COLORADO CLIENT ASSESSMENT RECORD
INTERRATER RELIABILITY STUDY

FINAL REPORT

Deborah B. Altschul, Ph.D.
Jack Wackwitz, Ph.D.
Anita Saranga Coen, LCSW
Dick Ellis, Ph.D.

Colorado Mental Health Services

BACKGROUND

The Colorado Client Assessment Record (CCAR) is a clinical instrument designed to provide a standardized method of documenting clinical impressions of cognitive and behavioral functioning, symptoms, and strengths and resources at various points of service delivery. The CCAR was designed for completion by clinicians with broad levels of training and/or experience.

The CCAR has been used by Colorado Mental Health Services (MHS) for over 20 years as the primary source of data collection documenting clinical evaluations of individuals served by the public mental health system. Recently, Child Welfare has adopted the tool for RTC (Residential Treatment Center) placements and the Division of Youth Corrections has adopted the tool for RTC placements, detention, and special studies.

Due to the CCAR's widespread current and planned use as a measure of outcome, and cost effectiveness, the Colorado Legislature requested and funded an instrument reliability study to be conducted by Mental Health Services. Subsequently, a series of studies were designed to examine the reliability of the CCAR. Study 1 was designed to determine the interrater reliability of the CCAR between raters who were trained extensively on CCAR scoring and completion. The goal of this study was to assess the instruments' reliability across raters in a scientifically controlled environment. Study 2 was designed to determine the test-retest reliability within this same rater group to determine whether individuals in a controlled environment scored the CCAR consistently over time. Finally, Study 3 was designed to assess the interrater reliability of the CCAR across two different rater groups. One group was comprised of the Study-Raters (those participating in Studies 1 and 2), and the other was Agency-Raters. The Agency-Rater group consisted of individuals working in Community Mental Health Centers who complete CCARs routinely as part of their job description and participated in the standard MHS sponsored CCAR training. The goal of Study 3 was to estimate the reliability of the instrument in a natural setting.

The following report is intended to describe the design, implementation, and results of Study 1: The Study-Rater CCAR Interrater Reliability Study. Study 1 was selected for completion first because it is a study of instrument effectiveness, and extraneous variables have been subtracted from the equation by ensuring that all Study-Raters have identical training in CCAR administration and scoring. In addition, all Study-Raters completed CCARs under identical conditions. The goal of the study was to determine the reliability of the CCAR in a controlled setting and then move to Study 3 were CCAR scores of Study-Raters can be compared with the CCAR scores of clinicians working in public mental health. The hope of this multi-tiered investigation is to determine the overall reliability of the CCAR in order to assess its' usefulness and generalizability for Colorado's public mental health system.

Rationale for Study 1

Over the past 20 years there have been several studies conducted evaluating the reliability of the CCAR's primary scales. Generally, these studies have focused on **internal consistency**, which is a measure of how clear items are, how well raters understand response options, and how items in a scale work together. Internal consistency is measured by a reliability coefficient (a number between 0 and 1), with desirable coefficients ranging from .60 to .80. Previous investigation determined that all CCAR reliability coefficients lie within this acceptable range (Internal MHS CCAR Studies).

Although internal consistency studies provide useful information, it is equally important to assess other types of reliability. **Interrater reliability**, for instance, is a statistical means of determining the level of agreement between two independent raters who are evaluating the same clinical information with an identical assessment instrument. Interrater reliability is measured by a correlation coefficient, which compares pairs of ratings made by different people. Correlation coefficients range from -1 to 1, with 1=perfect agreement, 0= no agreement, and –1=perfect disagreement. Acceptable ranges for correlation coefficients vary, depending on the type of information being collected. Reliability coefficients for behavioral observations, for instance, are generally expected to range between .6-.8 (Fleenor, Fleenor & Grossnickle, 1996)[1], while reliability coefficients for clinical judgments are expected to range between .5-.8 (Dow & Ward, 1996)[2]. Because the CCAR collects clinical judgment data, the target range for the current study was set at .5-.8. This range provides a useful method of determining whether items are likely to be endorsed consistently across raters.

