Brief Rating of Aggression by Children and Adolescents (BRACHA): A Reliability Study

Drew Barzman, MD, Douglas Mossman, MD, Loretta Sonnier, MD, and Michael Sorter, MD

The Brief Rating of Aggression by Children and Adolescents (BRACHA) is a 14-item instrument scored by emergency room staff members to assess aggression risk during an upcoming psychiatric hospitalization. In this study, we investigated the inter-rater reliability of the BRACHA 0.9, the latest version of the instrument. After receiving training based on the BRACHA 0.9 manual, 10 intake workers viewed 24 ten-minute videos in which child and adolescent actors portrayed pediatric emergency room patients with low, moderate, or high levels of risk for aggression during an upcoming hospitalization. We then evaluated inter-rater reliability for individual BRACHA items, using three measures of agreement, and reliability for total BRACHA 0.9 scores, using conventional (frequentist) methods and Bayesian techniques for calculating the intraclass correlation coefficient ICC (2,1). Inter-rater reliability for individual items ranged from good to almost perfect, with Kendall's *W* exceeding 0.75 for eight of 14 BRACHA items. The ICC (2,1) for the total BRACHA 0.9 score was 0.9099, with both conventional and Bayesian methods (95% credible interval 0.8530–0.9533), suggesting an excellent level of overall agreement. The BRACHA appears to be an accurate, highly reliable instrument for assessing the risk of aggression by children and adolescents who are about to undergo psychiatric hospitalization.

J Am Acad Psychiatry Law 40:374–82, 2012

Aggression in children during psychiatric hospitalization is a common phenomenon that can harm the physical and mental health of both patients and clinical staff members.[1–5] Better methods for assessing inpatients' risk of acting aggressively may help clinicians institute treatment measures that would improve hospital safety. Although some instruments appear useful for rating severity and type of pediatric aggression[6] (e.g., the Overt Aggression Scale (OAS)) and for assessing violence risk in adult psychiatric inpatients,[7,8] mental health professionals do not yet have a well-validated tool for assessing the potential for violence among children during short-term psychiatric hospitalization.[9]

Recently, Barzman and colleagues[9] showed that the Brief Rating of Aggression by Children and Adolescents (BRACHA) may help clinicians rapidly assess the risk of aggression by child and adolescent psychiatric inpatients. The BRACHA is a 14-item instrument scored by emergency room staff members using information that is consistently available, even during short, high-pressure evaluations. In this initial accuracy study, Barzman and colleagues showed that the sum of the 14 BRACHA items was directly related to the risk of aggression by children and adolescents during psychiatric hospitalization. This finding suggests that the BRACHA may help admitting clinicians differentiate between patients of relatively low and high aggression risk, which could help inpatient staff members plan treatment, reduce injuries, and reduce the need for restraint. To date, however, no information about the BRACHA's inter-rater reliability is available, an important consideration, given that the usefulness of the instrument would

Dr. Barzman is Assistant Professor of Psychiatry and Pediatrics and Director of the Child and Adolescent Forensic Psychiatry Service, and Dr. Sorter is Professor of Clinical Psychiatry and Clinical Director of Psychiatry, Cincinnati Children's Hospital Medical Center, Cincinnati, OH. Dr. Sonnier is a Forensic Psychiatry Fellow, University of Cincinnati, Cincinnati, OH. Dr. Mossman is Director, Glenn M. Weaver Institute of Law and Psychiatry, University of Cincinnati College of Law, and Professor of Psychiatry, Department of Psychiatry and Behavioral Neurosciences, University of Cincinnati College of Medicine, Cincinnati, OH. This study was supported by the American Academy of Psychiatry and the Law's Institute for Education and Research. Address correspondence to: Drew H. Barzman, MD, MLC 3014, 3333 Burnett Avenue, Cincinnati, OH 45229-3039. E-mail: drew.barzman@cchmc.org.

