Expert versus Youth Raters on Measuring Social and Therapeutic Climate in Secure Juvenile Placement

Sarah Cusworth Walker, PhD, Asia S. Bishop, MSW, Henry Schmidt III, PhD, Terry G. Lee, MD, and Jeff A. Indermark

Growing concern about the use of incarceration is driving significant reform in juvenile legal system decision-making and is likely to have a substantial impact on the role residential options play in the future continuum of care. It appears inevitable that surviving institutions or alternative residential models will be increasingly scrutinized for their impact on youth development. While rehabilitative models focused on youth development are a promising and growing part of residential institutions, few tools are available to measure quality. For institutions to sustain a focus on quality assessment, programs should use an organized and specified treatment model against which staff behavior can be assessed. This study examined the concurrent validity and item functioning of corresponding youth and expert ratings of social and therapeutic climate across multiple sites in a state-wide juvenile residential setting (n = 225 paired observations). Results suggest that the reliability of expert ratings of therapeutic climate exceeds the reliability of youth ratings, whereas reliability for other indicators of social climate are roughly equal between rater types. In addition, youth and expert ratings had weak concurrent validity. Implications for the use of youth versus expertly trained raters for measuring social and therapeutic environment are discussed.

J Am Acad Psychiatry Law 50(2) online, 2022. DOI:10.29158/JAAPL.210064-21

Key words: environmental adherence; social climate; juvenile legal system; youth raters; secure placement

Secure placement continues to be a common sanction for youth involved in the legal system. According to the Census of Juveniles in Residential Placement (United States), over 60,000 youth were placed in some type of secure facility in 2015.[1] About half of these youth spent time in facilities self-classified as residential treatment centers or long-term secure placements, and, as a result, spent a significant amount of time exposed to institutional programming. Youth in secure placements tend to have more severe behavioral health needs than the general population[2] and many institutions are not adequately prepared to meet these needs.[3,4] Growing concern about the effectiveness and ethics of incarceration heightens the importance of attending to the impacts of these settings.[5] To date, little research is available on the quality of forensic institutions generally, and for youth settings in particular.[6]

Although rehabilitative models focused on youth development (e.g., skills-based) are having a growing impact on juvenile corrections,[7] little is known about the effective components of these models and the environments in which they are implemented.[8–11] The available research typically evaluates therapeutic residential programs as black box interventions yielding little information on the impact on effectiveness of variation in staff competencies, youth characteristics, clinical components, dosage, and adherence to the

Published online April 20, 2022.

Dr. Walker is Research Associate Professor and Ms. Bishop is Doctoral Candidate, University of Washington, Seattle, WA. Dr. Schmidt is with Behavioral Affiliates, Inc., Seattle, WA. Dr. Lee is Senior Behavioral Health Medical Director, Community Health Plan of Washington, University of Washington, Seattle, WA. Mr. Indemark is Associate Superintendent, Green Hill School, Washington State Division of Juvenile Rehabilitation, Chehalis, WA. Address correspondence to: Sarah Cusworth Walker, PhD. E-mail: secwalkr@uw.edu.

Data for this study was collected with the support of a grant from the National Institute of Justice 2012-IJ-CX-0040).

Disclosures of financial or other potential conflicts of interest: None.

model.[9,12–14] Program quality monitoring tools can serve the dual function of supporting onsite implementation while contributing broader knowledge regarding the elements that drive successful outcomes in residential programs, including social climate and specific intervention techniques.

Social Climate

Social climate, or the "feel" of a unit's social environment, is considered an important aspect of a rehabilitative milieu in adult and youth psychiatric and forensic psychiatric settings.[15–17] Positive climate in these settings is associated with higher staff and resident satisfaction,[18–20] lower institutional violence,[21] stronger therapeutic alliance[22] and negative attitudes toward offending.[23] Perceived safety and order are also important for facilitating positive social climate and outcomes. For example, research suggests that overcrowding is associated with youth violence toward staff and suicidal behavior.[24] The correspondence between social climate and institutional adjustment suggests this is a valuable measurement construct to use for quality performance monitoring and research.[17] An early measure of social climate for forensic settings, the Correctional Institutions Environment Scale[25] is widely used but is also criticized for poor internal consistency and unreliable factor structure.[26–27]

A more recently introduced measure, the Essen Climate Evaluation Scheme (EssenCES) was designed to address these shortcomings in addition to being shorter and easier to administer.[28] The EssenCES is administered to both clients and staff and measures three domains of social climate in forensic settings: Therapeutic Hold (TH) measures client perceptions of staff support and care; Experience Safety (ES) measures how safe staff and clients feel in the unit; and the Prisoners' Cohesion and Mutual Support subscale measures whether clients exhibit care toward each other. These three domains were developed to reflect face validity[28] and provide evidence of the value of peer support in therapeutic communities on outcomes.[29] A validation study of the EssenCES across adult prison and secure psychiatric settings found that social climate can be reliably measured using these domains.[17]