In the case of public mental health, it is important that clinical evaluations are conducted consistently, regardless of the mental health setting or the clinician conducting the evaluation. Consistent evaluation permits the monitoring of an individual's therapeutic process. It also provides a framework for mental health practitioners for conducting overall assessments of client functioning. Finally, the collection of CCAR data is critical for State,

[1] Fleenor, J.W., Fleenor, J.B., & Grossnickle, W.F. (1996). Interrater reliability and agreement of performance ratings: A methodological comparison. Journal of Business & Psychology, Vol 10(3), 367-380.

[2] Dow, M.G. & Ward, J.C. (1996). Program evaluation and outcome assessment project: HRS District 7 phase three summary, fiscal year 1995-1996. Tampa, FL: Department of Community Mental Health, Florida Mental Health Institute.

Local, and Federal agencies for identifying statewide and regional trends that require unique policy development and service implementation. The analysis of statewide CCAR data allows for conclusions to be drawn about specific populations and policies to be formed based on data findings.

In recent years, the CCAR has undergone considerable change and has been adopted for use in several states. Its use in Colorado has spread to other service systems, and a standardized, accessible training program has also been implemented. The purpose of the current study was to determine the interrater reliability of the CCAR and to identify potential areas warranting improvement. Thus, the findings from the CCAR Interrater Reliability Study will provide critical information for informing both policy and practice.

STUDY 1 RESEARCH DESIGN

Instrumentation

The CCAR is comprised of three sections: Problem Severity, Strengths and Resources, and Level of Functioning. The Problem Severity domain is the first section and documents clinical rating in 21 behavioral, clinical, and cognitive areas. Each item included in the Problem Severity domain consists of a scale (rated 1-9, with higher numbers indicating greater severity) and a set of accompanying checklist items that are endorsed if the problem is present. The Problem Severity section also includes an Overall Problem Severity scale and a Change in Overall Problem Severity scale.

Section two of the CCAR documents Strengths and Resources in 8 areas and follows the same format as section one, with each item consisting of a Strengths and Resources scale (rated 1-9, with higher numbers indicating lower Strengths and Resources) and a set of accompanying checklist items. The Strengths and Resources section also includes an Overall Strengths/Resources scale and a Change in Overall Strengths/Resources scale.

The third and final section of the CCAR consists of Level-Of-Functioning (LOF) scales in 6 areas (rated 1-9, with higher numbers indicating lower functioning). In addition, the section includes an Overall Level of Functioning Scale and a Change in Overall Level of Functioning Scale.

Study-Raters

Study-Raters were clinicians working outside of Colorado's public mental health system who were hired specifically for the study. There were two pairs of Study-Raters (four individuals), with one pair focusing on adult consumers, and the other pair focusing on youth consumers.

Interviewees (Adult/Youth Consumers)

Interviewees for the current study included 143 individuals who accessed mental health services at one of two community mental health centers over the course of a six-month period. Forty-seven participants were aged 4-17, and 96 participants were 18 or over. Participation was limited to individuals who were at either the admission or update point of

service delivery and, therefore, routinely required an assessment and subsequent CCAR completion by a mental health center staff member.

Procedure

Video Observations: One hundred and forty three observations of actual interviews with 47 youth, and 96 adults were video-recorded at the two CMHCs. Prior to beginning the interview, adult and youth consumer participants were told of the study and asked whether they agreed to participate. Youth participation required additional consent from a parent or guardian. All individuals who agreed to participate in the study signed a written consent form, and then participated in a routine assessment that was videotaped by the clinician. Consumers who agreed to participate in the study received monetary incentives in appreciation of their participation. CMHC staff received incentives in the form of a gift certificate.