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

Table 1 Abbreviated BRACHA 0.9 Items and Response Options*

Item	Abbreviated BRACHA Items	Response Options	
1	Previous psychiatric hospitalization or day treatment placement	□ Yes	□ No
2	School suspension or expulsion	□ Yes	□ No
3	Trouble accepting adult authority at home or at school	□ Little or none	
		□ Some	□ A lot
4	Frequency of physical aggression toward others (e.g., hitting, kicking punching, biting, slapping, fights at school, throwing objects at others)	□ Never	
		□ Occasionally	
		□ Often	
5	Impulsiveness in the emergency department (e.g., often needing redirection, throwing objects, running out of the room, yelling at the interviewer, extremely talkative, etc.)	□ No Incidents	
		□ One or more incidents	
6	Intrusiveness in the emergency department (e.g., invading personal space, asking personal questions, etc.)	□ No incidents	
		□ One or more incidents	
7	Attempts to harm others or violent acts with intent to seriously harm others (includes all weapons use, even without injury, if used with harmful intent)	□ Never	
		□ Once	
		□ More than once	
8	Violent ideation towards others (i.e., thoughts, wishes, or desires to harm other people)	□ Never	
		□ Occasionally	
		□ Often	
9	Actual expressions of violent intentions or plans to hurt others (includes text messages and e-mails)	□ Never	
		□ Occasionally	
		□ Often	
10	Acts that intentionally destroyed property (e.g., breaking objects, vandalism, fire setting, making holes in the walls; does not include accidents or throwing things)	□ Never	
		□ Occasionally	
		□ Often	
11	Threats or physical aggression towards self or others in the past 24 hours	□ Yes	□ No
12	Pattern of either verbal or physical aggression towards self or others	□ Yes	□ No
13	Aggressive behavior before age 10 years (e.g. firesetting, destruction of property, stealing, trying to seriously hurt a person or animal, bullying, frequent fights; does not include lying)	□ Never	
		□ Occasionally	
		□ Often	
14	Signs of remorse (such as responsibility, shame, and/or guilt) after violence or aggressive acts	□ Not aggressive, or if aggressive, displays remorse, guilt, shame, or responsibility	
		□ If aggressive, displays no remorse, guilt, shame, or sense of responsibility	

* A full version of the BRACHA 0.9 and scoring instructions may be obtained by contacting the first author.

depend in part on there being consistency among ratings performed by various emergency room workers. We investigated the inter-rater reliability of the latest version of Barzman and colleagues' instrument, the BRACHA 0.9.

Methods

This study received approval from the Institutional Review Board at Cincinnati Children's Hospital Medical Center, which granted a waiver of consent because the study was deemed an exempt research project.

Instrument

The BRACHA 0.9 contains some minor modifications from previously described versions of the BRACHA.[9] As was true of previous versions of this instrument, the BRACHA 0.9 is a 14-item instrument that directs evaluators to score 12 historical and behavioral items and two clinical observations. However, the BRACHA 0.9 items are reworded or rephrased to improve clarity and reliability. In previous versions of the BRACHA, evaluators simply assigned ratings of present/yes or absent/no, but in the BRACHA 0.9, several items allow three levels of response (Table 1), a scoring option often used in psychological instruments. We hoped that by allowing graded or intermediate response options, evaluators would not have to shoehorn ambiguous findings or mischaracterize mild expressions of clinical problems as simply present or absent. We also believed that graded responses might be conducive to improved single-item and full-scale reliability and (as we hope to investigate in future studies) that item intensity might be used to increase the BRACHA's predictive power.

Materials and Procedures

We developed a BRACHA 0.9 training manual to help psychiatric intake personnel apply criteria and interpret information consistently. (Readers may obtain a copy of the manual by writing to the first author). We also produced 38 short videos in which the first author interviewed actors who portrayed child and adolescent patients and their legal guardians. The child, adolescent, and adult actors came from two acting classes at local schools. Actors received short descriptions of clinical scenarios derived from actual clinical presentations of children and adolescents, along with instructions to respond to the interviewer's questions using improvisation. Two acting coaches were available to provide assistance and directing to make the videos more realistic. The videos were approximately 10 minutes long and portrayed minors with low, moderate, or high levels of risk for aggression or violence. In making the videos, participants attempted to simulate conditions and clinical practices applied during a short emergency room interview. Although the interviewer made inquiries about clinical problems with BRACHA items in mind, the interviewer did not specifically read and request responses to items in the BRACHA.

Ten emergency room social workers from the Psychiatric Intake Response Center (PIRC) of Cincinnati Children's Hospital Medical Center volunteered to be raters in the study. These raters had 2 to 12 years of experience in conducting evaluations on children and adolescents in our hospital's emergency department. Each rater received one hour's training from the first author that included discussion of the training manual and the BRACHA 0.9 items, coupled with viewing a sample video that was not used for the study. The raters scored the sample video with the first author to foster consensus on ratings, and he answered questions that the raters had about the instrument.

Data Sample

A power analysis (a two-tailed test with 80% power and $\alpha = .05$) showed that 10 raters would have to view 23 videos to detect a reliability difference between intraclass correlation coefficients (ICCs) of 0.4 (which would be considered poor) and 0.7 (the lowest overall reliability level that we considered acceptable).[10] We ultimately selected 24 of the 38 videos for this rating study because they showed a range of aggression risk levels for an age group ranging from preschoolers to adolescents. We culled 14 videos, either because the coverage of risk factors was incomplete or because risk level and age overlapped with the other videos. The 10 raters viewed all 24 videos and scored the 14 BRACHA items solely on the basis of the content of each video (and without other collateral information), using an in-house computer for viewing and recording item scores electronically. Raters received $25 as compensation for their participation time.

