., & de Roten, Y. (in press). Change in Defense Mechanisms and Coping over the Course of Short-Term Dynamic Psychotherapy for Adjustment Disorder. *Journal of Clinical Psychology.*

Kramer, U., & Drapeau, M. (in press). Etude de validation de la version française des échelles de codage du coping et des erreurs cognitives (CE-CAP) sur une population non-clinique. *Annales Médico-Psychologiques.*

Kramer, U., & Drapeau, M. (2009). Cognitive Errors and Coping Action Patterns in Child Molesters as assessed by external observers. *The Open Criminology Journal, 2*, 24-28.

Kramer, U., Bodenmann, G., & Drapeau, M (2009). Cognitive Errors Assessed by Observer-Ratings in Bipolar Affective Disorder, Symptoms and Therapeutic Alliance. The *Cognitive Behaviour Therapist, 2*(2), 92-105.

Kramer, U., Drapeau, M., Khazaal, Y., & Bodenmann, G. (2009). Coping Specificities in Bipolar Affective Disorder: Relation with Symptoms, Symptom Evolution and Therapeutic Alliance. *German Journal of Psychiatry, 12*(1), 19-27.

Lefebvre, M. F. (1981). Cognitive Distortion and Cognitive Errors in Depressed Psychiatric and Low Back Pain Patients. *Journal of Consulting and Clinical Psychology, 49*, 517-525.

Lorber, M. F. (2006). Can Minimally Trained Observers Provide Valid Global Ratings? *Journal of Family Psychology, 20*(2), 335-338.

Luborsky, L., & Crits-Christoph, P. (1998). *Understanding Transference: The Conflictual Relationship Theme Method (2nd edition).* Washington: D. C.: American Psychological Association Books.

Perry, J. C., Drapeau, M., Dunkley, D., & Blake, E. (2005). *The Coping Action Patterns Rating System*. Montréal: McGill University.

Perry, J. C., Drapeau, M., Dunkley, D., Foley, E., Blake, E., & Banon, E. (2007). *Changes in Cognitive Errors and Coping Action Patterns Following 20 Sessions of Either Dynamic Psychotherapy or CBT of Acute Recurrent Depression.* Paper presented at the 38th Annual Meeting of the Society for Psychotherapy Research. Madison, USA: June.

Perry, J. C., & Henry, M. (2004). Studying Defense Mechanisms in Psychotherapy using the Defense Mechanisms Rating Scales. In U. Hentschel, G. Smith, J. G. Draguns, & W. Ehlers (Eds.), *Defense Mechanisms. Theoretical, Research and Clinical Perspectives* (pp. 165-192). Amsterdam: Elsevier.

Schanche, E., Hostmark Nielson, G., McCullough, L., Mykletun, A., & Valen, J. (2008). *The Effects of Varying Amounts of Training on the Reliability of Graduate Students' Rating of Complex Clinical Phenomena*. Paper presented at the 39th Annual Meeting of the Society for Psychotherapy Research. Barcelona: June 18-21.

Shrout, P. E., & Fleiss, J. L. (1979). Intra-Class Correlations: Uses in Assessing Rater Reliability. *Psychological Bulletin, 86*(2), 420-428.

Skinner, E. A., Edge, K., Altman, J., & Sherwood, H. (2003). Searching for the Structure of Coping: A Review and Critique of Category Systems for Classifying Ways of Coping. *Psychological Bulletin, 129*, 216-269.

Tractenberg, R. E., Schafer, K., & Morris, J. C. (2001). Interobserver Disagreements on Clinical Dementia Rating Assessment: Interpretation and Implications for Training. *Alzheimer Disease and Associated Disorders*, *15*(3), 155-161.

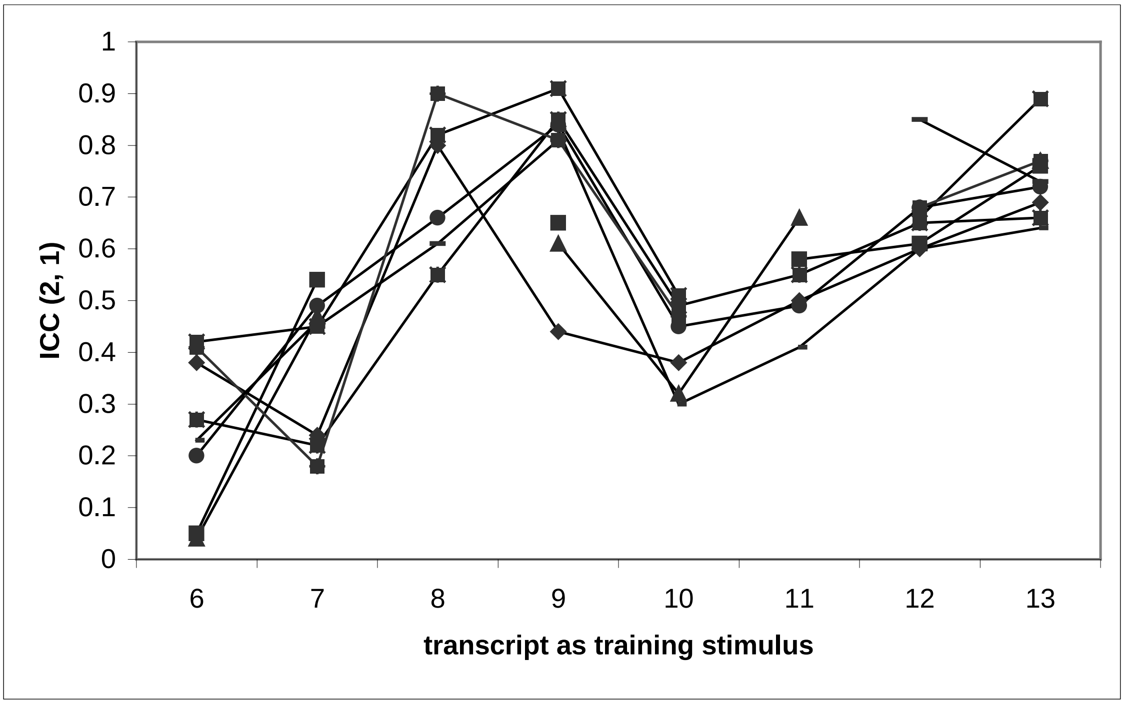
Wirtz, M., & Caspar, F. (2002). *Beurteilerübereinstimmung und Beurteilerreliabilität*. Göttingen: Hogrefe.

Table 1

Mean reliability at the end of rater training and calibration phases

Category	ICC (2, 1) training	ICC (2, 1) calibration
CAPs		
Problem-Solving	.77	.71
Information-Seeking	.76	.81
Helplessness	.61	.87
Escape	.65	.67
Self-Reliance	.89	.88
Support-Seeking	.79	.85
Delegation	.76	.71
Isolation	.66	.74
Accommodation	.88	.86
Negotiation	.65	.78
Submission	.70	.71
Opposition	.71	.87
CEs		
Positive Errors		
Overgeneralization	.61	.71
Selective Abstraction	.67	.66
Personalization	.71	.61
Negative Errors		
Overgeneralization	.88	.81
Selective Abstraction	.82	.86
Personalization	.67	.69

Figure 1



Note. Transcripts 6-9: basic training phase (identical pre-selected transcripts); transcripts 10-13: rater calibration phase (different non-selected transcripts). ICC (2, 1) > .60 considered as sufficient.

Figure Caption

Figure 1. Evolution of reliability coefficients during CE-CAP basic training and calibration phases