UNITED STATES DISTRICT COURT
DISTRICT OF MASSACHUSETTS
Western Division

ROSIE D., et al.,	)	
Plaintiffs,	)	
v.	)	C.A. No. 01-30199-MAP
DEVAL L. PATRICK, et al.,	)	
Defendants	)	

Affidavit of Jack Simons

I, Jack Simons, hereby depose and state as follows:

1. I am the Assistant Director of Children's Behavioral Health Interagency Initiatives for the Massachusetts Executive Office of Health and Human Services (EOHHS), and as such, I report to EOHHS' Compliance Coordinator for implementation of the judgment in the above-captioned matter.
2. I hold a Masters in Education from Harvard University and a Ph.D. in Clinical and Community Psychology from Boston University. My training with regard to research and evaluation methodology includes advanced courses in social science research methodology, which includes the use of both qualitative and quantitative data in social science research, and requires an understanding of psychological measurement and the construction of valid measures. My clinical training includes a predoctoral fellowship in psychology Children's Hospital Boston / Harvard Medical School, and over 20 years of clinical, supervisory and administrative work in child and family mental health, including inpatient, hospital

outpatient, community outpatient, school consultation and court clinic settings, primarily with a low-income population. My experience with regard to systems of care and high-fidelity Wraparound includes being the principal investigator on a SAMHSA grant for Service Systems Change in 2000, starting and managing one of the five Coordinated Family Focused Care pilot sites (Lawrence, MA) from 2002 to 2005, and currently managing the MassHealth contract with Vroon VanDenBerg LLC for training and coaching in high-fidelity Wraparound.

3. I have reviewed the CSR tool and the Court Monitor's report of the findings from the September 2010 sample (*Rosie D. Community Services Review – Western Massachusetts Regional Report*). I have also attended the public presentation on 11/19/2010, at which Karen Snyder and other reviewers summarized initial findings from the November 2010 sample. For comparison purposes I have also reviewed a case review tool that is similar in many ways to the CSR, known as the System of Care Practice Review (SOCPR) developed at the University of South Florida and available at http://logicmodel.fmhi.usf.edu/SOCPR.html.

4. I have concluded that the Community Service Review (CSR), as currently implemented and reported, would be a limited and possibly misleading tool for assessment of system performance. Better tools exist, such as the SOCPR.

5. I have four concerns about the current CSR methodology that lead me to conclude that the review is not a valid means by which to assess or improve practice: (1) reporting of quantitative ratings without supporting estimates of error is not meaningful at best, and in this case is misleading; (2) the rating process is opaque and not subject to verification, making it less credible, less subject to improvement and less useful for quality improvement; (3) The CSR process confounds child or family status with the perceived adequacy of the

intervention; and (4) CSR reviewers are not, to my knowledge, consistently trained in high-fidelity Wraparound and are evaluating Intensive Care Coordination on a clinical model that it was not designed to reflect.

Reporting of quantitative ratings without supporting estimates of error is not meaningful at best, and in this case is misleading.

6. Case studies are a valid source of information for research and evaluation, and can provide information, insights and hypothesis not readily obtainable from quantitative approaches. The CSR reflects a hybrid methodology using multiple qualitative multi-informant case studies, while also rating each case quantitatively on a variety of domains. This approach does allow gathering of rich information, but by selecting cases at random it sacrifices up a major benefit of case studies, which is the ability to select cases in order to illuminate specific issues of interest (for example, by intentionally selecting extreme cases or those that illustrate a critical process). Random sampling is presumably used in order to generalize the quantitative findings from the sample to the population from which it is drawn. But such generalizations must be accompanied by error estimates, and the CSR fails to do this.

7. The current reports of the CSR do not address two significant sources of error: *sampling error* and *measurement error*. All quantitative findings from the CSR should be disregarded unless and until they are reported in a way that takes both error sources into account. Sampling error is easily estimated from sample size (assuming that the sample is random and that each observation is independent of the others). Measurement error is more challenging since it must be estimated through empirical work, such as by measuring the agreement among multiple judges rating the same cases. The developers of the CSR could have, estimated interrater *reliability* for various rating domains of the CSR. Knowing sample

size, reliability, and the sample variance of scores, the evaluators could then have reported inferred population values accompanied by estimates of likely error. They have not done this, however

8. *Sampling error* occurs because samples are usually not perfectly representative of the population from which they are drawn. For example, in a sample of 22 (as reported by the monitor for the Western region sample in September, 2010) if 45 percent are known with certainty to be Latino, we can construct a *confidence interval* which has a 95 percent likelihood of containing the true percentage of Latinos in the population sampled. Using a standard formula for the 95 percent confidence interval for the proportion suggests that the true proportion of Latinos is in the range from 34 to 56 percent. It is a sizable range, but could be useful information nonetheless.