Adherence and Treatment Quality

Although social climate appears to be a strong predictor of institutional adjustment and treatment outcomes, it does not specifically capture adherence or the delivery of expected treatment components. Adherence to an expected treatment approach is strongly related to youth outcomes in general, including treatment addressing disruptive disorders,[30] complex behavioral health treatment,[31] and reoffending.[32] Adherence in clinical trials and real-world monitoring is typically measured through the use of expert clinicians;[33] however, the use of external raters to assess social or therapeutic climate is infrequent. We did not find any published forensic studies examining the validity of observational climate ratings of secure youth placements by external raters. As secure settings increasingly consider delivering complex, multicomponent therapeutic models, it will be important to ensure that self-report or more passive models of adherence assessment have adequate sensitivity. For example, dialectical behavioral therapy (DBT), a cognitive–behavioral therapy (CBT)-based intervention originally developed to treat borderline personality disorder, is now being widely used in rehabilitative placements and implemented at a pace that has exceeded the ability of research to assess its appropriateness for forensic contexts.[34] Research on the sensitivity and validity of tools to assess quality implementation of therapeutic models is needed.

Current Study

The current study examines the concurrent validity and item functioning of two approaches used to assess the therapeutic environment across multiple sites within a state-run residential placement system for legally involved youth. As the extant literature has established the validity of youth report as a measure of social climate, we examined the correspondence of youth and trained external raters on social climate and related domains to replicate previous research[39] and examine the validity of expert rater scores. We further examine the reliability of items specifically related to therapeutic environment to assess the relative performance of youth versus expert rater measures of social climate. We hypothesized that item functioning, reliability, and associations between youth and expert raters would be comparable on measures of social climate and that expert rater scores would demonstrate superior reliability on therapeutic domains.

The study procedures were approved by the Washington State Institutional Review Board.

Methods

Sample

The study sample included 1,740 youth from 13 youth residential placements from December 2008 through December 2013 in Washington State. All youth in the sample had either a Category A felony (e.g., manslaughter, assault in second degree, robbery in second degree) or had numerous prior criminal adjudications. The demographics of the sample included 979 youth of color (56%), including African American ($n = 297$, 17%), Latino/a ($n = 325$, 19%), American Indian/Alaskan Native ($n = 67$, 4%), Asian/Pacific Islander ($n = 42$, 2%), mixed race ($n = 233$, 13%), and other ($n = 15$, 1%). Approximately 44 percent of the sample was White ($n = 758$) and 0.2 percent did not report race/ethnicity in the sample ($n = 3$). Youth were between 11 and 20 years old ($M = 16.62$, $SD = 1.62$), and primarily male ($n = 1{,}571$, 90%). Institutional settings included four secure facilities ($n = 2{,}115$, 82%), eight community group homes ($n = 441$, 17%), and one boot camp ($n = 14$, 0.5%).

Data and Procedures

Data for the study came from two administrative databases managed by the state juvenile residential agency, hereafter referred to as JR. The first database included youth survey ratings of environmental quality. The second database included environmental quality ratings conducted by highly trained quality assessment staff employed by JR who were external to the residential setting. During the study timeframe, all youth living in the residential units within each facility (i.e., secure facility, community group home/transitional program, or boot camp) were administered institutional quality surveys every two months. The surveys were developed to align with the institution's externally-rated environmental quality assessment. These items were created by JR but align with published measures of institutional quality.[6,35] Youth completed the surveys by hand, which were then collected by the environmental assessment team and subsequently entered into a centralized database. Collection by the environmental assessment team was expected to provide a higher level of anonymity than collection by the unit staff. Youth survey forms were not entirely anonymous, however, in that they included the youth's JR number. How youth perceived the confidentiality of these forms is unknown, but we judge the risk of bias is low as unit staff did not receive punishment or reward as a result of youth responses.

Residential units were assessed by three different expert raters approximately every two months. Raters were full-time employees who received formal training in the assessment process by first observing and then being shadowed by existing raters until they achieved acceptable interrater reliability as determined by the assessment supervisor. The environmental rating process included a day-long site visit by trained staff raters to observe unit climate and residential staff practices. Each living unit was rated by two experts who then compared ratings, discussed discrepancies, and came to a consensus score.

For this study, items from the youth survey and quality assessment tool were organized to match the social climate domains of previous published studies of institutional climate (Table 1). Items from the JR tools were grouped to match the domains of these previous studies by the first and second authors who sorted items individually and then met to compare results and develop the final grouping.[36] This was followed by confirmation from the remaining coauthors regarding the final item placement. Item sets aligned with four subscales of organizational functioning, including overall organization, staff connectedness, social support, and future orientation of the program. In addition to these validated domains, items reflective of the therapeutic orientation of the unit were grouped within a new domain the authors termed "treatment milieu." These items focused on clinical components of the institutional treatment program and staff readiness to support treatment in the therapeutic milieu.