Previous work showed that children's ages had an inverse relationship to risk of aggression, independent of the 14 BRACHA interview items, and age was therefore an independent factor in previous BRACHA formulae.[9] The present study included mock patients with ages ranging from 6 to 18 years. In this study, we focused primarily on agreement regarding the 14 items shown in Table 1, but we also used Bayesian techniques (explained further in the Data Analysis section) to evaluate the potential impact of age on the ratings.

Statistical Analysis

After raters finished viewing and scoring the vignettes, the first author and support staff printed each set of results and entered the data by using a double-entry checking method that allowed for verification and correction before analysis. Data sheets were then stored in a secure area.

Because statisticians disagree on optimal methods of evaluating inter-rater agreement, we used three measures of inter-rater reliability for individual BRACHA items: Kendall's coefficient of concordance W,[11,12] Fleiss' generalized kappa (κ_F),[13] and Gwet's AC statistic.[14] Kendall's W is a nonparametric measure of agreement applicable to ranked outcomes or judgments. As originally developed, κ_F and Gwet's AC statistic apply to nominal rather than ranked or ordinal data, so for three-level items, we evaluated these statistics applying the quadratic weighting method recommended by Gwet. To calculate Kendall's W, we used the on-line StatTools calculator developed by Allan Chang, available at http://www.stattools.net/StatToolsIndex.php. We calculated κ_F and AC statistics using the downloadable Excel-based Agreestat software, available at http://www.agreestat.com/agreestat.html.

For purposes of this study, we calculated total BRACHA 0.9 scores using equal weightings for each questionnaire item, with two-option items scored 0 or 1, and three-option items scored 0, ½, or 1.

BRACHA 0.9 scores could thus range from 0 to 14. In cases in which a rater did not score one or more BRACHA items (which occurred 11 times in $14 \times 10 \times 24 = 3{,}360$ instances; 0.33%), we computed the rater's prorated total BRACHA 0.9 score for that vignette as 14 times the average of the answered items.

We evaluated BRACHA 0.9 reliability using conventional (frequentist) methods and Bayesian techniques implemented with WinBUGS.[15,16] We obtained the frequentist ICC (2,1), the appropriate statistic when, as in this case, all subjects are rated by the same raters who are assumed to be a random subset of all possible raters,[17] for total-score agreement using the on-line calculator available at the Chinese University of Hong Kong website (http://department.obg.cuhk.edu.hk/researchsupport/IntraClass_correlation.asp).

Bayesian estimation methods summarize knowledge about unknown parameters using posterior distributions of the probability that a parameter has a particular value, given the observed data and a prior probability of the parameter's value. When priors are noninformative, Bayesian and frequentist methods yield similar inferences.[18] An advantage of Bayesian estimation, however, is that it provides a proper basis for statements such as: the probability that the ICC is between x and y is 95 percent.

WinBUGS allows users to specify a Bayesian model and generate draws from the joint posterior distribution of unknown parameters using Markov chain Monte Carlo iteration methods.[19–21] One discards values from an initial set of burn-in iterations (chosen to be large enough to assure model convergence) to make inferences about model parameters from subsequent iterations. For this study, we adopted methods described by Broemeling[22] and modified his WinBUGS code to make several inferences about Bayesian measures of agreement, sampling from the last half of a three-chain, 200,000-iteration run for each tested model. An example of our WinBUGS code appears in Appendix A.

Results

Individual Items

Table 2 summarizes results of our evaluation of inter-rater agreement for individual BRACHA items. For all items, the three test statistics imply that agreement exceeds chance levels, in most cases by a large margin. Item 7 appears to be an outlier, if one looks at the Fleiss kappa value κ_F alone, because the 95 percent confidence interval for this statistic just exceeds the random range. We note, however, that the base rate for Item 7 was low (2 of 24 videos), a situation in which κ_F is known to perform poorly. The AC statistic does not share this flaw,[14] and, along with Kendall's *W*, it suggests that Item 7 has respectable reliability.

Table 2 Inter-rater Reliability for Individual BRACHA Items

	Kendall's W			
Item	*W*	χ^2 (*df* = 23)*	Fleiss' $\kappa \pm$ SE†	Gwet's AC $\pm$ SE‡
1	0.932	213.4	0.905±0.051	0.937±0.034
2	0.885	194.0	0.851±0.065	0.917±0.040
3	0.794	182.6	0.698±0.053	0.683±0.050
4	0.818	188.1	0.780±0.045	0.782±0.047
5	0.814	146.8	0.585±0.075	0.729±0.076
6	0.773	127.9	0.492±0.153	0.886±0.036
7	0.484	90.6	0.349±0.141	0.844±0.057
8	0.661	152.0	0.568±0.108	0.743±0.070
9	0.802	180.6	0.671±0.087	0.884±0.038
10	0.621	117.2	0.385±0.095	0.849±0.038
11	0.616	141.6	0.565±0.084	0.565±0.083
12	0.671	154.3	0.625±0.095	0.627±0.094
13	0.754	171.9	0.536±0.135	0.840±0.046
14	0.763	168.4	0.703±0.091	0.826±0.051