The videotaped interviews were conducted by Agency-Raters who participated in the 4-hour standardized CCAR training session provided by Mental Health Services. After completion of the training session, Agency-Raters solicited participation in the study from consumers who accessed mental health services at their agencies. Agency-Raters then videotaped clinical interviews with consumers who agreed to participate, and completed a CCAR following each videotaped interview. Finally, Agency-Raters sent a copy of the CCAR, along with the original videotaped interview, to a designated staff member at Mental Health Services, who then stored all videotapes in a locked cabinet.

Study-Raters: Over a three-week period, the four Study-Raters were trained in CCAR scoring and video observation. All Study-Raters met simultaneously with the training instructor for three hours per day, three days per week, for a total of 27 hours. CCAR training consisted of an introduction to the Revised CCAR 2000 through general overview and practical applications, followed by hands-on exposure to the instrument via videotaped case studies and written case vignettes. The training also included video observation training, which consisted of exposure to methods and techniques regarding video observation and coding. The goal was to train Study-Raters until they reached agreement on definitions, such that their scores on identical cases were consistent across Raters. This method is called **criterion training**, and is necessary for estimating the highest possible interrater reliability.

After completion of the training, Study-Raters viewed videotapes independently and scored CCARs. Distinct pairs viewed youth and adult observations, such that both members of the Adult Focused Study-Rater pair viewed all 96 adult videotapes, and both members of the Youth Focused Study-Rater pair viewed all 47 youth videotapes. Throughout the course of the study, all videotapes remained in a locked cabinet at Mental Health Services. Study-Raters viewed videotapes in a private room designated specifically for this purpose, and were paid an hourly wage for time-spent training, viewing videotapes, and scoring CCARs.

DATA ANALYSIS AND FINDINGS

CCAR interrater reliability was assessed by comparing Study-Rater scores within the Adult Focused Study-Rater pair and Study-Rater scores within the Youth-Focused Study-Rater

pair. Pearson Product Moment Correlations were used to estimate the correlation between ratings within the Adult Study-Rater pair and the Youth Study-pair. The following is a description of the data analysis procedure and findings.

Data Analysis

Due to the number of items and domains that comprise the CCAR, 105 different assessments of reliability were possible. The task of assessing interrater reliability, therefore, first involved the summarization of this information into a smaller number of coefficients. This process allowed for the generation of overall reliability coefficients for the CCAR as a whole, as well as overall reliability coefficients for each of the three domains: Problem Severity, Strengths and Resources, and Level of Functioning. This was followed by interrater analysis of each scale within each domain.

Three Pearson Product Moment Correlations were computed comparing scores across the two Adult Focused Study-Raters, and the two Youth Focused Study-Raters for each of the Problem Severity, Strengths and Resources, and Level of Functioning scales. This statistical process provided a single correlation coefficient for each CCAR item (e.g., emotional withdrawal, and depression) within the Problem Severity, Strengths and Resources, and Level of Functioning domains. These correlation coefficients were then averaged to create correlation coefficients for the CCAR overall, and for each domain (problem severity, strengths and resources, and level of functioning).[3] The results of this analysis are presented in Table 1.

Table 1: Overall Interrater Reliability Correlation Coefficients

	Adult Focused Study-Raters	Youth Focused Study-Raters
Problem Severity	.644*	.592*
Strengths and Resources	.457	.517*
Level of Functioning	.538*	.591*
Overall CCAR Instrument	.588*	.578*

* Correlations falling within the expected range (.5 -.8)

Using the expected range for interrater reliability (.5 - 8) discussed earlier, it can be seen that 7 of the 8 summary correlation coefficients fell within this range. It is important to note that the correlation that fell below this range is the Strengths and Resources domain for adult consumers. Notably, this correlation is approaching the .5 lower bound of the identified range.

[3] As correlation coefficients cannot be directly averaged, the process of averaging requires three steps. Correlations are first transformed into z-scores, Z-scores are then averaged, and averaged z-scores are then transformed back into correlation coefficients.

