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 DR. RAY FOSTER

I. Education and Experience.

1. I have served for the past 14 years as a Director of Human Systems and Outcomes, Inc., a human services consulting organization based in Tallahassee, Florida. I hold a Ph.D. from the College of Education in the Florida State University. My studies included advanced courses in administration, research, measurement, and program evaluation. I am a member of the American Evaluation Association. Early in my career, I worked as an evaluation consultant in a research, development, and demonstration program in a state education agency. Then I moved to a senior associate position in private consulting firm specializing in survey research and program evaluation where I conducted special studies for public agencies and private sector organizations. About 25 years ago I, along with others, began designing, developing, and applying case-based review methods and tools used for measuring organizational performance in human service areas (e.g., developmental disabilities, mental health, special education, and child welfare services) that focused on the connections

between frontline case practice and its impact on persons receiving services. In time, these tools and processes were found to be useful for measuring service system performance and monitoring practice improvement in agencies involved in class action lawsuits. Using such tools and processes, public agencies in three states (Alabama, Utah, and Hawaii) succeeded in meeting and sustaining acceptable levels of practice performance that enabled them to exit from their lawsuits. These tools and processes became know as the Community Service Review (CSR) and the Quality Service Review (QSR) in various jurisdictions. For example, the Washington, D.C. Department of Mental Health uses the CSR for measuring practice performance for lawsuit exit purposes in the Dixon Case. Its use is supported by agency leadership to monitor performance and to guide system improvements. CSR and QSR tools and processes are used on a statewide basis in various jurisdictions for quality measurement and system improvement purposes.

II. The Development and Implementation of the CSR

2. Agencies using the CSR/QSR tools and processes move through a series of ongoing design, development, and refinement steps that produce qualitative case review tools that trained professionals use to appraise the quality of case practice in relationship to outcomes for persons receiving services. These steps involve a design team composed of content experts, stakeholders, and end users of results who shape the content and design of the tools and processes. A prototype case review protocol is developed based on the design requirements recommended by the design team. The prototype protocol undergoes a technical review and refinement process involving a wider circle of content experts and end users selected by the agency and its practice partners. Following the technical review refinements, an initial reviewer training is conducted for new reviewers. Then, a pilot test is conducted using the new protocol on a small sample (usually 12-24 cases). Certified mentor reviewers from other systems are assigned to support the newly trained reviewers during the pilot. A mentor and new reviewer

are paired together to review two or three cases in the course of on-site pilot activity that usually takes a week to complete.

3. The pilot test serves the purposes of testing the new protocol using highly experienced reviewers and training new reviewers who begin their own work toward certification as a qualified CSR/QSR reviewer. Reviewer trainees prepare their case ratings independently from the mentors. Results are compared to check inter-rater reliability and to provide feedback to trainees on their performance. Pilot test experiences lead to further refinements in the case review protocol, database design and management, reporting displays and documents, review logistics, and strategies for supporting the use of results by frontline staff, local managers, and the state agency.

4. Once a pool of certified reviewers is built by new user agency, formal reviewer rating agreement studies may be undertaken. Using a well-developed case simulation, members of the reviewer pool independently conduct a case review by rating the qualitative indicators in the case review protocol. Rated results are compiled and analyzed for the reviewer pool to check rating agreement among reviewers and to perform item analyses. Recent studies of this nature conducted in Wisconsin and Alabama revealed Intra-Class Correlation Coefficients and Alpha Coefficients in the .79-.99 ranges (using SPSS). These ranges are considered to be within acceptable limits for inter-rater reliability. Results can be used to identify and strengthen qualitative indicators that have lower levels of reviewer agreement and to identity any reviewers who tend to be outliers within the reviewer pool so that next steps can be taken to strengthen reviewer agreement with the reviewer pool.

5. These developmental steps produce protocols that are clear and useful to reviewers, practitioners, providers, and end users of results. These steps produce well-trained reviewers who are consistent in their interpretation and rating of qualitative indicators contained in the protocols. These steps progressively build tools, inquiry processes, reviewer pools,

and well-prepared end users at all levels (frontline practitioner, supervisor, manager, state agency leadership) that can produce and use results for positive practice change.