* All values are significant ($p < 0.0001$).
† All 95% confidence intervals lie above the critical value of 0.033.[14]
‡ AC_1 for two-category items (3, 4, 7–10, 13); AC_2 for three-category items (1, 2, 5, 6, 11, 12, 14). All 95% confidence intervals lie above the critical value of 0.160.[14]

Figure 1 depicts the 10 raters' prorated total BRACHA 0.9 scores as box-and-whisker plots of the five-number summaries (smallest score, lower quartile, median score, upper quartile, and largest score) for all 24 videos. The mean $\pm$ SD of the raters' scores for each video appears along the vertical axis. (To facilitate apprehension of these data, we ordered the 24 videos along the vertical axis from smallest to largest mean BRACHA 0.9 score.) Figure 1 shows that the largest standard deviation of raters' scores was 1.6 points, and for 13 videos, the standard deviation was less than 1 point, findings that informally suggest good inter-rater agreement.

Total BRACHA 0.9 Score

Table 3 provides the results from our ICC (2,1) calculations for the BRACHA 0.9. Using conventional (frequentist) statistical methods, the ICC (2,1) was 0.9099, which implies excellent overall agreement among video raters. Figure 2, however, suggests that Raters 1 and 6 tended to assign lower scores than

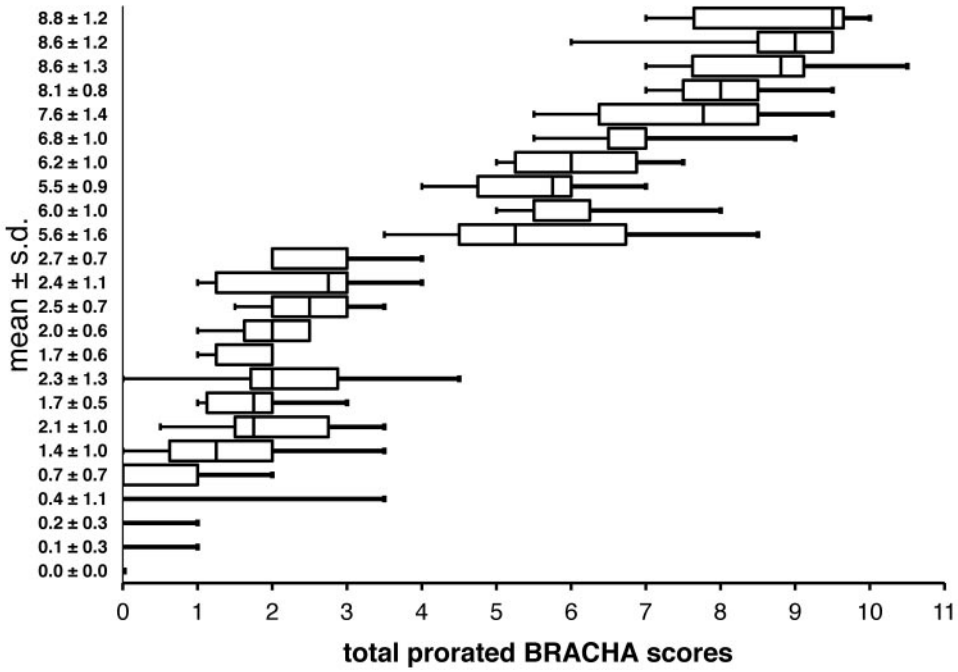

Figure 1. The 10 raters' prorated total BRACHA 0.9 scores for the 24 videos. The scores were a summary of five numbers: smallest score, lower quartile, median score, upper quartile, and largest score.

did the other raters. We evaluated this possibility and its impact by using Bayesian methods,[22] implemented via three data models for the raters' prorated BRACHA 0.9 scores.

In Model 1,

$$y_{ij} = \theta + a_i + e_{ij},$$

where $y_{i,j}$ represents the score assigned to the ith video by the jth reader; θ is a constant representing the mean of the scores; and a_i, the variation in scores attributable to the ith video, has a mean of 0 and a variance of σ_a^2. Note that $e_{i,j}$ represents the residual error, which is the portion of the variation in scores that is not explained by the other variables and has a mean of 0 and a variance of σ_w^2. Broemeling[22] notes that the covariance between two raters can be shown to equal σ_a^2, and consequently, ICC $= \sigma_a^2/(\sigma_a^2 + \sigma_w^2)$. As Table 4A shows, the 95 percent Bayesian credible interval for the ICC is 0.8530–0.9533, confirming the excellent level of agreement suggested by the frequentist calculations.