Measures

Overall Facility Organization

Overall organization was created using three items from the youth survey and three items from the expert tool (Likert scales). Example items include "Do you know what structure/activities to expect on a daily basis?" (youth, ranging from 0 [*never*] to 4 [*always*]) and "Important treatment specific information is communicated among staff daily" (staff, ranging from 0 [*poor implementation*] to 3 [*strong implementation*]).

Table 1 Matching Dimensions of Organizational Functioning[6] with Environmental Adherence (EA) Items

Dimension[a]	Definition	Subscales	Example Items	EA Youth Items	EA Staff Items
Institutional order[b]	Set schedules, clear lines of responsibility, and cohesion among staff about shared mission. Clear rules and routines; consistency of messages provided by staff about normative values; positive modeling and connection between staff members and residents	Overall organization	We follow a regular schedule every day.	Item 6. Did staff explain to you how to earn privileges? Item 7. Do you know what structure/activities to expect on a daily basis? Item 8. Do staff lead activities in the program?	Item 5. Program is structured in a way that ensures treatment is occurring Item 6. Youth have structured programming on the floor (behavior permitting) Item 7. Important treatment specific information is communicated among staff daily
		Staff connectedness	Staff worked with kids who were failing.	Item 1. Does the staff's voice remain firm and supportive when a youth is not following directions?	Item 1. Staff are respectful in their communication with youth Item 9. Staff structures milieu to actively engage youth in generalizing skills Item 10. There is a clear program matic structure that pairs privilege to treatment performance
		Staff negative behavior	How much do you see staff using disrespectful language?	None	None
Caring adults	Concerned connection between staff members and residents; Positive relationships with caring adults; Social support provided by key staff members	Social support: Domains	Is there an adult here with whom you can talk about important decisions in your life?	Item 2. Would you describe staff as "excited to work with youth" during interactions? Item 3. Are staff working with you to accomplish your treatment goals? Item 4. Do staff assist you in resolving treatment concerns you may have?	Item 2. Staff convey genuine regard and liking toward youth Item 3. Staff demonstrate that they listen to youth Item 4. Behavior is described in an empathetic, objective and nonjudgmental way Item 11. Staff help youth accomplish treatment goals that are important to the youth
		Social support: Diversity	How many different adults can you talk to about important decisions?	None	None
Reentry planning[b]	Gains made during institutional stays may deteriorate unless reinforced or built on in the community; Institutions provide focused resources to assist in the community transition	Future orientation of the program	Staff help individuals here get jobs.	Item 10. Do staff work with you on how to apply your skills to your community/home setting?	None
		Release counselor	Did you have a person assigned to you to help you out with making	None	None

Table 1 Continued

Dimension[a]	Definition	Subscales	Example Items	EA Youth Items	EA Staff Items
			arrangements for you to return to the community?		
Treatment milieu[c]	Staff members coach youth on emotional and behavioral skills and reinforces use of these skills in the milieu	Not applicable	Not applicable	Item 5. Are you practicing new skills to earn reinforcements (token incentives) from staff? Item 9. Do staff help coach you on how to use your skills?	Item 8. Program effectively reinforces behaviors Item 12. Staff apply DBT strategies in the milieu Item 13. Staff support each other in delivering the treatment with fidelity

[a] Mulvey[6] included eight selected dimensions of organization functioning, including safety; institutional order; harshness; caring adults; fairness; antisocial peers; services; and reentry planning. Dimensions included here are based on availability of EA items that are conceptually similar and align with Mulvey's definitions and example items.
[b] Institutional order, harshness, level of service provision, and release planning distinguished the most among different institutions in Mulvey et al.[6] Intraclass correlations are examined for the subscales.
[c] Extends Mulvey's findings with the addition of a new "treatment milieu" dimension.

Staff Connectedness

Staff connectedness was measured with one item from the youth survey and three items from the expert tool. Example items include "Does the staff's voice remain firm and supportive when a youth is not following directions?" (youth) and "Staff are respectful in their communication with youth" (staff).

Social Support

Social support was measured using three youth items and four expert tool items. Example items include "Are staff working with you to accomplish your treatment goals?" (youth) and "Staff convey genuine regard and liking toward youth" (staff).

Future Orientation of the Program

The youth survey contained one item that directly aligned with the "future orientation of the program" subscale: "Do staff work with you on how to apply your community/home setting?" This item was added to the youth survey when the tool underwent minor revisions after a pilot testing phase (January–March 2012). Consequently, youth ratings for this item are only available for a subset of cases (n = 908).

