Interrater Reliability Study
CCAR 2006
Colorado Division of Mental Health

A Report from the Colorado Department of Human Services
Division of Mental Health

DMH

cdhs
Colorado Department of Human Services
people who help people

Prepared by the Division of Mental Health, Data and Evaluation Section

In June 2005 the Division of Mental health (DMH) conducted a study to assess the interrater reliability of the CCAR 2006. Testing the psychometric properties of the CCAR 2006 was necessary because it is a substantive redesign of the existing instrument. The revisions were based on previous studies of validity and reliability, focus groups, and user feedback. The interrater reliability study of the previous CCAR instrument utilized highly trained observers who rated multiple videotaped case studies. The present study created a more representative trial of the instrument by employing a large number of experienced CCAR users to rate a small number of videotaped client interviews. This approach yields a representative picture of the tool's reliability in the field.

Method

Participants

A total of 54 raters participated in the study. Of those 23 individuals viewed and rated child case studies and 31 rated adult case studies. Two of the participants had associate degrees, 20 had bachelor's degrees, 27 had master's degrees, three had doctoral degrees, and two raters did not report their educational level. There were eight representatives from Behavioral Health Organizations (BHOs), eleven from Community Mental Health Centers (CMHCs), nine from the Division of Youth Corrections (DYC), six from Residential Treatment Centers (RTCs), seven from specialty mental health clinics, five from state mental hospitals, and six who identified themselves as from "other" types of agencies. All raters had previously completed at least five CCARs.

Procedure

Experienced CCAR users from all systems that use the CCAR (DYC, CW, and MH) were invited to participate in the interrater reliability study. All participants attended a 45-minute training session to review all items on the CCAR and to discuss changes from the old instrument. Following this training, participants viewed videotaped intake interviews. Participants who identified themselves as primarily working with youth watched three videotapes of adolescents' initial interviews with a clinician. Participants who identified themselves as primarily working with adults watched four videotapes of adults' initial interviews with a clinician. These particular videotapes were selected so that every CCAR domain would be elevated in at least one case. This was true for all domains except for psychoses in youth. A case with demonstrable psychosis for a child or adolescent could not be identified. Of the three youth focused videotapes one was an actual intake interview with a CMHC client who agreed to have his intake interview released for training purposes. The videotapes of adults included two actual clients and two actors portraying clients.

At the conclusion of each videotape, participants were asked to score a CCAR according to the information provided in the video. Scores from each CCAR were analyzed and compared to the mean for the overall group. The percent of raters whose scores were within one, two, three, four or five or more points of the mean were calculated.

Results and Discussion

Percent agreement was used to estimate interrater reliability. A commonly used index of consistency, percent agreement indicates the extent to which raters assign identical or approximate scores. This method is particularly advantageous when the number of scale points is small and/or when scale categories are coarsely grouped, conditions that both limit the meaningfulness of correlation coefficients (Brown, Lucero, & Foss, 1962). Additionally, examination of percent agreement helps to identify areas of disagreement that can then be focused upon in training (Goodwin, 2001).

Results represent the averages between child and adult CCAR assessments. Different levels of agreement were conceptualized based upon a score's distance from the mean of all scores. For instance, a score was classified as being in the first and highest category of agreement if it was within one point in either direction of the average. This method was believed to best capture the score ranges encompassed by each of the five categories (or anchors) for the nine-point scales. Thus, agreement levels were thought to reflect the variability of scores in much the same way that a confidence interval represents the range into which a true score falls.

Across all domains, 70.8% of raters assigned ratings within one point of the mean, and 87% of raters assigned ratings within two points of the mean. Thus, close to three quarters of all ratings were in the highest level of agreement, represented by the conservative range of one point in either direction of the mean. Because five defined anchors (e.g. 3 = "Occasional," 5 = "Frequent") comprise each of the nine-point scales, the second level of agreement (two points from the mean) likely reflects a difference of one categorical ranking (e.g. occasional vs. frequent legal difficulties). Thus, across all domains, 87% of raters seem to have assigned ratings that were either in the same or the most approximate category.

Table 1 presents the percent of ratings within each group (based on distance from the mean) across all domains.