In Model 2,

$$y_{ij} = \theta + a_i + b_j + e_{ij},$$

we let b_j represent the potential between-rater variation in BRACHA scores hinted at in Figure 2. Comparing the results with this model (Table 4B) with those of Model 1, we found only a tiny decrement in

Table 3 Calculation of the ICC, Conventional Method

	df	Sum of Squares	Mean Square	*F*
Between raters	9	2,030.67	225.63	5.62
Between cases	23	112,035.20	4,871.09	121.41
Within cases	216	10,335.69	47.85	
Residual	207	8,305.02	40.12	
Total	239	122,370.80		

ICC = 0.9099

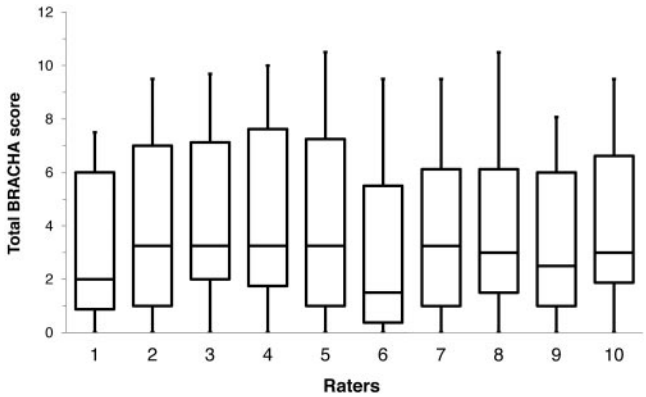

Figure 2. Plots of scores assigned by each rater show that Raters 1 and 6 tended to score videos lower than the others.

Table 4 Bayesian Calculation of the ICC

A. Model 1:

$$y_{ij} = \theta + a_i + e_{ij}; \quad \text{ICC} = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_w^2}$$

Node	Mean	SD	2.5%	Median	97.5%
ICC	0.9097	0.02577	0.8530	**0.9119**	0.9533
σ_a^2	10.3	3.414	5.655	9.733	18.78
σ_w^2	0.9461	0.09194	−0.7828	0.9402	1.142

Deviance: $\overline{D}$ = 666.619; DIC = 691.498.

B. Model 2:

$$y_{ij} = \theta + a_i + b_j + e_{ij}; \quad \text{ICC} = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_b^2 + \sigma_w^2}$$

ICC=σ2aσ2a+σ2b+σ2w

Node	Mean	SD	2.5%	Median	97.5%
b_1	−0.5125	0.2166	−0.9567	−0.5054	−0.1074
b_2	0.1487	0.21	−0.2577	0.146	0.5705
b_3	0.3413	0.2131	−0.06605	0.3368	0.7739
b_4	0.4272	0.2153	0.02044	0.4213	0.8669
b_5	0.1998	0.2106	−0.2077	0.1973	0.6226
b_6	−0.5122	0.2172	−0.9586	−0.505	−0.1055
b_7	0.0672	0.2094	−0.3421	0.06645	0.4842
b_8	0.0267	0.2092	−0.3856	0.0263	0.4399
b_9	−0.3972	0.2136	−0.8346	−0.3919	0.006401
b_{10}	0.2163	0.2102	−0.1886	0.2131	0.6413
ICC	0.9062	0.02869	0.8416	0.9093	0.9526
σ_a^2	10.39	3.409	5.690	9.766	18.74
σ_b^2	0.1952	0.1454	0.05085	0.1585	0.5565
σ_w^2	0.7953	0.07901	0.6558	0.7901	0.9643

Deviance: $\overline{D}$ = 624.948; DIC = 657.596.

C. Model 3:

$$y_{ij} = \theta + a_i + b_j + cx_{ij} + e_{ij}; \quad \text{ICC} = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_b^2 + \sigma_c^2 + \sigma_w^2}$$

Node	Mean	SD	2.5%	Median	97.5%
b_1	−0.5156	0.2222	−0.9651	−0.5113	−0.09033
b_2	0.1659	0.2191	−0.2592	0.1629	0.607
b_3	0.3656	0.2216	−0.0563	0.3603	0.8178
b_4	0.4529	0.2232	0.03095	0.447	0.9103
b_5	0.2184	0.2199	−0.2049	0.2143	0.6642
b_6	−0.5157	0.2222	−0.9666	−0.5109	−0.08992
b_7	0.08211	0.219	−0.3448	0.08011	0.5194
b_8	0.03979	0.2188	−0.3872	0.0379	0.477
b_9	−0.3965	0.2203	−0.8403	−0.3933	0.02854
b_{10}	0.236	0.2196	−0.1873	0.2321	0.6808
c	−0.5156	0.2222	−0.9651	−0.5113	−0.09033
$\text{ICC}_{\text{Model 2}}$	0.9035	0.02976	0.8367	0.9066	0.952
$\text{ICC}_{\text{Model 3}}$	0.8901	0.03456	0.8125	0.8939	0.9462
Diff_{3-2}	−0.01335	0.01244	−0.04756	−0.009631	−0.001874
σ_a^2	9.04	3.154	4.767	8.427	16.87
σ_b^2	0.226	0.1568	0.06714	0.1861	0.6212
σ_w^2	1.35	1.775	0.3154	0.9307	4.887
σ_w^2	0.7942	0.07886	0.6543	0.789	0.9631