Treatment Milieu

Two items were used from the youth survey and three items from the expert tool to measure treatment milieu. Example items include "Do staff help coach you on how to use your skills? (youth) and "Staff apply [treatment] strategies in the milieu" (staff).

Analytic Strategy

During the study period, youth provided 2,570 living unit ratings and expert raters provided 677 living unit ratings across 36 living units (hereafter, units). This included 26 units across four secure institutions, nine community group homes, and one boot camp. Because youth and expert rater assessments were not always captured in the same month over the two-month period, scale scores for the expert rater and youth data were aggregated at three-month intervals to ensure the time period captured at least one mean rating from each source. For example, if youth survey scores were conducted at Month 1 and 3 and expert rating scores were available in

Month 2, youth survey scores were averaged for Months 1 and 3 and the expert rating score from Month 2 was assigned to the aggregated 3-month time period (Months 1, 2 and 3). This aggregated score was then treated as a single time point (hereafter, time). After aggregation, the total number of observations included 243 time/unit expert ratings and 228 time/unit youth ratings (termed "analysis units"). Of these, 225 analysis units had both expert and youth ratings. As a result, correlation analyses between youth and expert scales were conducted using a sample size of $n = 225$ time/unit cases. Within this sample, there was an average of 3.49 expert ratings per time/unit ($SD = 2.66$; median = 3, range = 1–13), and 12.25 youth ratings per time/unit ($SD = 8.83$; median = 10, range = 1–48). Individual item performance as well as composite scores aggregated across living unit and three-month time periods (total of 14 time points) separately for youth and expert ratings were examined.

The analytic strategy was consistent with previous assessments of measure reliability for institutional quality (for example, see Ref. 6). Descriptive statistics were used to assess item functioning using SPSS version 24. Intraclass correlation coefficients (ICCs) were used to assess rater reliability for youth within the five subscales of organizational functioning. ICCs were computed using the following formula: ICC = covariance intercept variance/(covariance intercept variance + covariance residual estimate).[37,38] ICCs were not computed for expert ratings because the environmental rating process required that raters reach consensus even though both rater scores are recorded as separate assessments.

Bivariate correlations were used to assess inter-relations among subscales as well as the relationship between youth ratings and expert ratings of organizational functioning. Independent samples *t*-tests were used to determine whether youth and expert ratings varied by institution type.

Results

Reliability × Rater Type

ICC results suggested substantial variation among youth as raters of organizational functioning across living units, with ICCs for individual items ranging from poor (ICC = .10) to excellent (ICC = .97).[39] Compared with previously published youth ratings,[6] our sample of youth raters demonstrated higher consistency in ratings of Overall Organization (ICC = .37 vs. .16 respectively), Social Support (.50 vs. .36 respectively), and Future Orientation of the Program (.50 vs. .37 respectively), and less consistently in ratings of Staff Connectedness (ICC = .29 vs. .50). Youth in the current sample were found to be relatively consistent raters of the treatment milieu (ICC = .34).

In aggregate (mean of all quality assessment items), expert raters had high interrater reliability (ICC = .93) across living units and time, which closely resembled findings from a prior study using a subset of the same data (ICC = .98).[40]

Reliability by Facility Type

Independent samples *t*-tests using standardized means (*z* scores) indicated significant differences in ratings of the treatment milieu by facility type (secure vs. community group homes) among both youth and expert raters with community group homes scoring lower on treatment milieu when rated by youth, $t(222) = -3.68$, $p < .0001$ and expert raters $t(240) = -2.23$, $p < .05$ (see Fig. 1). Descriptively, youth scored both secure facilities ($M = .19$, $SD = .57$) and community group homes ($M = -.11$, $SD = .49$) lower on treatment milieu compared with expert ratings of these facility types (secure facilities: $M = .30$, $SD = .98$; community group homes: $M = .03$, $SD = .50$).

We found no significant differences in youth or expert ratings of institutional order, caring adults, or reentry planning across facility type (secure vs. group home).

Item Functioning within Scales

Table 2 displays the descriptive statistics and reliability coefficients for the individual items and subscales of organizational functioning for both youth and expert ratings. The subscales demonstrated good reliability in our sample, with Cronbach's α coefficients ranging from .61 to .78 for youth ratings and .69 to .83 for expert ratings. Youth ratings demonstrated comparable reliability for the Overall Organization subscale ($\alpha = .67$) compared with the extant literature using youth ratings ($\alpha = .63$),[6] while expert rater scores demonstrated stronger reliability ($\alpha = .80$). The new treatment milieu scale demonstrated acceptable reliability for both the youth ($\alpha = .61$) and expert rater (.78) samples, with