6. Various versions of the CSR/QSR were jointly relied upon by plaintiffs and defendants in other jurisdictions (Hawaii, Utah, Alabama) to successfully measure case-level changes in service system performance over time and to exit lawsuits based on the aggregate use of CSR/QSR results. Similar uses are presently underway in lawsuits in Washington, D.C. and in Los Angeles County, California. These successes were achieved without providing the measurement details that Simons asserts are required for success. The SOCPR has no such record of success in exiting lawsuits.

7. Statewide systems (e.g., child welfare systems in Wisconsin, Pennsylvania, Virginia, Indiana, Utah) that provide children's services are presently using CSR/QSR for measuring case-level practice change and improvements in service system performance for quality improvement purposes. These system-wide uses are being successfully conducted without providing the measurement details that Simons asserts are required for success.

8. Mr. Simons did not participate in the protocol development, refinement, pilot testing, or reviewer training processes. He only used the paper protocol to base his judgments. Therefore, one might suspect that he is uninformed about the process of protocol development as well as how reviewers are trained to use the protocol.

9. The SOCPR, recommended by Mr. Simons, has no such record of use for statewide quality improvement or for monitoring of service system performance for use by a court.

III. The Specific Objections to the CSR Described in the Affidavit of Jack Simons

10. Mr. Simons asserts, without benefit of evidence or direct experience with the tool and process, that the "CSR, which in practice - as well as in theory - fails to measure

appropriately the impact and quality of the remedy services in this case." The remedy of practice improvement in the Rosie D case is of the same nature, substance, and magnitude as required for service system change in Hawaii and other jurisdictions. CSR/QSR measured the quality and impact of services successfully in these other jurisdictions.

A. Error Rates

11. Evaluations that rely on samples have some error associated with the degree to which a sample represents the population from which it was drawn. No sample is ever a perfect representation of the population. The goal is to design a method of sampling that is perceived as credible, feasible, and useful by those involved. Finding a good balance between precision (confidence) and efficiency (cost) of measurement is important. Unlike survey sampling that might use a large random sample of 1,000 informants for a 10-item survey that takes five minute to compete, CSR/QSR involves in depth case-based reviews that may take a day to complete following a record review, a series of interviews, and use of a detailed protocol. The time and effort involved limits the number of cases that can be practically reviewed in a service site during a review.

12. When used for court monitoring purposes, CSR/QSR is based on small samples taken on a recurring cycle at service sites. This is based on an efficient spot-check monitoring approach rather than traditional survey research or experimental research design models. Use of small, proportionate, stratified random samples enables users to discover patterns that emerge from the case review data without having to build the high-cost redundancies associated with traditional scientific samples. For example, in such a sample of 12 cases it is possible to detect low and high levels of performance of core practice functions (e.g., teaming or assessing) across cases reviewed.

13. The focus and purpose of small samples is placed on pattern detection, not on point estimates of proportions or treatment and control group comparisons. Further, the

case stories that accompany the case-based ratings for qualitative indicators reveal the patterns of interactions and interconnections that help explain what is working and not working at the practice points in the service system. Users in other jurisdictions have found this approach to be compelling and affordable. These features distinguish the power of CSR/QSR to surface key patterns for action. Thus, CSR/QSR is based on an *action research model,* not on the inquiry logic used in survey research or experimental research.

14. Care is taken in the sampling processes used in CSR/QSR inquiry when it is used in court monitoring situations. A sampling strategy used in court monitoring situations usually begins with a random selection of a sample of cases to ensure that each case in a service site had a chance of being drawn for review and to ensure that key aspects of the sample are in proportion to the local service population. Usually a sample is composed of cases that have been open for services for at least six months and will likely remain open at the time of review. A simple random draw of cases is entered into a sampling matrix to ensure that the distribution of certain factors (e.g., life stage, current placement, time in service, etc.) within the sample match the local service population distribution from which the cases were drawn. Then, a smaller, stratified sample of a planned size is selected from the sampling. For example, a small service agency might have a sample of 12 cases selected, a middle size agency might have a sample of 24 selected, and a large service agency might have a sample of 36-48 cases selected from the matrix.