Deviance: $\overline{D}$ = 624.572; DIC = 657.308.

agreement (ICC = 0.9062, 95% credible interval = 0.8416–0.9526). Yet the 95 percent credible interval for σ_b^2 does not include 0, so we conclude that between-rater variation (largely attributable to Raters 1 and 6) contributed to the overall variance under Model 2.

As we noted earlier, Barzman and colleagues[9] found that age had an inverse relationship to inpatient aggression, and we therefore wondered whether subjects' age might be a factor in the scores that raters assigned videos. Model 3 expresses this potential relationship as

$$y_{ij} = \theta + a_i + b_j + cx_{ij} + e_{ij},$$

where $x_{i,j}$ represents the potential impact on the jth rater of the subject's age for the ith video. Here, the ICC reflects that portion of the variance arising from the a_i and $cx_{i,j}$ terms. Obtaining convergence for this model required use of vaguely informative priors for the variance terms. The results appear in Table 4C, where the 95 percent credible interval for the difference between ICCs for Models 2 and 3 suggests that the ages of subjects played a small but detectable role in overall inter-rater agreement.

Discussion

We found that the inter-rater reliability of individual BRACHA items ranged from good to almost perfect, when using the criteria of Landis and Koch,[23] and agreement for the total BRACHA score (ICC (2,1) = 0.9099) qualified as excellent according to the criteria of Cicchetti and Sparrow.[24] We note also that our reliability findings concerning the BRACHA 0.9 compare favorably to the inter-rater agreement reported for other well-studied adult risk assessment instruments. For example, Douglas and Reeves[25] reported that in 36 studies of the HCR-20, the median reliability was 0.85 (range, 0.67–0.95); Anderson and Hanson described studies reporting ICCs of 0.87 for the Static-99.[26] Our findings also compare favorably to those of Almvik and colleagues[27] who, in their study of the Brøset Violence Checklist (a six-item adult inpatient assessment instrument), reported individual item kappas of 0.48−1.0 and 0.44 for the total BVC score.

Our reliability evaluation methods differed from those in most other reliability studies of instruments used to assess risk of aggression. In our study, intake workers viewed video recordings of actors portraying emergency room pediatric patients and their adult guardians. This study design had several advantages: it allowed us to examine reliability under evaluation

scenarios with children and adolescents of various ages, with diverse levels of aggression and with a large (and therefore more representative) group of raters. In addition, the raters based their judgments on clinical scenarios derived from actual case presentations of child and adolescent psychiatric patients. This study thus showed that scoring of BRACHA items is reasonably reliable when raters receive information of the sort typically obtained in an emergency room or an urgent office consultation.

Because the BRACHA is intended for use in rapidly assessing emergency room patients for whom hospitalization is anticipated, it would have been impossible to carry out a multirater reliability study under conditions of actual use. By giving raters the same information on which to base ratings, our study's video-scenario design eliminated potential errors in information-gathering that might obscure intrinsic reliability of individual items themselves (and the resulting total BRACHA score).

A real-life reliability study would require having multiple raters interview the same patients in the emergency room, together or separately, something that would probably be impossible to carry out and that would certainly raise questions of ethical practice. Yet we recognize that this unavoidable limitation in our study design prevented us from learning how the information-gathering process affected ratings, which is an important feature of inter-rater reliability. Using actors who worked from scenarios presented in videos designed specifically to capture items relevant to the BRACHA may have enhanced inter-rater agreement above what one would obtain if multiple raters could elicit information from the same subjects under the typical battle conditions of the emergency room.

However, the BRACHA items should form part of an initial psychiatric interview, so we believe that the use of videos depicting an initial psychiatric interview was an appropriate means of assessing important features of the instrument's reliability. Also, most of our raters viewed multiple consecutive videos during work hours in a noisy environment that was at least as distracting as being in a quiet interview room in the emergency department. Had the raters conducted their own detailed, individual interviews of actual patients and been able to seek information specific to the BRACHA items, and had the raters then combined this information with the types of collateral information that often is available (e.g., chart reports on previous hospitalizations), they might have improved factual ascertainment and achieved better reliability parameters than the results we developed from having raters view multiple consecutive short videos.