An improvement on this assessment of reasonableness for interrater reliability involved creating a 95% confidence interval around a parameter estimate, and assessing each correlation against this standard. A 95% confidence interval is an interval constructed from a sample so that there is a 95% chance that the interval contains the parameter. With repeated sampling, 95% of intervals constructed would include the population parameter. Therefore, a confidence interval provides an estimate of how accurately the population parameter has been approximated. If a correlation falls above the lower bound then it is considered "not statistically different" from the parameter estimate. In contrast, if it falls below the lower bound then it is considered "significantly different" and warrants further investigation.

For this analysis, the parameter of .65 was chosen for the following three reasons: (1) It is the mid-point of the .5 to .8 range for expected inter-rater reliability; (2) It represents a reasonable goal for the desired reliability for the CCAR; and, (3) It is approximately equal to the largest average correlation in Table 1, thus providing a high standard of comparison. Table 2 presents the 95% confidence intervals for the Adult-Focused-Study-Raters and the Youth-Focused-Study-Raters.

Table 2: 95% Confidence Intervals Using a .65 Parameter Estimate

	Lower Limit	**Upper Limit**
Adult	.515	.757
Youth	.446	.789

The following is a description of findings based on the examination of the correlation coefficients generated in Table 1 with regard to where they fell around the bounds of the confidence intervals provided in Table 2. Findings are reported for each of the rater pairs (Adult Focused and Youth Focused).

Adult Focused Study-Raters Interrater Reliability

Using .65 as the parameter estimate, it was determined that the 95% confidence interval for the Adult Focused Study-Rater CCAR correlation coefficients was **.515 to .757** (see Table 2). As mentioned earlier, this confidence interval signifies that all correlation coefficients falling within these bounds are not statistically different from a parameter estimate of .65, and all items falling below the lower limit of the confidence interval are statistically different. (See Table 2)

Overall Correlations: The correlation coefficient of the Adult Focused Study-Raters on the Overall CCAR Instrument was **.588**, which falls within the 95% confidence interval for Adult Focused Study-Raters. The overall correlation coefficient for the Problem Severity domain was **.644**, which also falls within the 95% confidence interval for Adult Focused Study-Raters. The overall correlation coefficient for the Strengths and Resources domain was **.457**, which falls below the lower limit of the 95% confidence interval for Adult Focused Study-Raters. And, the overall correlation coefficient for the Level of Functioning domain was **.538**, which falls within the 95% confidence interval for Adult Focused Study-Raters. Thus, three of

the four correlation coefficients fell within the 95% confidence interval using a parameter estimate of .65. (See Tables 1 and 2)

These findings support the notion that the CCAR appears to be a reliable instrument for adult consumers, particularly regarding Problem Severity and Level of Functioning. The Strengths and Resources domain, however appears to be less reliable for evaluating adult consumers, and warrants further investigation.

Item-Specific Correlations: For the 21 Problem items, correlation coefficients for Adult Focused Study-Raters ranged from .318 to .826, with 15 items falling within the 95% confidence interval, and 6 items falling below the lower bounds of the confidence interval. Correlation coefficients for the Strengths and Resources domain ranged from .161 to .803, with two items falling within the 95% confidence interval, and six items falling below the lower bounds of the confidence interval. Similarly, Level of Functioning correlation coefficients ranged from .351 to .795, with 2 items falling within the 95% confidence interval, and four items falling below the confidence interval. (See Table 3)

This focus on item-specific correlations suggests differential reliability across CCAR domains. While the majority of the Problem Severity items met acceptable standards, this was not the case with the Strengths and Resources and Level of Functioning Domains. In order to gain a greater understanding of the usefulness of specific CCAR items, it is important to examine items falling below the lower limit of the confidence interval.