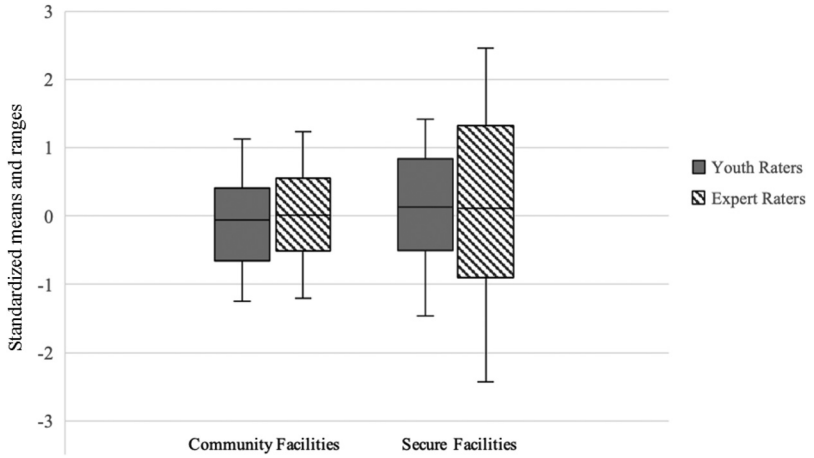

Figure 1. Distributions of standardized mean ratings of treatment milieu for community and secure facilities by rater type. Means are standardized (converted to a scale in which the mean of the responses is zero, z score) to make direct comparisons among the youth and expert ratings. The box plots represent the minimum, maximum, and interquartile (25–75th percentile) range of the scale distributions.

stronger reliability for expert raters. The range among subscales scores and items was greater for expert ratings, suggesting greater precision in measurement by experts compared with youth. This is demonstrated by comparting the distance of the lowest mean scale score from the average score aggregated across all scales. For expert raters, the lowest item score was "program effectively reinforces behaviors" (M = 1.41, SD = .71). Standardized, this item was .58 standard deviations from the mean of all expert rating items. In comparison, the lowest item score for youth ratings was "Would you describe staff as 'excited' to work with youth during interactions?" (M = 2.21, SD = 1.07). Standardized, this item was .36 standard deviations from the mean of all youth items.

Convergent and Concurrent Validity

Scale intercorrelations are reported in Table 3. All subscales of the organizational functioning dimensions were moderately inter-correlated between rater groups, demonstrating weak concurrent validity: Overall Organization (r = .41), Staff Connectedness (r = .43), Social Support (r = .40), and Treatment Milieu (r = .49). Within raters, all scales were expected to demonstrate convergent validity given the interdependence and conceptual overlap among domains. Convergent validity was strongest for the expert raters with all scales highly correlated (r = .75 to r = .82). Youth ratings were moderately to highly correlated (r = .66 to r = .78).

Discussion

This study examined the concurrent validity of youth and expert ratings of social climate and treatment milieu in a state-wide juvenile residential system. Consistent with the study hypothesis, we found that the reliability of expert ratings of treatment milieu exceeded the reliability of youth ratings, although both were in the acceptable range and concordant with previous reliability studies of institutional climate ratings by youth (References 6 and 35, for example). The analysis also revealed that youth and expert rating were only modestly correlated at the lower bound for acceptable concurrent validity.

Low concurrent validity raises questions about the adequacy of measurement. External staff ratings demonstrated higher observed reliability in scale scores. This strongly suggests that expert ratings were the more accurate measure of institutional social and therapeutic climate. At the same time, the findings also replicated previous analyses demonstrating acceptable consistency and reliability of youth as raters of environmental quality. As between-youth reliability scores (as measured by intraclass correlations) were comparable with previously published youth scores on the same domains,[6] we see the results of the current study as replicating and confirming extant research on youth as adequate raters of institutional quality. Together, results suggest that while youth ratings are adequate measures of institutional social climate, greater reliability will likely be achieved with expert raters.

Table 2 Descriptive Statistics and Scale Reliability for Staff and Youth Environmental Adherence Measures