Conclusions

The BRACHA appears to be a highly reliable instrument for assessing the risk of aggression in children and adolescents in hospitals. This finding, coupled with earlier findings of good accuracy,[9] suggests that the BRACHA can help mental health professionals identify children and adolescents with heightened risk of aggression during psychiatric hospitalization.

References

1. Connor DF, Melloni RH Jr, Harrison RJ: Overt categorical aggression in referred children and adolescents. J Am Acad Child Adolesc Psychiatry 37:66–73, 1998
2. Garrison WT, Ecker B, Friedman M, *et al*: Aggression and counteraggression during child psychiatric hospitalization. J Am Acad Child Adolesc Psychiatry 29:242–50, 1990
3. Ryan EP, Hart VS, Messick DL, *et al*: A prospective study of assault against staff by youths in a state psychiatric hospital. Psychiatr Serv 55:665–70, 2004
4. Sukhodolsky D, Cardona L, Martin A: Characterizing aggressive and noncompliant behaviors in a children's psychiatric inpatient setting. Child Psychiatry Hum Dev 36:177–93, 2005
5. Vivona JM, Ecker B, Halgin R, *et al*: Self- and other-directed aggression in child and adolescent psychiatric inpatients. J Am Acad Child Adolesc Psychiatry 34:434–44, 1995
6. Yudofsky SC, Silver JM, Jackson W, *et al*: The overt aggression scale for the objective rating of verbal and physical aggression. Am J Psychiatry 143:35–9, 1986
7. Almvik R, Woods P, Rasmussen K: The Brøset Violence Checklist (BVC): sensitivity, specificity and inter-rater reliability. J Interpers Violence 15:1284–96, 2000
8. Almvik R, Woods P, Rasmussen K: Assessing risk for imminent violence in the elderly: the Brøset Violence Checklist. Int J Geriatr Psychiatry 22:862–7, 2007
9. Barzman DH, Brackenbury L, Sonnier L, *et al*: Brief Rating of Aggression by Children and Adolescents (BRACHA): development of a tool for assessing risk of inpatients' aggressive behavior. J Am Acad Psychiatry Law 39:170–9, 2011
10. Altaye M, Donner A, Klar N: Inference procedures for assessing interobserver agreement among mulitple raters. Biometrics 57: 584–8, 2001
11. Kendall MG, Babington Smith B: The problem of *m* rankings. Ann Math Stat 10:275–87, 1939
12. Seigel S, Castellan NJ Jr: Nonparametric Statistics for the Behavioral Sciences (ed 2). New York: McGraw-Hill, 1988
13. Fleiss J: Measuring nominal scale agreement among many raters. Psychol Bull 76:378–82, 1971
14. Gwet KL: Handbook of Inter-rater Reliability: The Definitive Guide to Measuring the Extent of Agreement Among Raters (ed 2). Gaithersburg, MD: Advanced Analytics, 2010
15. Lunn DJ, Speigelhalter D, Thomas A, *et al*: The BUGS project: evolution, critique and future directions. Stat Med 28:3049–67, 2009

16. Lunn DJ, Thomas A, Best N, *et al*: WinBUGS: a Bayesian modeling framework—concepts, structure, and extensibility. Stat Comput 10:325–37, 2000
17. Shrout PE, Fleiss J: Intraclass correlations: uses in assessing rater reliability. Psychol Bull 86:420–8, 1979
18. Carlin BP, Louis TA: Bayes and Empirical Bayes Methods for Data Analysis (ed 2). London: Chapman & Hall, 2000
19. Gelfand AE, Smith AFM: Sampling-based approaches to calculating marginal densities. J Am Stat Assoc 85:389–409, 1990
20. Geman S, Geman D: Stochastic relaxation, Gibbs distributions, and the Bayesion restoration of images. IEEE Trans Pattern Anal Mach Intell 6:721–41, 1984
21. Metropolis N, Rosenbluth A, Rosenbluth M, *et al*: Equations of state calculations by fast computing machines. J Chem Phys 21: 1087–91, 1953
22. Broemeling LD: Bayesian Methods for Measures of Agreement. Boca Raton, FL: Chapman & Hall/CRC/Taylor and Francis, 2009
23. Landis JR, Koch GG: The measurement of observer agreement for categorical data. Biometrics 33:159–74, 1977
24. Cicchetti DV, Sparrow SS: Developing criteria for establishing inter-rater reliability of specific items in a given inventory. Am J Ment Defic 86:127–37, 1981
25. Douglas KS, Reeves KA: Historical-Clinical-Risk Management-20 (HCR-20) violence risk assessment scheme: rationale, application, and empirical overview, in Handbook of Violence Risk Assessment. Edited by Otto RK, Douglas KS. New York: Routledge/Taylor & Francis, 2010, pp 147–85
26. Anderson D, Hanson RK: Static-99: an actuarial tool to assess risk of sexual and violent recidivism among sexual offenders, in Handbook of Violence Risk Assessment. Edited by Otto RK, Douglas KS. New York: Taylor & Francis, 2009, pp 251–67
27. Almvik R, Woods P, Rasmussen K: The Brøset Violence Checklist: sensitivity, specificity and interrater reliability. J Interpers Violence 15:1284–96, 2000