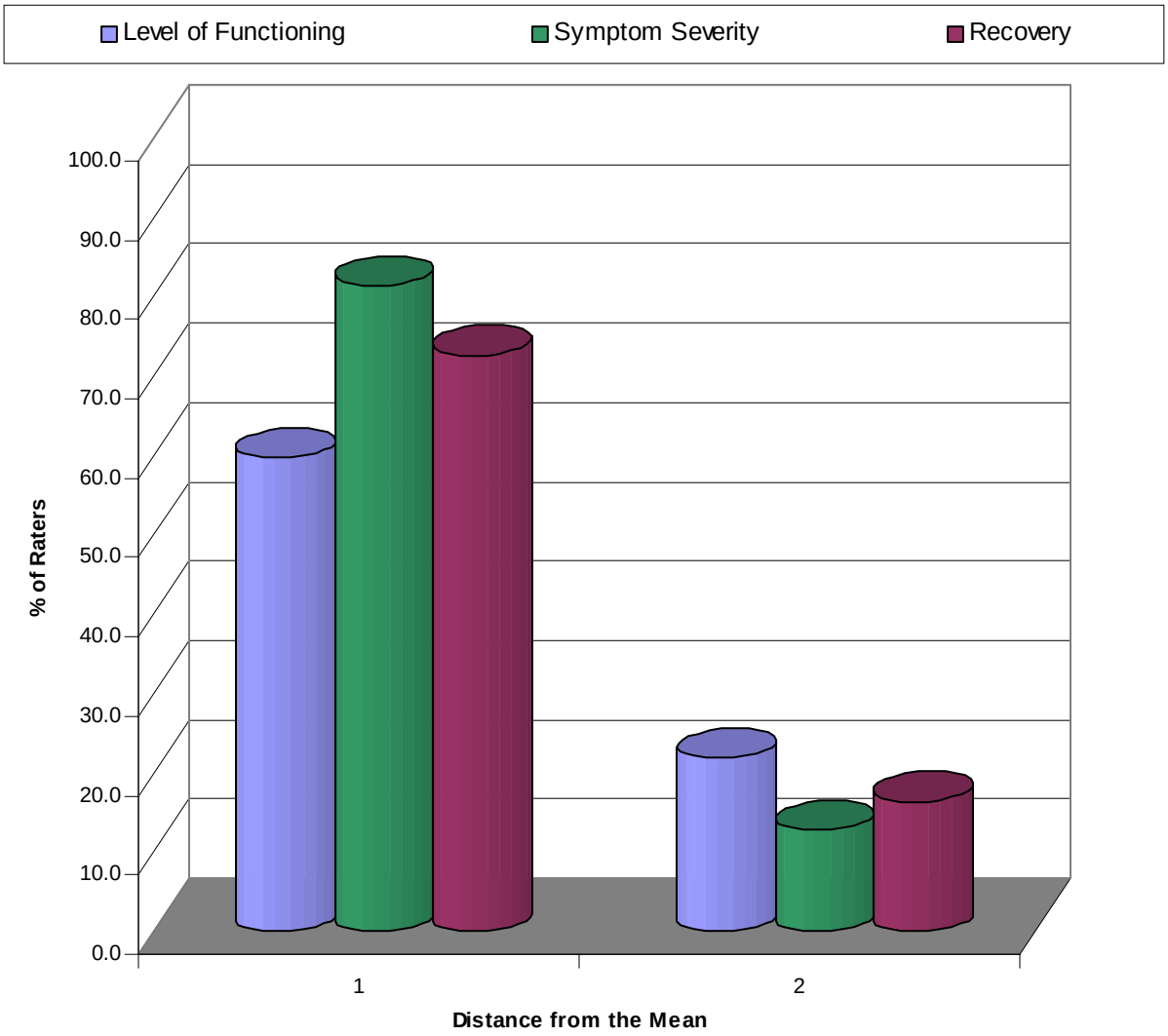
Table 1
Percent of Ratings by Distance from the Mean across all Domains

Domains	Percent of Raters within Specified Number of Points from the Mean			
	1 pt.	2 pts.	3 pts.	4 pts.
Psychosis	92.5	97.7	98.9	99.7
Drug Use	86.6	93.4	96.5	98.5
Physical Health	86.3	98.1	98.5	98.9
Suicide/Danger to Self	84.2	93.9	98.9	98.9
Socialization	84.2	96.2	98.2	99.4
Overall Symptom Severity	81.4	94.2	97.8	98.6
Legal	81.1	91.7	96.3	99.3
Depressive Issues	79.8	94.5	99.2	99.6
Self Care/Basic Needs	76.3	92.4	97.8	99.4
Security Supervision	75.7	91.3	96.9	97.7
Manic Issues	75.0	90.8	97.3	99.3
Cognition	74.5	88.4	97.9	99.1
Role Performance	72.9	91.4	97.7	98.9
Overall Recovery	72.6	89	97.7	98.5
Family	72.3	90.7	94.9	99.2
Aggression/Danger to Others	71.4	88.7	95.5	98.2
Empowerment	70.7	86.6	95.7	99.3
Attention	66.3	90.4	98.2	100
Hope	65.1	89.8	94.7	97.4
Alcohol Use	64.3	88.8	95.9	99
Interpersonal	63.9	89.5	97.3	98.9
Anxiety Issues	63.4	90.8	96.6	98.5
Activity Involvement	62.8	83.4	93.3	97.2
Overall Level of Functioning	59.7	81.6	93.8	96.2
Social Support	56.4	76.9	91.9	97.1