Subscale	Environmental Adherence Measurement Tool	Descriptive Statistics			Scale Reliability	
		Range	*M* (SD)	scale mean (sd)	alpha	ICC
Youth ratings (*n* = 2,570)						
Overall organization	Item 6. Did staff explain to you how to earn privileges?	0–4	2.78 (1.30)	2.65 (0.91)	0.63	0.37
	Item 7. Do you know what structure/activities to expect on a daily basis?	0–4	2.52 (1.15)			
	Item 8. Do staff lead activities in the program?	0–4	2.60 (1.11)			
Staff connectedness	Item 1. Does the staff's voice remain firm and supportive when a youth is not following directions?	0–4	2.37 (1.01)	–	–	0.29[a]
Social support	Item 2. Would you describe staff as "excited to work with youth" during interactions?	0–4	2.21 (1.07)	2.57 (0.89)	0.78	0.39
	Item 3. Are staff working with you to accomplish your treatment goals?	0–4	2.79 (1.07)			
	Item 4. Do staff assist you in resolving treatment concerns you may have?	0–4	2.67 (1.07)			
Future orientation of the program	Item 10. Do staff work with you on how to apply your skills to your community/home setting?[b]	0–4	2.39 (1.24)	–	–	0.50[a]
Treatment milieu	Item 5. Are you practicing new skills to earn reinforcements (token incentives) from staff?	0–4	2.53 (1.24)	2.57 (1.01)	0.61[c]	0.34
	Item 9. Do staff help coach you on how to use your skills?	0–4	2.59 (1.11)			
Expert ratings (*n* = 677)						
Overall organization	Item 5. Program is structured in a way that ensures treatment is occurring	0–3	1.96 (0.91)	1.77 (0.79)	0.80	–
	Item 6. Youth have structured programming on the floor (behavior permitting)	0–3	1.63 (0.95)			
	Item 7. Important treatment specific information is communicated among staff daily	0–3	1.72 (0.95)			
Staff connectedness	Item 1. Staff are respectful in their communication with youth	0–3	2.40 (0.74)	1.83 (0.60)	0.69	–
	Item 9. Staff structures milieu to actively engage youth in generalizing skills	0–3	1.66 (0.80)			
	Item 10. There is a clear programmatic structure that pairs privilege to treatment performance	0–3	1.44 (0.76)			
Social support	Item 2. Staff convey genuine regard and liking toward youth	0–3	2.06 (0.76)	2.03 (0.56)	0.83	–
	Item 3. Staff demonstrate that they listen to youth	0–3	2.23 (0.64)			
	Item 4. Behavior is described in an empathetic, objective and nonjudgmental way	0–3	2.15 (0.67)			
	Item 11. Staff help youth accomplish treatment goals that are important to the youth	0–3	1.69 (0.67)			
Future orientation of the program	–	–	–	–	–	–
Treatment milieu	Item 8. Program effectively reinforces behaviors	0–3	1.41 (0.71)	1.66 (0.55)	0.78	–
	Item 12. Staff apply treatment strategies in the milieu	0–3	1.68 (0.67)			
	Item 13. Staff support each other in delivering the treatment with fidelity	0–3	1.89 (0.59)			

[a] Based on single item.
[b] *n* = subsample of 908 ratings from new version of tool only (i.e., item was not included on old version).
[c] Pearson correlation coefficient = 0.44, $p < .001$.

Greater range in standardized item scores among expert ratings also suggests their reviews yielded higher precision in assessment. This was most apparent in the measurement of the new treatment milieu construct as measured by deviation from the averaged subscale scores. The mean expert rating of treatment milieu was lower than other subscales while the youth rating of treatment milieu did not differ from other subscales.

Table 3 Bivariate Correlations between Standardized Youth and Expert-Rated Scales[a]

Ratings	1	2	3	4	5	6	7	8
Youth ratings								
1. Overall organization	–							
2. Staff connectedness	0.67**	–						
3. Social support	0.66**	0.78**	–					
4. Future orientation of program	0.77**	0.70**	0.84**	–				
5. Treatment milieu	0.76**	0.65**	0.77**	0.78**	–			
Expert ratings[b]								
6. Overall organization	0.41**	0.34**	0.36**	0.33**	0.46**	–		
7. Staff connectedness	0.40**	0.43**	0.40**	0.43**	0.49**	0.82**	–	
8. Social support	0.36**	0.43**	0.40**	0.29*	0.49**	0.75**	0.84**	–
9. Treatment milieu	0.42**	0.39**	0.36**	0.30*	0.49**	0.75**	0.85**	0.82**

[a] Standardized scales aggregated to living unit and 3-month time intervals.
[b] No future orientation scale for staff ratings.
* $p < .01$. ** $p < .001$.

The reliability of measurement was also higher for experts than youth in this domain, which supported our hypothesis that experts would have more knowledge of acceptable levels of fidelity or adherence to the therapeutic approach, which led them to be more precise, and consequently, harsher raters. This aligns with other clinical research in which trained, expert coders of specialized treatment approaches rate therapist competence more harshly than ratings using self-report or client assessment.[41]

At the same time, more precise measurement may not yield meaningful difference in predicting youth skill improvement or outcomes. Although there is some indication that treatment fidelity is related to client symptom improvement,[17,22,30] the overall literature is mixed. A number of studies demonstrate the importance of nonspecific factors, like therapeutic rapport, independent of specific skills as predictors of client recovery. For juvenile residential settings, there is a small literature demonstrating the predictive strength between youth ratings of institutional climate and recidivism outcomes.[6] Further research is needed to determine whether more precise and reliable ratings of social and therapeutic climate made by external raters within an established quality assurance infrastructure translate into more robust indicators of youth outcomes.