15. The sample sizes are based on agreements between the parties that take confidence and feasibility into consideration. Several rules are applied to the cases selected for actual review: 1) informed consent is given by the parent or guardian for the child's participation, 2) no more than one case is selected per local practitioner (e.g., caseworker or care coordinator), and 3) all informants in a case must be available for interviews on the same

day. These criteria reduce the number of cases available in each cell of the sampling matrix. For this reason, the initial sample is usually three or four times the number of cases to be reviewed. For example, if the sample size for a small local agency is 12, then the initial random draw is may be 36 to 48 cases. The approach agreed to by the parties provides the model used for sampling across the service sites to be monitored.

16. The logic of the strategic sampling approach is based on the following points:

1. Every child and family served provides a unique and valid test of the service system at point in time at a given service location.

2. No two cases are identical, although some cases may have sufficient similarities to be grouped into defined and useful categories for examination.

3. Each case served will have different levels of child well-being and daily functioning as a result of changing patterns of practice-related interactions at different points in the life of the case.

4. The service system should provide an individualized and adequate response to each child and family being served at every point in the life of the case, regardless of the child/family circumstances presented.

5. Within the service population, there are frequently recurring patterns of child/family circumstances to which the service system must respond effectively. Cases representing these patterns provide useful tests of the service system's responsiveness and capacity to meet needs. These patterns can be defined and used strategically for drawing cases to form a small and useful sample at specific service sites.

6. Major patterns of child/family circumstances and needs can be targeted in the

selection of a small test sample of cases. Each case represents a test of how well the system responds to each of the selected patterns.

7. A sampling strategy that targets and tests service system responses to a defined set of case patterns or profiles contained in a small stratified random sample (e.g., 12-24 cases per site) that is done repeatedly across review sites will efficiently produce useful patterns of results that will be of high interest and utility to end users. In the Western Massachusetts CSR Report, for example, patterns surrounding the reliability of Crisis Services were identified through the review findings.

17. Stakeholders and end users define the case pattern profiles to be contained in the sample design that will best meet their needs for new learning about current local performance 8 and next step actions to improve that performance. Advantages of using small, stratified random samples are relevance, immediate utility to end users, and feasibility in the time and effort required for data gathering.

18. Mr. Simons asserts that methodological flaws exist in the CSR and that all qualitative findings should be disregarded unless and until they are reported in a way that takes all sources of sampling error into account. His argument assumes that CSR is a form of survey research that is only focused on making point estimates of proportions on indicator ratings. Mr. Simons misunderstands that CSR is focused on pattern recognition, not point estimation per *se.*

19. Mr. Simons offers no alternative tool or method that meets the requirements that he asserts as necessary. Taken to its logical end, Mr. Simon's argument results in the absence of any tool or method that can be applied because no such tool or method currently exists that meets all aspects of the academic argument offered. Of course, his assertion does not consider the successful use of CSR in other jurisdictions that have demonstrated consistent levels of adequate service system performance that enabled their exit from lawsuit

requirements.

B. Reliability of Clinical Judgments

20. Clinical judgment requires basic content knowledge of an area of practice to be examined and practical case-level experience in how practice actually works in real life situations. Qualifications for CSR reviewers require a working knowledge of the area of practice to be reviewed. Clinical judgment is based on pattern recognition. For example, a physician learns to recognize a pattern of signs and symptoms of a disease in order to make a diagnosis.

21. The CSR protocol uses qualitative indicators as measures of child status (e.g., emotional well-being or behavioral functioning) and of core practice functions (e.g., engagement, teamwork, and assessment). Specifications for each indicator detail the central construct to be measured, explanation of situations in which the indicator may or may not apply, key elements for building a fact pattern related to the central construct, and descriptions of six differing fact pattern situation levels ranging across a qualitative continuum (i.e., optimal, good, fair, marginal, poor, adverse). All determinations made by reviewers for status and practice indicators are fact based or fact derived.

22. The only exception to this statement is the six-month forecast that adds the reviewer's projection of the near-term future trajectory of the case. The forecast is based on current case status (known facts), history and tendency patterns in the case (known facts), known events that will occur in the near future (e.g., discharge from residential treatment within the next 30 days), and the expectation that the service system will continue to perform business as usual. The forecast adds speculation to known facts to estimate the most likely near-term path in the case. This element is similar to a medical prognosis made by a physician based on a diagnosis and expected course of a disease. Ratings of practice performance indicators are not based upon the six-month prognosis.