9. But this way of constructing confidence intervals assumes that we can measure our target variable (e.g. the ethnicity or school status of youth in the sample) with zero error. This assumption is questionable for ethnicity, and is certainly untenable when using subjective measures such as ratings. Because of measurement error in ratings, confidence intervals will generally be wider, and potentially very much wider, than those computed based on sampling error alone.

10. Measures such as ratings always involve some degree of *measurement error*. Scientists in any field, whether in the physical or behavioral sciences, require known precision in their instruments. Social scientists have established ways of quantifying the error inherent in their measures. Typically, the *interrater reliability* of a measure is reported, ranging from 0 to 1. Reliability, in conjunction with an estimate of the standard deviation of scores in the population, allows us to calculate the *standard error of measurement*, and from this we can

estimate the range (or confidence interval) that is likely to contain the true score for the individual case.

11. The developers of the CSR clearly understand the importance of reliability and make efforts to train raters to rate reliably. Nonetheless, ratings on the CSR involve considerable subjectivity. Even when the raters are not biased (tending to rate too high or too low), their use of judgment will introduce variability relative to the "true" value. This is measurement error.

12. Measurement error applies to every measured case, and it therefore introduces error into sample averages computed from multiple case ratings. When inferences are made from sample averages to population averages, *both* measurement error and sampling error contribute to the error in estimating the population averages. (The contributions are not strictly additive; but both contribute to the total.)

13. Therefore, measurement error and sampling error *both* affect the confidence that we can have in generalizing ratings from samples to the general population. Low reliability could have a much greater negative impact than sampling error in a sample of 20 to 25 cases. And while we can estimate the sampling error of the CSR ratings from the sample size, we rely on the developers of the CSR to inform us of the measurement error, or reliability. Currently, when CSR ratings are reported they are not accompanied by any information quantifying their reliability. Nor is there, to my knowledge, any published literature establishing the reliability levels of the CSR ratings. Therefore we cannot know what level of confidence, if any, to place in the reported ratings. If the observed rating for a case on one domain is 4, we don't know how likely it is that true score should be 2 or 6. Similarly, if the sample has an average observed rating of 4, we don't have a way to construct the confidence interval around that

estimate of the true mean score for the sample. And without knowing the range of the mean for the sample, we cannot use the sample size to estimate the confidence interval for the population of interest. For all we know, it might contain the whole range from 1 to 6. Subdivisions of the range into "Favorable" versus "Unfavorable" categories, or "Improvement", "Refinement", "Maintenance" in this case become meaningless. Similarly, lacking information about the reliability of the tool, comparisons across domains of practice, over time, or across regions of the state could also involve so much error as to be both meaningless and misleading.

14. The System of Care Practice Review provides an instructive contrast, with published data on the reliability of its ratings (Hernandez et al)[1]. The tool was calibrated as follows: two reviewers participated in the same case review process, conducting interviews and record reviews for 15 cases across 4 sites. They then independently rated each case across all domains. They reported the following inter-rater reliabilities:

Level	Name	Reliability
Domain	Child Centered And Family-Focused	0.91
Subdomain	Individualization	0.90
Subdomain	Full participation	0.86
Subdomain	Case management	0.89
Domain	Community-Based	0.83
Subdomain	Early Intervention	0.48
Subdomain	Access	0.85
Subdomain	Restrictiveness	0.54
Subdomain	Integration	0.80

[1] Mario Hernandez, Angela Gomez, Lodi Lipien, Paul E. Greenbaum, Kathleen H. Armstrong, and Patricia Gonzalez, "Use of the System-of-Care Practice Review in the National Evaluation: Evaluating the Fidelity of Practice to System-of-Care Principles", J. Emotional and Behavioral Disorders, 9:43-52 (2001)

Domain	Culturally Competent	0.90
Subdomain	Awareness	0.83
Subdomain	Sensitivity and Responsiveness	0.85
Subdomain	Informal Supports	0.88
Subdomain	Agency Culture	0.70
Domain	Impact	0.75
Subdomain	Improvement	0.75
Subdomain	Appropriateness	0.70

15. This table tells us that measures within the SOCPR are not all equally reliable; the rating for Child Centered And Family-Focused (0.91, meaning only 9% of the variance in scores due to measurement error) has far less measurement error than that for Restrictiveness (0.54, with 46% of the variance in the score due to measurement error). This is important information for two reasons. First, conclusions based on the less reliable measure merit less deference than those from the more reliable measure. Second, we would look to improve the reliability of our rating process for the less reliable measure. And if we know the variability of each score we can also compute the Standard Error of each measure, which allows us to create confidence intervals for the sample ratings and for comparisons (differences) across domains.
16. The developers of the CSR could use a similar process to estimate error in the CSR ratings, but do not.