Within the Problem Severity domain, for instance, emotional withdrawal, resistiveness, socialization issues, role, security/management issues, and overall Problem Severity all fell below the lower limit of the confidence interval. Further item-level analyses indicate that Resistiveness, Socialization, and Role are all multifactorial variables that lack convergent validity. This suggests that each item contains more than one construct, and therefore may result in lowering of Interrater reliability. Further examination of the checklist items may help with identifying items that may contribute to the low reliability.

Additionally, analysis of the Security/Management Issues variable indicates that 90 of the 96 adult consumers were given a rating of 1 (indicating “no problem”). This finding suggests a restriction of range issue, which inherently results in lowering and destabilizing correlations. Further investigation of this construct may help to understand the reasons underlying the restriction of range phenomenon.

Within the Strengths and Resources domain, Basic Resources, Education, Support, Personal Judgment, Personal Strengths, Overall Strengths, all fell below the lower limits of the 95% confidence interval. More in-depth analyses of the Strengths and Resources items also suggest issues related to lack of convergent validity and restriction of range. The Personal Strengths item, for instance, contains checklist items representing multiple constructs. Thus, further research and development of the Strengths and Resources section is in order.

Similarly, within the Level of Functioning domain, Interpersonal, Daily Living, Cognitive/Intellectual, and Overall Level of Functioning also fell below the lower limit.

Because Level of Functioning scales do not contain checklist items, it is difficult to determine their divergent and concurrent validity. Further exploration of these items is clearly indicated.

Youth Focused Study-Raters Interrater Reliability

Using .65 as the parameter estimate, it was determined that the 95% confidence interval for Youth Focused Study-Rater CCAR correlation coefficients was **.446 to .789** (see Table 1). Thus, all correlation coefficients falling below the lower bound of this confidence interval are considered statistically different than the parameter estimate and warrant further investigation.

Overall Correlations: The overall correlation coefficient for Youth Focused Study-Raters on the Overall CCAR Instrument was **.578**, which falls within the 95% confidence interval for Youth Focused Study-Raters. The overall correlation coefficient for the Problem Severity domain was **.592**, which also falls within the 95% confidence interval for Youth Focused Study-Raters. The overall correlation coefficient for the Strengths and Resources domain was **.517**, which falls within the 95% confidence interval for Youth Focused Study-Raters, and the overall correlation coefficient for the Level of Functioning domain was **.591**, which also falls within the 95% confidence interval for Youth Focused Study-Raters. Thus, all four of the Youth Focused Study-Rater correlation coefficients fell within the 95% confidence interval using a parameter estimate of .65. (See Table 1)

Given that the Overall CCAR Instrument correlation, and the three domain-specific correlations (Problem Severity, Strengths and Resources, and Level of Functioning) all fell within the 95% confidence interval, it appears that the CCAR is a reliable instrument across all four domains. Thus, these findings suggest that the CCAR is a reliable clinical evaluation instrument for youth involved in the public mental health system.

Item Specific Correlations: Problem Severity rating correlation coefficients for Youth Focused Study-Raters ranged from .316 to .82, with 19 items falling within the 95% confidence interval, and three items falling below the lower limits of the confidence interval. In addition, correlation coefficients for the Strengths and Resources domain ranged from .262 to .645, with five items falling within the 95% confidence interval, and three items falling below the lower limits of the confidence interval. Level of Functioning correlation coefficients ranged from .30 to .717, with 5 items falling within the 95% confidence interval, and one item falling below the lower limits of the confidence interval. (See Table 3)

Careful examination of the items within the Problem Severity domain suggest that emotional withdrawal and manic issues warrant further investigation, as both fell below the lower bound of the 95% confidence interval. Similarly, within the Strengths and Resources domain educational/skill resources, personal strengths and empowerment fell below the lower bound of the confidence interval, suggesting the need for further investigation of these items. Finally, within the Level of Functioning domain, Cognitive/Intellectual Functioning fell below the lower bound of the confidence interval, warranting further consideration.