23. Thus, a reviewer's judgment is grounded in the qualitative construct, based on a relevant facts gathered and relied upon in selecting a rating level, and guided by the fact based rating descriptions that accompany each indicator. Furthermore, reviewers work in pairs so that fine discriminations in patterns can be resolved through careful consideration of the facts and criteria applied in the case situations being evaluated.

24. In the Massachusetts' CSR process, additional reliability checks are being applied. The CSR team leader gathered inter-rater reliability data for each site review. Furthermore, every protocol is eventually rated according to 100% agreement between the paired reviewers and the team. That's why the reviewers spend so much time on each indicator and talking through the issues until a consensus is eventually reached.

25. There is a second check made during the debriefings of individual cases to the assembled review team. As reviewers present their case stories and indicator ratings, they are challenged if any rating does not match the data presented orally to the group.

26. A third reliability check occurs with the written case review narrative, in which the facts reported must substantiate the ratings given for indicators. These various checkpoints of reliability and validity far exceed those found in typical research.

27. These various quality control elements work together to provide a high level of consistency among reviewers examining the same fact patterns in the same case. End users have confidence in the accuracy and utility of findings reported. Evidence from inter-rater reliability studies in Wisconsin and Alabama demonstrate that acceptable levels of reviewer agreement when rating the same case.

C. Use of Status Scores to Assess Practice

28. Mr. Simons asserts that: "The CSR conflates appraisals of subject children's status with measuring the effectiveness of the remedial services." At the time a child enters into services his status (e.g., emotional well-being, behavioral functioning and risk, safety,

stability, school performance) may be impaired by the presence of psychiatric symptoms and behavioral challenges that interfere with his daily functioning.

29. The purpose of intervention (i.e., service provision through case practice) is helping the child to get better, do better, and stay better in key status areas that demonstrate he is achieving and maintaining adequate levels of well-being and daily functioning. Positive changes in a child's status that occur over the course of intervention represent important outcomes that benefit the child and family and, more generally, benefit society too.

30. There is a logical association between positive changes in child status (i.e., outcomes achieved) and the provision of services via case practice to achieve adequate levels of child functioning and well-being. This association is the linkage between means (i.e., interventions delivered via case practice) and desired ends (i.e., adequate child well-being and functioning). CSR recognizes the positive association between improvements in child status and effective practice performance by a service system. Many factors may contribute to status improvements for a child (e.g., placement in safe and stable home, strong afterschool and evening services to provide age-appropriate supervised activities, experiencing the benefits of trauma-informed care, the child's maturation), not just the contribution of a single program or service. This is a reason for taking a holistic view of integrated and coordinated children's services that is promoted by system of care principles on which CSR is based. Historically, in court monitoring uses of CSR, the court monitor looks to the aggregate overall system performance rating (across cases in a sample) to broadly evaluate service system performance at a given site and point it time.

31. The most troublesome aspect of Mr. Simons' argument is that CSR should only evaluate the service process, and not how well the child is actually doing during and after receiving services. That would be merely a check on compliance with operation process requirements, which we know is insufficient for evaluating the actual effectiveness of practice.

D. Using the CSR to Assess Wraparound Practice

32. Mr. Simons presents no evidence to support his assertion that CSR is "an inept tool for evaluating the effectiveness of specific remedy services" or that "CSR is particularly ill-suited to evaluate services delivered pursuant to a 'Wraparound,' as opposed to a traditional clinical model." The Hawaii reform was based on system of care principles and practices that were successfully measured and monitored using a CSR tool and process. This demonstrates that CSR is actually well-suited for this purpose.

33. In a case review, the CSR reviewers look for positive changes in child status that occur over time as a combination and sequence of intervention strategies and supports are being delivered to the child. The CSR is looking for *what is working* in a case as determined by positive changes in child status without regard to such artificial distinctions. It is not necessary to make a false dichotomy between forms of intervention strategies and supports; however, it is necessary to determine whether the combination and sequence of all strategies and supports being used is making a positive difference in the life the child and family receiving these services.

34. In summary, CSR has been used successfully and repeated for demonstrating that child-serving agencies are achieving and sustaining consistently adequate levels of frontline practice and service capacity necessary to meet the needs of children receiving services. Federal courts and parties to lawsuits in three statewide jurisdictions have shown that CSR results can be used to exit lawsuits based on meeting required levels of demonstrated performance.

Signed under the pains and penalties of perjury this 30th day of December 2010.

Dr. Ray Foster