Limitations

The study is limited to the therapeutic environments in six residential units across one state. The findings are expected to generalize only to those youth institutions that are using a structured therapeutic approach in which staff are expected to engage in routine positive reinforcement and coaching of youth behavior and social-emotional skills (e.g., problem-solving, emotion regulation, stress management). Although more than half of the youth were youth of color, the largest single demographic was White, and the youth ratings may reflect perceptions that align with systematic differences in experiences of these environments by race/ethnicity and gender that are not fully accounted for in our models. We note the lack of multivariate models to examine possible moderators of youth or staff response by race/ethnicity as a limitation and the results should be generalized with caution. Given the significant variability in youth inter-rater reliability, it is likely that variance due to youth factors not accounted for may affect some of the associations identified between youth and expert rater validity.

Conclusion

Youth rehabilitation and treatment models are growing in importance in the operations of secure juvenile placements. This study found that expert raters provide more precise and reliable assessments of institutional social and therapeutic climate, suggesting that the investment in expert led quality monitoring is important for valid measurement of therapeutic placements. As youth ratings were still in the acceptable range of reliability, institutions not able to invest in expert led review should continue with youth ratings of quality. As expert rating requires more workforce training and resources, the benefit of more precision may be outweighed by the greater feasibility of obtaining youth ratings. Future research will need to examine whether more precise ratings of institutional quality by experts are also better predictors of youth outcomes.

