

Chapin Hall Issue Brief

Policy research
that benefits
children, families,
and their
communities

December 2009

Chapin Hall at the University of Chicago 1313 East 60th Street Chicago, IL 60637 T: 773.753.5900 F: 773.753.5940 www.chapinhall.org

The Child and Family Services Review Composite Scores: A Critique of Method

By John R. Schuerman and Barbara Needell

This brief summarizes our examination of shortcomings in the Child and Family Services Review process (CFSR), the federal government's effort to hold state child welfare systems accountable for outcomes of children under their care.¹

The goal of improving the performance of state child welfare systems is clearly desirable. Too often these programs are failing our society's most vulnerable children, and it is important that child welfare systems be held accountable for the outcomes they achieve. It is also desirable that these efforts at accountability be data driven, insofar as possible, and it is laudable that the federal government has attempted to do that.

Under the CFSR, states are held responsible for meeting what the Administration for Children and Families (ACF) of the federal Department of Health

and Human Services (DHHS) refers to as "substantial conformity" on seven "outcomes" and seven "systemic factors," under three broad domains of safety, permanency, and well-being.

Forty-five items are reviewed in CFSR, using information from the State Data Profile, the Statewide Assessment, Case Record Reviews, and Stakeholder Interviews.² States that are not in substantial conformity on one or more of the seven CFSR outcomes or the seven systemic factors must develop a Program Improvement Plan (PIP). To date, no state has "passed" the CFSR at the initial stage, and every state reviewed has had to develop a PIP.³ Although the PIP is "negotiated" between ACF and the state, ultimately, the goals and action steps must meet ACF approval. Financial penalties may be incurred if the PIP is not successfully completed.

¹ The full report, *The Child and Family Services Review Composite Scores: Accountability off the Track*, by John R. Schuerman and Barbara Needell is available at www.chapinhall.org

² http://www.acf.hhs.gov/programs/cb/cwmonitoring/tools_guide/procedures/appendixb.htm

³ According to ACF, "All 50 States, the District of Columbia, and Puerto Rico completed their first review by 2004. No State was found to be in substantial conformity in all of the seven outcome areas or seven systemic factors. Since that time, States have been implementing their PIPs to correct those outcome areas not found in substantial conformity. The second round of reviews began in the spring of 2007." (<http://www.acf.hhs.gov/programs/cb/cwmonitoring/recruit/cfsrfactsheet.htm>)

In part because of severe criticism of the initial CFSR, HHS revised the process in the second round of reviews. Despite the fact that CFSR2 attempted to overcome some widely recognized limitations of CFSR1, many of the CFSR1 flaws are repeated and additional problems were introduced. In CFSR2, four national standards are derived from 15 measures in the areas of reunification from foster care, adoption from foster care, children in long-term foster care, and placement stability for children in foster care.

This report focuses on these four national standards and their associated measures (“composites”). We have identified a number of significant problems with the standards and composites:

- **They overlook important variation among states in the demographics of the children and families served.**

ACF treated all states the same in establishing standards for the CFSR2, just as it did in CFSR1.⁴

Because the data from various states are not comparable, it is inappropriate to lump them together to form national standards.⁵

- **They fail to account for systemic state differences in caseload inclusion criteria.**

States differ in the inclusion of certain populations in their child welfare caseloads, such as mentally ill and developmentally disabled children or children in juvenile corrections systems.

- **They draw on data derived from a database (AFCARS) that was not designed to measure**

longitudinal performance, and is still not of the quality to justify imposing fiscal penalties.

State submissions to AFCARS include data about children served at any time during a six-month period. ACF attempts to stitch together these files in order to construct a pseudo-longitudinal database. The result is prone to record linkage errors and still fails to include all of the data needed to accurately track performance over time. Improvements in data have been made in the last few years, but concerns about data quality persist.⁶ Judgments about the adequacy of data should be governed by the consequences of decisions made on the basis of those data. The more severe the consequences, the better the data should be. Because the CFSR process can lead to quite severe monetary consequences for states, the data ought to be quite good.