Table 3: Item-Specific Correlations			
CCAR Domain	**CCAR Item**	**Adult Focused Study-Raters**	**Youth Focused Study-Raters**
Problem Severity	Emotional Withdrawal	0.367	0.426
	Depression	0.726*	0.637*
	Anxiety	0.733*	0.522*
	Manic	0.809*	0.316
	Attention	0.66*	0.613*
	Suicidal	0.717*	0.563*
	Thought	0.707*	0.478*
	Cognitive	0.557*	0.562*
	Self Care	0.627*	0.641*
	Resistive	0.426	0.548*
	Social	0.394	0.608*
	Legal	0.826*	0.561*
	Aggressive	0.641*	0.564*
	Family	0.648*	0.715*
	Interpersonal	0.572*	0.638*
	Role	0.489	0.64*
	Alcohol	0.685*	0.82*
	Drug	0.83*	0.655*
	Medical	0.802*	0.639*
	Security	0.318	0.538*
	Overall PS	0.49	0.545*
Strengths and Resources	Public Benefits.	0.803*	0.642*
	Basic Resources	0.42	0.464*
	Education	0.39	0.407
	Support	0.432	0.645*
	Personal Judgment	0.24	0.559*
	Personal Strengths	0.218	0.262
	Empowerment	0.714*	Unable to score
	Overall Strengths	0.161	0.573*
Level of Functioning	Societal/LOF	0.573*	0.661*
	Interpersonal LOF	0.351	0.717*
	Daily Living LOF	0.446	0.487*
	Physical LOF	0.795*	0.679*
	Cognitive LOF	0.443	0.3
	Overall LOF	0.498	0.614*

* Correlations falling within the 95% confidence interval

SUMMARY AND RECOMMENDATIONS

Summary

The goal of the CCAR Interrater Reliability Study was to determine the reliability of the CCAR across rater groups, and to identify potential areas needing improvement. Findings

from this study indicate that overall, the CCAR is a reliable instrument, although some domains and items appear to be more reliable than others.

Problem Severity Domain

Problem Severity scores appear to have the highest reliability across all rater groups. This may be due to the fact that treatment often focuses on areas where mental health consumers are experiencing discomfort. As a result, clinical interviews tend to focus on problem areas. The problem-focused nature of clinical interviewing, therefore, may have inadvertently resulted in videotapes that were predominantly problem-focused. The incorporation of strengths-based models into clinical interviewing may result in increased interrater reliability for the strengths domains. In addition, a content analysis of the videotapes would provide information regarding the amount of time spent talking about problems versus strengths and resources.

Another explanation for the higher reliability in the Problem Severity section may be due to the fact that the current CCAR training (as well as the specific training for Study Raters) focuses largely on the Problem Severity section of the CCAR. In addition to more time spent on Problem Severity, greater leeway is suggested for the Strength and Resources section. Although Study-Raters were trained to criterion on all sections of the CCAR, the greatest portion of time was spent focusing on the Problem Severity section. Therefore, it is possible that these raters had less practice completing the Strengths and Resources section. Increasing time and precision during training of the Strengths and Resources section may result in increased interrater reliability.

Item-specific correlations within the Problem Severity domain indicated that several items fell below the lower limits of the 95% confidence interval. Emotional withdrawal, for instance fell below the confidence interval for both adult and youth rater pairs, suggesting that there appears to be something about the emotional withdrawal construct that is difficult to evaluate. One explanation may be that emotional withdrawal is a highly internal process, which is more difficult to evaluate from an external perspective. Another explanation is that several of the checklist items that are found under the emotional withdrawal item (e.g., passive and subdued) overlap resulting in a lack of discriminant validity. This is consistent with anecdotal observation during criterion training of the Study-Raters, were all raters clearly experienced difficulty differentiating between emotional withdrawal items Thus, it seems likely that lack of discriminant validity of some items may impact the CCAR's reliability across raters.