References

1. Sickmund M, Sladky A, Kang W. Easy access to juvenile court statistics: 1985–2019 [Internet]; 2015. Available from: https://www.ojjdp.gov/ojstatbb/ezajcs/asp/display.asp. Accessed December 21, 2020
2. Adams Z, McCart M, Zajac K, *et al*. Psychiatric problems and trauma exposure in non-detained delinquent and non-delinquent adolescents. J Clin Child Adolesc Psychol. 2013; 42:323–31
3. Loughran T, Schubert C, Fagan J, *et al*. Estimating a dose-response relationship between length of stay and future recidivism in serious juvenile offenders. Criminology. 2009; 47:699–740
4. Winokur K, Smith A, Bontrager S, *et al*. Juvenile recidivism and length of stay. J Crim Just. 2008; 36:126–37
5. Maggard SR. Assessing the impact of the Juvenile Detention Alternatives Initiative (JDAI): Predictors of secure detention and length of stay before and after JDAI. Just. 2015; 32:571–97
6. Mulvey E, Schubert C, Odgers C. A method for measuring organizational functioning in juvenile justice facilities using resident ratings. Crim Just & Behav. 2010; 37:1255–77
7. Lipsey M, Wilson D: Practical Meta-Analysis. Thousand Oaks, CA: Sage; 2001
8. Nagin D, Cullen F, Jonson C. Imprisonment and reoffending. In Tonry M, editor. Crime and Justice: A Review of Research (Vol 38). Chicago, IL: Unversity of Chicago Press; 2009. p. 115–200
9. Underwood L, Warren K, Talbott L, *et al*. Mental health treatment in juvenile justice secure care facilities: Practice and policy recommendations. J Forensic Psychol Pract. 2014; 14:55–85
10. Weis R, Whitemarsh S, Wilson N. Military-style residential treatment for disruptive adolescents: Effective for some girls, all girls, when, and why? Psychol Serv. 2005; 2:105–22
11. Woolgar M, Scott S. Evidence-based management of conduct disorders. Curr Opin Psychiatr. 2005; 18:392–6
12. Alonzo-Vaughn N, Bradley R, Cassavaugh M. PBIS in Arizona Department of Juvenile Corrections: How tier II practices build upon tier I. Resid Treat Child Youth. 2015; 32:321–33
13. Cannaday A. Effectiveness of DBT in the milieu regarding increased therapy progress with at-risk adolescents. Ann Arrbor, MI: ProQuest Dissertations Publishing; 2015
14. Jolivette K, Boden L, Sprague J, *et al*. Youth voice matters: Perceptions of facility-wide PBIS implementation in secure residential juvenile facilities. Resid Treat Child. 2015; 32:299–320
15. Moos R. Person-environment congruence in work, school, and health care settings. J Vocat Behav. 1987; 31:231–47
16. Schalast N, Groenewald I. Ein kurzfragebogen zur einschätzung des sozialen klimas im strafvollzug [A short questionnaire for assessing the social climate in correctional institutions-First findings in general prison units and social therapeutic units]. In Haller R, Jehle J-M, editors. Drogen, Sucht, Kriminalitäty [Drugs, Addiction, Crime]. Mönchengladbach, Germany: Forum Verlag Godesberg GmbH; 2009. p. 329–52
17. Tonkin M, Howells K, Ferguson E, *et al*. Lost in translation? Psychometric properties and construct validity of the English Essen Climate Evaluation Schema (EssenCES) social climate questionnaire. Psychol Assess 2012; 24:573–80
18. Marsh S, Evans W. Youth perspectives on their relationships with staff in juvenile correction settings and perceived likelihood of success on release. Youth Violence & Juv Just. 2009; 7:46–67
19. Marsh S, Evans W, Williams M. Social support and sense of program belonging discriminate between youth-staff relationship types in juvenile correction settings. Child Youth Care Forum. 2010; 39:481–94
20. Rossberg J, Friis S. Patients' and staff's perceptions of the psychiatric ward environment. Psychiatr Serv. 2004; 55:798–803
21. Friis S, Helldin L. The contribution made by the clinical setting to violence among psychiatric patients. Crim Behav & Ment Health. 1994; 4:341–52
22. Fox E, Anagnostakis K, Somers J, *et al*. The social climate of a women's forensic pathway of care according to level of security, diagnosis and therapeutic alliance. Eur Psychiatry. 2010; 25:1396
23. Beech A, Hamilton-Giachritsis C. Relationship between therapeutic climate and treatment outcome in group-based sexual offender treatment programs. Sex Abuse. 2005; 17:127–40
24. Parent D, Lieter V, Kennedy S, *et al*. Conditions of Confinement: Juvenile Detention and Corrections Facilities. Washington, DC: Office of Juvenile Justice and Delinquency Prevention; 1994. Available from: https://ojjdp.ojp.gov/library/publications/conditions-confinement-juvenile-detention-and-correctional-facilities. Accessed November 27, 2020
25. Moos R, Schaefer J. Evaluating health care work settings: A holistic conceptual framework. Psychol Health. 1987; 1:97–122
26. Røssberg J, Friis S. A suggested revision of the Ward Atmosphere Scale. Acta Psychiatr Scand 2003; 108:374–80
27. Wright K, Boudouris J. An assessment of the Moos Correctional Institutions Environment Scale. J Res Crime & Delinq. 1982; 19:255–76
28. Schalast N, Redies M, Collins M, *et al*. EssenCES, a short questionnaire for assessing the social climate of forensic psychiatric wards. Crim Behav & Ment Health. 2008; 18:49–58
29. Beech A, Fordham A. Therapeutic climate of sexual offender treatment programs. Sex Abuse. 1997; 9:219–37
30. Schoenwald S, Garland A, Southam-Gerow M, *et al*. Adherence measurement in treatments for disruptive behavior disorders: Pursuing clear vision through varied lenses. Clin Psychol (New York). 2011; 18:331–41
31. Bruns E, Walker S, Zabel M, *et al*. Intervening in the lives of youth with complex behavioral health challenges and their families: The role of the wraparound process. Am J Community Psychol. 2010; 46:314–31
32. Barnoski RP. Providing Evidence-Based Programs with Fidelity in Washington State Juvenile Courts: Cost Analysis. Olympia, WA: Washington State Institute for Public Policy [Internet]; 2009. Available from: https://www.wsipp.wa.gov/ReportFile/1058/Wsipp_Providing-Evidence-Based-Programs-With-Fidelity-in-Washington-State-Juvenile-Courts-Cost-Analysis_Full-Report.pdf. Accessed November 27, 2020
33. Schoenwald S. It's a bird, it's a plane, it's . . . fidelity measurement in the real world. Clin Psychol (New York). 2011; 18:142–7
34. Tomlinson M. A theoretical and empirical review of dialectical behavior therapy within forensic psychiatric and correctional settings worldwide. Int J Forensic Ment Health. 2018; 17:72–95
35. Schubert C, Mulvey E, Loughran T, *et al*. Perceptions of institutional experience and community outcomes for serious adolescent offenders. Crim Just & Behav. 2012; 39:71–93
36. Miles MB, Huberman AM, Saldana J. Qualitative Data Analysis: A Methods Sourcebook. Thousand Oaks, CA: SAGE Publications; 2014
37. Shrout PE, Fleiss JL. Intraclass correlations: Uses in assessing rater reliability. Psychol Bull. 1979; 86:420–8
38. Singer J. Using SAS Proc Mixed to fit multilevel models, hierarchical models, and individual growth curves. J Educ Behav Stat. 1998; 23:323–55
39. Cicchetti D: Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. Psychol Assess. 1994:284–90
40. Walker SC, Bishop AS. Length of stay, therapeutic change, and recidivism for incarcerated juvenile offenders. J Offender Rehabil. 2016; 55:355–76
41. Brosan L, Reynolds S, Moore R. Self-evaluation of cognitive therapy performance: Do therapists know how competent they are? Behavioral and Cognitive Psychotherapy. 2008; 36(5): 581–587