The rating process is opaque, making it less subject to improvement and less useful for quality improvement.

17. Summarizing the wealth of qualitative data that is harvested from a case review can be daunting, and using summary ratings by trained raters is one way to deal with this challenge. But for ratings to be credible they should be as transparent as possible. Currently the only

information reported on the ratings is sample means; the ratings are essentially a “black box” of judgment. Furthermore, the developers of the CSR present no empirical evidence that CSR ratings are valid predictors of real-world events. The results of the CSR, therefore, necessarily rank very low on credibility.

18. A large body of research demonstrates that clinical judgment is typically less reliable and less accurate than clinicians think (Dawes, Faust and Meehl, 1989)[2]. In an example particularly relevant to the CSR, Bickman, Karver and Schut (1997) found that interjudge reliability for clinicians deciding what level of care children should receive had an was “close to zero”, while their chance-corrected accuracy in predicting levels of care that the child would actually experience was “very low”.[3] The problem of clinical judgment is particularly great when clinicians making predictions receive no feedback about the accuracy of those predictions. The CSR requires raters to speculate about how each child or youth will be functioning in six months (“Six-month Forecast”). It is doubtful whether their judgment has been informed by feedback on accuracy through six-month follow-ups in children rated in past reviews. In the absence of empirical data supporting the accuracy of the “Six-month Forecast”, there is no reason to believe confidence that raters can predict the future for these youth.

19. By comparison, ratings on the System of Care Practice Review have been shown to be valid predictors of child experience. Stephens, Holden, and Hernandez (2004) found that SOCPR rating s were higher for children in communities with funded systems of care than for

[2] “Clinical Versus Actuarial Judgment”, Science, <u>243</u>:1668-74 (1989).

[3] “Clinician Reliability and Accuracy in Judging Appropriate Level of Care”, J. of Consulting and Clinical Psychology, <u>65</u>: 515-520 (1997).

children in comparison communities, and also found that SOCPR ratings predicted child clinical functioning after one year of service as measured by the Child Behavior Checklist.[4]

20. One way to enhance reliability and confidence in the ratings, utilizing a strength of the case study method, would be to solicit feedback from interviewees regarding the ratings. This would entail some additional work, but would give interviewees an opportunity to enlarge reviewers' understanding of the facts and to correct possible misunderstandings.

21. An improvement in reporting would be to disclose all ratings for all cases, along with a narrative summary of the case. This would give the consumer of the CSR data the opportunity to understand the pattern of ratings, the amount of variation, and the amount of agreement between the two reviewers on each case. It would help, for example, to be able to see whether the patterns of ratings for ICC and IHT are the same.

The CSR process confounds child or family status with the perceived adequacy of the intervention.

22. One example appears in the CSR tool on page 26, in rating the child's Education Status, where one determinant of "Adverse status" is "School personnel have made no effort to provide behavioral interventions or supports that might have helped maintain the youth in school." Here, the efforts of the school personnel are clearly confounded with assessment of the youth's status. This is like confusing the aggressiveness of medical care with the health status of the patient. An example in the Western Massachusetts summary appears on page 18, under Emotional and Behavioral Well-Being: "Further review of the data indicators that reviewers found nearly all of the youth in the sample needed refinement or improvement in the emotional / behavioral well being… The finding of reviewers was that many teams lack

[4] System-of-Care Practice Review Scores as Predictors of Behavioral Symptomatology and Functional Impairment", J. Child and Family Studies, 13:179–191 (2004)

the ‘clinical presence’ and oversight in order to successfully understand, plan, implement and track strategies and supports that would result in more favorable emotional status for youth.” No one disputes that youth in ICC and IHT have a high level of emotional / behavioral need; that is why they are receiving the service, after all, and we all agree with the moral proposition that children need to be well and have good lives. But it is a logical error to confuse the “need” in “children need to be well” with the “need” in “the service needs to be changed or improved”. (The CSR scale categories for status, “Maintenance”, “Refinement”, “Improvement” embody this confusion since they imply conclusions about practice based solely on status.) This logical confusion may explain why this section of the report jumps from a discussion of emotional and behavioral well-being into recommendations about practice. Practice recommendations should be based upon practice observations and should be reported in sections that present evidence about observed practice. Both the tool and the reporting process should be refined to distinguish conclusions about ***status*** from conclusions about ***practice***.