- **There are inherent practice and policy conflicts between measures.**

The principles of avoiding placement and effecting reunification as quickly as possible conflict with the measure of placement stability. That is, a state that prevents placement whenever possible and effects early reunifications will be left with a placement population that is more difficult, with greater problems, and therefore more likely to encounter multiple placements. States that bring many children (often unnecessarily) into care for very short time periods will show inflated performance on placement stability because those children will almost always have fewer placements. A state moving toward increased

⁴ For a discussion of the importance of adjusting for population and caseload characteristics, see Ramesh Raghavan, “Risk Adjustment Practices,” paper presented to Planning Meeting on the Metrics of Performance Assessment in Child Welfare Systems, National Research Council, Institute of Medicine Board on Children, Youth, and Families. Washington, March 13, 2009.

⁵ The Children’s Bureau website, http://www.acf.hhs.gov/programs/cb/pubs/cwo05/state_data/, contains the *Child Welfare Outcomes 2002–2005: Report to Congress*. This includes comments by some states regarding their submissions of data (presumably for FFY 2005). A number of states comment on the uniqueness of their definitions of data elements and changes in how they have handled data over time, and warn of problems in comparisons both between states and over time.

⁶ In *Child Welfare Outcomes 2002–2005: Report to Congress*, cited above, states also comment on improvements in data quality as well as ongoing problems with it.

permanence for children who have already been in care for long periods of time will demonstrate *decreased* performance on some measures but *improved* performance on others.

- **They count/weight states equally despite enormous differences in the size of the child population.**

In CFSR2, ACF treats all states equally in establishing national standards, as it did in CFSR1, with no weighting by population or child welfare caseload size. Vermont and California are weighted equally, despite the fact that there are nearly 60 times as many children served in foster care in California as there are in Vermont. Thus, some states have inordinate effects on the national standards.

- **They employ a complicated statistical method, principal components analysis (PCA), in the absence of any evidence that such a method is in any way required or superior to simpler and more transparent approaches to measurement.**

PCA is sometimes thought of as a “data reduction technique,” used to simplify a large number of similar measures. It is not clear why ACF used a statistical manipulation to obtain combinations of variables rather than a more straightforward approach, such as a simple average (perhaps after standardizing the original variables). It is not evident that PCA produces a superior result for the purposes for which the combinations were intended. The linear combination constituting a composite measure does not have a concrete, intuitive meaning; it is an abstraction. As a result, it is not immediately evident to policymakers what actions should be taken to improve “scores.” Although one would hope that states’ policy and program decisions would be driven by the best

interests of children, using linear combinations for composites in this way may encourage gaming of the system or other inappropriate actions by the states. In an attempt to “pass” their PIPs, states can analyze the linear combinations weighing two factors: those measures that have the most impact on a composite and those measures that are easiest to affect with policy and program changes. Choosing measures to focus on in this manner may not lead a state to address changes most important for children and families.

- **They make many arbitrary and statistically inappropriate decisions in the use of the PCA procedure, thereby undermining the ranking of states that the method produced.**

Through a Freedom of Information Act request, we obtained the data on which ACF performed PCA. Our examination of these data revealed troublesome issues. Among these is the use of a rotation step that is not discussed in any ACF documentation that we have been able to find. Furthermore, after obtaining principal components in each area, ACF proceeded to average these components, thereby abrogating the idea of PCA in the first place, which is to find independent dimensions among the measures. Our full report details our analysis of these data and our critique of ACF’s use of PCA.⁷

- **They arbitrarily set the national standard at the 75th percentile, and then rely on ill-conceived rules that adjust the standard to a different level.**

In the effort to improve the functioning of child welfare systems, the data on these composites could have been used in various ways. The government could have decided to focus on those states that were performing quite badly, say, below the 25th percentile. Alternatively, a triage approach could have been used:

⁷ Schuerman, J. R. & Needell, B. (2009). *The Child and Family Services Review Composite Scores: Accountability off the Track*. Chicago: Chapin Hall at the University of Chicago.

focusing efforts on those in the middle, giving high-performing states a pass while essentially giving up on low-performing states as unredeemable. Obviously, such a course of action is politically problematic, but triage is often the best use of resources. Still another approach would be to focus on those states with declining scores, requiring them to at least stabilize their performance. Instead, ACF established national standards at the 75th percentile. This is an arbitrary number, apparently chosen to “set the bar high,” and we know of no other justification for it. But, the 75th percentile for the states was not actually used. Rather, what would be the 75th percentile of the normal curve was used (i.e., the 75th percentile of a normal curve with this mean and standard deviation). The reasons for using a normal distribution for this determination are not clear.⁸ The normal curve is a theoretical statistical distribution. Very few quantities in nature are normally distributed, and there is no reason to believe that state outcome data are or should be normally distributed. It would have been perfectly legitimate to simply determine the 75th percentile of the distribution as it stood.