Appendix A. WinBUGS Code for Bayesian Estimation of the ICC

```
# BRACHA2-1
# calculation of intraclass correlation coefficient and examination of inter-
# rater agreement, based on Broemeling (Ref. 22, pp 206-208)
# implements Model 1, 10 raters, 24 videos
# responses y[i,j] are total prorated BRACHA scores

model{
      for(i in 1:videos) {
      m[i] <- theta + a[i]
         for(j in 1:raters) {
            y[i,j] ~ dnorm(m[i], tauw) }
      }
      for(i in 1:videos) {a[i] ~ dnorm(0, taua)}
      sigmaw <- 1/tauw
      sigmaa <- 1/taua
      tauw ~ dgamma(0.001, 0.001)
      taua ~ dgamma(0.001, 0.001)
      theta ~ dnorm(0, 0.001)
      icc <- sigmaa/(sigmaa + sigmaw)
   }

# data
list(videos=24, raters=10, y = structure(.Data =
       c(6.5,  7.5,     7.0,  6.5,  7.0,  5.5,  6.5,  5.5,  6.5,  9.0,
         7.0,  7.0,     9.7, 10.0, 10.0,  9.5,  9.5,  7.5,  8.1,  9.5,
         7.0,  8.5,     7.5,  9.5,  8.0,  9.0,  8.0,  7.0,  7.5,  8.5,
         5.0,  3.5,     7.5,  8.5,  4.5,  7.0,  5.5,  5.9,  4.5,  4.0,
         1.5,  2.0,     2.0,  3.5,  1.5,  0.5,  3.5,  1.5,  3.0,  1.5,
         0.0,  1.0,     0.0,  0.0,  0.0,  0.0,  0.0,  0.0,  0.0,  0.0,
         0.0,  0.0,     0.0,  0.0,  0.0,  0.0,  0.0,  0.0,  0.0,  0.0,
         0.0,  1.0,     0.0,  0.0,  0.0,  0.0,  0.0,  0.0,  0.0,  0.5,
         7.5,  8.5,     8.5,  8.0,  9.0,  5.5,  6.0,  7.5,  6.0,  9.5,
         0.0,  4.0,     2.0,  2.0,  4.5,  1.0,  2.0,  2.5,  1.6,  3.0,
         0.0,  0.5,     3.5,  0.0,  0.0,  0.0,  0.0,  0.0,  0.0,  0.0,
         0.5,  1.0,     2.0,  0.5,  1.0,  0.0,  3.5,  1.5,  2.0,  2.0,
         2.5,  4.0,     3.0,  3.5,  1.0,  1.0,  3.0,  3.0,  1.0,  2.0,
         1.5,  1.0,     1.5,  2.0,  3.0,  1.0,  2.0,  2.0,  1.0,  2.0,
         2.0,  3.0,     4.0,  3.0,  2.0,  2.0,  2.0,  3.0,  3.0,  3.0,
         4.0,  7.0,     6.0,  5.5,  5.0,  4.5,  6.0,  6.0,  4.7,  6.0,
         6.0,  9.5,     8.5,  9.5,  9.5,  8.5,  9.5,  9.0,  7.0,  9.0,
         2.0,  2.0,     2.0,  2.0,  2.0,  1.0,  1.0,  2.0,  1.0,  2.0,
         5.5,  5.5,     5.5,  7.5,  8.0,  5.5,  5.0,  6.5,  5.4,  5.5,
         1.0,  0.0,     2.0,  1.0,  0.0,  0.0,  1.0,  1.0,  1.1,  0.0,
         7.5,  9.0,     8.0,  9.2, 10.5,  7.0,  8.6, 10.5,  7.0,  9.0,
         1.5,  2.0,     3.0,  2.5,  3.5,  2.0,  2.5,  3.0,  1.5,  3.5,
         6.0,  7.0,     6.5,  7.5,  5.0,  5.0,  7.5,  5.0,  6.0,  6.0,
         2.0,  1.5,     2.0,  2.5,  2.5,  1.0,  1.0,  2.5,  2.0,  2.5),
 .Dim = c(24,10)))

# inits
list(theta = 0, tauw = 1, taua = 1)
list(theta = 5, tauw = 2, taua = 2)
list(theta = 3, tauw = 0.5, taua = 0.5)
```