<u>CSR does not call for reviewers to be trained in high-fidelity Wraparound and are evaluating Intensive Care Coordination on a clinical model that it was not designed to reflect..</u>

23. Intensive Care Coordination implements high fidelity Wraparound, which departs from traditional clinical practice in several important ways. Wraparound uses a different assessment process that follows a different path and timeline from traditional clinical assessment. Wraparound uses a broader array of expertise than traditional clinical practice, placing the family’s expertise in the center and adding other expertise as needed. Unless safety is an issue, a Wraparound team is supposed to choose interventions based on the family’s prioritization of needs rather than on priorities established by clinical experts.

24. This contradicts the experience of traditional clinicians who have been trained to take a leadership role in selecting interventions. Traditional clinicians often do not "get" Wraparound. As a psychologist who spent many years doing neuropsychological assessments of clinically complex children and youth, I did not find it easy to discard my "expert clinician" hat when I began to practice Wraparound. We do not embrace Wraparound because it fits our clinical preconceptions, and in fact we resist it. In the end, however, Wraparound works better than traditional clinical approaches for working with children and families with complex and severe needs, and that is why we do it.

25. Drs. Jim Rast and John VanDenBerg describe a form of practice that superficially resembles Wraparound and that conforms more to traditional clinical thinking than Wraparound, which is expert-driven, high-intensity case management. They call this practice "turbo case management". Since the landmark system of care evaluation in the late 1990's known as the Fort Bragg Experiment, we have known that care coordination and a rich array of services does not necessarily lead to improved clinical outcomes. Subsequent research suggests that Wraparound fidelity as measured by the Wraparound Fidelity Index (WFI) does predict clinical improvement. "Turbo case management" produces a lot of expert-driven activity, but does not empower families to solve problems and does not necessarily link them to long-term sustainable natural supports. Programs practicing "turbo case management" may appear exemplary in traditional clinical terms (have excellent clinical assessments, for example), but have poor fidelity on the WFI, and indifferent long-term outcomes.

26. Given that Wraparound cannot be appropriately evaluated in traditional clinical terms, and that experienced clinicians often fail to appreciate the key processes that underlie Wraparound, falling back instead on assumptions that lead to "turbo case management",

there is a burden on the CSR to ensure that reviewers are trained to appreciate Wraparound practices and to distinguish them from less-effective practices, and that the CSR tool explicitly supports this distinction. Since I do not know the experience of all the reviewers with high-fidelity Wraparound, I do not know if this burden has been met. I am very concerned that the CSR reviewers (in the summary documents and the public presentation on November 19) appear to use the word "team" indiscriminately to apply to two very different entities: the Care Planning Team in Intensive Care Coordination and the treatment team in In-home Therapy. Care Planning Teams, when properly implemented, follow specific principles in terms of composition and process. Treatment teams, by contrast, may be constituted and may function very differently. The assessment process and timeline for Wraparound is also different from that in In-Home Therapy. The failure to appreciate the differences between these two services in the CSR process suggests inadequate understanding of Wraparound and a poor basis for evaluation of Wraparound.

27. In conclusion, while case studies can be a valuable tool for evaluation of practice, the current CSR method includes a series of quantitative ratings that, in my professional opinion, should not be reported without estimates of their amount of error. I am concerned that the CSR metholdology encourages raters to make judgments that are speculative, not informed by valid feedback, that equate behavioral health status with need for intervention, and that exceed the scope of the rater's expertise (in particular in evaluating Wrparound). It is not unrealistic to expect the CSR to meet the same standards of reliability and validity as the System of Care Performance Review (SOCPR), which tool also appears more deliberately aligned with the System of Care principles that underly Wraparound. On multiple points of

comparison I would consider the SOCPR to be a superior tool for the purposes for which the monitor currently employs the CSR.

Signed Under the Pains and Penalties of Perjury:

/s/ Jack Simons

Jack Simons

December 8, 2010