Figure 1 displays percent agreement for the first two levels of agreement for Level of Functioning, Symptom Severity, and Overall Recovery. These are of particular interest, as they represent the most global domains of the CCAR 2006. As shown, 81.4% of raters assigned scores within one point of the mean for Symptom Severity, indicating a strong level of agreement. Although somewhat lower, the 72.6% within this range for Overall Recovery is similar to the agreement level (70.8%) seen across all domains. In contrast, among the 25 domains, Overall Level of Functioning has the second lowest level of

agreement for the one-point range (59.7%). This is possibly due to its having been significantly redefined since the previous CCAR (see below).

Figure 1. Inter-rater agreement across selected domains.



In general, results indicate lower interrater reliability for new and significantly revised domains compared to those maintained with little change from the CCAR 2000. For instance, only two of the 25 domains, Social Support and Overall Level of Functioning, had less than 60% of their ratings within one point of the mean. Both domains represent revisions from the previous CCAR. Social Support is a new item, and the current rating criteria for Level of Functioning substantially departs from that in the CCAR 2000. Moreover, for the five Recovery domains added to the current CCAR (i.e. Social Support, Empowerment, Hope, Activity Involvement, and Overall Recovery), the average number of ratings within one point of the mean is 65.5% versus 75.6% for the average of those domains also included in the previous CCAR. Given that all raters had experience with the CCAR 2000, lower levels of agreement for new and altered items are not

surprising. Reliability for these items seems likely to improve with training specific to the CCAR 2006.

A statistically significant inverse relation appeared between average ratings and percent agreement within one point of the mean (r = -0.44, p < .05). That is, raters in general agreed more when client problems were ostensibly small or non-existent and, conversely, they showed greater variability as problems visibly increased in severity (see Table 2). This trend was somewhat expected, as it is sometimes seen in well-validated psychometric measures, such as the Clinically Administered PTSD Scale. Nonetheless, the relation suggests that increased training emphasis on recognizing and operationally defining problems will yield greater interrater reliability, particularly among more severely disturbed clients.

Table 2
Relation of Domain Average to Percent Agreement

Domain	Mean Rating	% Raters w/in 1 pt. of mean
Manic Issues	1.76	75.00
Self Care/Basic Needs	1.95	76.30
Psychosis	1.96	92.50
Drug Use	2.59	86.60
Aggression/Danger to Others	2.61	71.40
Physical Health	2.63	86.30
Suicide/Danger to Self	2.66	84.20
Cognition	2.73	74.50
Attention	3.13	66.30
Security Supervision	3.15	75.70
Alcohol Use	3.25	64.30
Suicide/Danger to Self	3.38	84.20
Anxiety Issues	3.47	63.40
Legal	3.66	81.10
Interpersonal	4.04	63.90
Family	4.11	72.30
Empowerment	4.29	70.70
Social Support	4.31	56.40
Overall Level of Functioning	4.77	59.70
Depressive Issues	4.91	79.80
Role Performance	5.16	72.90
Overall Symptom Severity	5.19	81.40
Hope	5.52	65.10
Activity Involvement	5.58	62.80
Overall Recovery	5.92	72.60

In summary, almost 90% of raters' scores were within two points of the domain mean, reflecting ratings that denoted the same or closely approximated clinical impressions. Examination of percent agreement indicated that new and significantly revised domains generally had poorer interrater agreement than those that remained unchanged from the CCAR 2000. For instance, Overall Level of Functioning had the second poorest level of agreement among the 25 domains, which seems likely due to it having been substantially redefined since the previous CCAR. Given raters' experience with the CCAR 2000, proactive interference may have contributed to the lower interrater agreement on this domain; reliability may therefore improve as raters gain experience with its current conceptualization. Regarding the additions to the CCAR 2006, raters had lower proportions of agreement on Recovery domains than on non-Recovery domains. This finding was not surprising, given that Recovery represents a conceptual change from the exclusively problem-oriented domains of the previous CCAR. Additionally, the lower reliability for Recovery domains may also reflect the relative newness in the mental health field of focusing on recovery concepts and principles. Increased reliability for the CCAR 2006 seems likely to result from trainings that focus on comprehension of domain and category definitions, particularly those belonging to Recovery domains.