Manic is another item in the Problem Severity domain warranting further attention. This variable appears to have differential usefulness for youth and adults, as the correlations fell within the confidence interval for adult raters and below the confidence interval for youth raters. One explanation for this discrepancy may be that checklist items related to this variable (e.g., overactive, elevated mood, accelerated speech) may have differential meaning within the context of adults versus youth. Thus, the manic item may have variable usefulness for evaluating adult and youth consumer populations.

Further examination of several items including Socialization Issues, Role, and Resistiveness indicate that all three variables are multifactorial in nature. As mentioned earlier, this suggests that each item contains more than one construct resulting in the failure of these items to hold together and measure a single unique construct. It would be important to review the checklist items accompanying each variable to determine which ones should be maintained because they adequately assess a particular construct, and which ones should be eliminated. This process would help resolve this multifactorial issue.

Another item warranting further investigation is Security/Management Issues. As mentioned earlier, closer analysis indicated that Security was rated virtually identically for all participants resulting in a restriction of range. One explanation for the lack of variation may be that the population sampled was from an outpatient setting. Checklist items related to security and management (e.g., seclusion/restraint, walk away/escape, and suicide watch) may not routinely present during an outpatient clinical interview. It would be helpful to conduct a similar study in an inpatient setting, in order to determine the reliability of the Security/Management Issues item.

Finally, the last item warranting further investigation was Overall Problem Severity. This item approaches the 95% confidence interval for Adult Raters and falls within the confidence interval for Youth Raters. It is important to note that this item is scored differently than all other items in the Problem Severity domain. The overall Problem Severity item has only one rating scale and does not contain checklist items. Raters, therefore, may have less information to assist them in determining the Problem Severity score for this item. In addition, a review of random data sets indicated that this item had more missing data than any other item on the CCAR. Therefore, the format of this item, training on this format, and attention to this area of the CCAR may be at issue.

Strength and Resources Domain

The Strengths and Resources domain appeared to have lower interrater reliability than the Problem Severity domain across all rater pairs. One explanation of this phenomenon may be that the data collection format for this domain is new to the CCAR. The Revised CCAR 2000 is the first version of the CCAR to include accompanying severity scores along with the previously used checklist item format. Thus, CCAR users had no prior familiarity with the process of rating Strengths and Resources on a 1-9 rating scale. Additionally, Strengths and Resources have not traditionally been rated on likert-type scales, rather data has generally been collected as qualitative information on intake summary forms. Therefore, even Study-Raters who lacked prior exposure to the CCAR, may have had equally limited familiarity with rating strengths and resources on an anchored scale.

Another issue potentially impacting the Strengths and Resources section is that the section was greatly expanded from the previous section, and many of the checklist items are new to the Revised CCAR 2000. These items have not undergone internal consistency investigations, and thus there is no prior research indicating whether the items have divergent and/or convergent validity.

Further consideration of the lower interrater reliability on the Strengths and Resources domain also suggests the need to examine current training practices. As mentioned earlier, the majority of the Study-Rater CCAR training focused on an examination and understanding of Problem Severity items. It is important to consider devoting more time and precision to conceptualizing the Strengths and Resources section so that clinicians obtain more practice and understanding of this important domain.

Finally, it is important to note that mental health consumers often enter the clinical process in crisis. Therefore, a great deal of the clinical intake interview generally focuses on problem severity. In addition, strength-based treatment approaches are relatively new to public mental health and therefore, standard clinical interviews may continue to be largely focused on problem areas and deficits. Thus, it is possible that the videotaped interviews used for this study were problem-focused rather than strength-focused, resulting in an in a restriction of the amount of information about consumer strengths and resources. Hopefully, as clinicians gain more training and exposure to evaluating strengths and resources, it is likely that the interrater reliability of this domain will increase.

Initiatives from the Consumer Movement have stressed the importance of strength-based therapeutic approaches. In addition, clinicians in the field have consistently requested an expansion of the Strengths and Resources domain of the CCAR. Therefore, Mental Health Services is committed to the continued investigation and improvement of this important clinical area.