- **They use a flawed method to develop a minimum improvement that is required to avoid millions of dollars in penalties.**

In setting improvement goals for CFSR2, ACF recognized the problem of requiring all states to achieve the same amount of progress, as in CFSR1. Hence, CFSR2 requires a percentage improvement by states, based on their baseline performance in a given year. A somewhat complex calculation was used to determine the amount of required improvement.

The process began with the sampling error (as in CFSR1). Using the sampling error in this context is not justifiable statistically. But then, the average performance of the five states just below the national standard was determined. The ratio of the sampling error to this average was then computed and used as the percentage improvement expected of states.⁹ The selection of five states is quite arbitrary; why not 2, or 10, or 20? Most importantly, why should such a procedure be used to determine the expected improvement of states?

This approach raises another problem. The four permanency national standards are based largely on percents of time some event happens (call this p). They could just as well be based on the percent of time that event does not happen ($100\% - p$). Whichever way one looks at it should not matter in further manipulations of the data. But this is not the case when we take proportions of proportions. For example, one of the national standards for safety in CFSR2 is the proportion of children who were victims of substantiated maltreatment in a particular period who were not repeat victims within the following six months. It should not matter if we were to phrase this in terms of the proportion who were victims. Suppose that a state had a baseline rate of 80 percent not subsequent victims (below the national standard of 94.6 percent) and was required to improve by 10 percent of its baseline. This would mean that the state would be required to improve by eight percent. But, if we look at this in terms of the percent of children who were subsequent victims (20 percent), they would have had to improve by two percent.¹⁰

⁸ Distributions are sometimes transformed to be more normal because some inferential statistical techniques assume a normal distribution in the population from which the sample was taken. It is possible that was the rationale here, since confidence intervals were later constructed (confidence interval construction is an inferential technique). There are, however, significant problems with this procedure.

⁹ This process is explained in ACYF-CB-IM-07-05. Our explanation of this process is simpler than that in the IM, which is unnecessarily complex.

¹⁰ The figure for required improvement is used for illustrative purposes; it is not the requirement derived by ACF. Although this argument is made in regard to a safety measure (where there is only one measure, not a composite), it also applies to the composites, since they are made up of individual measures that include percentages.

It is not at all clear that reasonable outcome indicators can be developed from currently available AFCARS data. A Notice of Proposed Rulemaking (NPRM), released in 2008, proposed that AFCARS be converted to a longitudinal database, along with other reforms.¹¹ At this time, the process appears to be on hold. A fully longitudinal national foster care database is necessary to adequately measure performance. Most, if not all,

states have the data needed for this kind of database, and a review of the public comments to the NPRM suggested that this change is welcome.¹²

A functional, useful CFSR process would encourage improvement with a clear and coherent use of available data, with an eye to ways to improve both the quality and quantity of data related to children and families who come to the attention of the child welfare system.

¹¹ *Federal Register Proposed Rules*, Vol. 73, No. 8, January 11, 2008.

¹² <http://www.regulations.gov/search/Regs/home.html#docketDetail?R=ACF-2007-0125>

ChapinHall at the University of Chicago

Established in 1985, Chapin Hall is an independent policy research center whose mission is to build knowledge that improves policies and programs for children and youth, families, and their communities.

Chapin Hall's areas of research include child maltreatment prevention, child welfare systems and foster care, youth justice, schools and their connections with social services and community organizations, early childhood initiatives, community change initiatives, workforce development, out-of-school time initiatives, economic supports for families, and child well-being indicators.

Recommended Citation

Schuerman, J. R., & Needell, B. (2009). *The Child and Family Services Review Composite Scores: A Critique of Method*. Chicago: Chapin Hall at the University of Chicago.

Related Publications

Schuerman, J. R. & Needell, B. (2009). *The Child and Family Services Review Composite Scores: Accountability off the Track*. Chicago: Chapin Hall at the University of Chicago.

Wulczyn, F. & Orlebeke, B. (2006). *Getting What We Pay For: Do Expenditures Align with Outcomes in the Child Welfare System?* Chicago: Chapin Hall at the University of Chicago.

For a complete list of Chapin Hall projects and to download publications, please visit our website.

Contact

Chapin Hall at the University of Chicago

1313 East 60th Street
Chicago, IL 60637
T: 773.753.5900
F: 773.753.5940
www.chapinhall.org