Level of Functioning Domain

Level of Functioning (LOF) also appears to be an area that is difficult to evaluate across all rater groups. This may be due to the fact that these items are similar to overall Problem Severity in that they simply require a numerical rating without checklist items. Less information may be gathered during a clinical interview, therefore, that is useful for determining Level of Functioning. Thus, including checklist items with Level of Functioning scores may result in higher interrater reliability for this domain.

Level of Functioning is also the most abstract and broad area of the CCAR. Interpersonal Functioning, for instance, may require more subjective judgment rather than objective information. In addition, although a clinician might query a parent directly regarding their child's interpersonal relationships, this question is asked more subtly to an adult consumer. Thus, this varied treatment approaches may result in differential scoring for youth and adult consumers.

Cognitive/Intellectual Functioning is a specific LOF item that appears problematic. This item is extremely broad and actually contains 2 different constructs (cognition and intellect). Therefore it might be difficult to assign a single severity rating to this item. Clearly, a review of the current LOF items is warranted to examine the usefulness of these broadly defined items.

Another issue to consider is the differential scoring across Youth and Adult Study Raters. One possibility for this disparity could be differential clinical abilities of Youth and Adult

Raters, which lead to a rater effect. The Youth Study Raters obtained consistently higher interrater reliability in the Level of Functioning and the Strengths and Resources domains. This was not the case, however, for the Problem Severity domain. It is important to note that correlations for Youth fell within a wider confidence interval (446 to .789) as compared to adults (.515 to .757). Therefore, the wider confidence interval may have inherently produced greater opportunities for correlation scores to fall within an adequate range. Closer examination of this discrepancy is clearly indicated.

Finally, each of the Level of Functioning scales requires the use of the CCAR Scoring Manual to determine numerical placement. It is possible that raters failed to use their manuals when scoring this domain, and therefore greatly decreased the interrater reliability of this section. However, manual use was strongly encouraged during the training of Raters. Therefore, it is unclear whether clinicians who participated in the study used their manuals consistently.

Suggestions

Overall, the CCAR demonstrates acceptable interrater reliability, particularly in the Problem Severity domain. The Strength and Resources and Level of Functioning domains demonstrate lower interrater reliability, suggesting the need for further analyses. The following is a list of suggestions to improve interrater reliability, based on the findings of this study.

1. Review items falling below the 95% confidence interval (e.g., emotional Withdrawal, Resistiveness, Cognitive/Intellectual Functioning) to determine whether they should be removed from the CCAR.

2. Revise Problem Severity items that are multifactorial (e.g., role, socialization) in order to develop items that hold together better and demonstrate convergent validity.

3. Conduct internal consistency and factor analyses on the Strengths and Resources checklist items to create a domain that is reliable and useful.

4. Integrate strength-based models into clinical interviewing.

5. Investigate the usefulness of the Level of Functioning scales and determine which items can be eliminated due to their broadness and ambiguity (e.g., Cognitive/Intellectual Functioning).

6. Encourage the use of manuals during CCAR completion in order to ensure definition consistency and accuracy.

7. Incorporate the CCAR into standard clinical evaluations so that it is used as a clinical instrument in addition to its' data collection purposes.

8. Afford more time to clinicians during their daily schedules for CCAR training and scoring.

9. Modify current training strategies to:

- Incorporate greater attention to Strengths and Resources, and Level of Functioning.
- Increase time spent on training to foster a consistent understanding of domain and item definitions.
- Develop a train-the-trainer model so that each CMHC has a designated CCAR trainer to provide on-going training on an as-need basis.

Future Research

1. Complete Analyses for Study 2: Test-Retest Analysis of Study-Raters

2. Complete Analyses for Study 3: Interrater Reliability Study of Agency vs. Study Rater Scores

3. Conduct content analysis of videotapes

4. Conduct study to examine differences between hospital and community settings