National Coalition for Child Protection Reform / 703-212-2006 / info@nccpr.info / www.nccpr.org

THE TROUBLE WITH CHILD AND FAMILY SERVICES REVIEWS

The Federal Government's Failed Attempt to Measure Child Welfare System Performance

Released August 2003, **Updated June 2008 and September 2025**

Once every six years, the federal government evaluates each state's child welfare system. The process is staggered over several years, but the first results are coming out now for round two of these Child and Family Services Reviews (CFSRs).

And once again, as the results are announced news media are making the same mistake they made last time: they're paying attention.

Some editorial writers even draw sweeping conclusions concerning how their states compare to others and whether their systems are good or bad, getting worse or getting better.

That's unfortunate. **As a way to measure an individual child welfare system's performance, CFSRs are almost entirely worthless. As a way to compare states, they *are* entirely worthless.**

And that's the best case. At worst, CFSRs can give an unearned seal of approval to lousy systems and penalize those that are doing relatively well.

It's not just publicity that is at stake. The federal government can take money from systems that supposedly fail. That means some systems might actually have to worsen their performance to keep federal aid as a result of the flawed, sometimes perverse means used to measure performance on CFSRs.

Using CFSRs to compare states is a difficult temptation to resist. In several cases, CFSRs could be used to lend support to NCCPR's point of view.

For example, "stakeholders" interviewed for Michigan's first CFSR singled out the state's Intensive Family Preservation Services (IFPS) program for praise. A system we consider a national model of safe, successful family preservation, Alabama's, was one of the few in which keeping children safe was rated as a "strength" on the first CFSR. It was rated that way again during the second round.

And in Arizona, where the Governor worsened a foster-care panic by demanding the removal of huge numbers of children, alleging the system did not do enough to keep children safe, the CFSR said, in effect, that the Governor is wrong. It gave Arizona a passing grade for safety before the panic.

But the truth is, one can't tell if any system is keeping children safe based on a review that combines an absurdly small sample size with highly subjective evaluations by parties who may lack objectivity. That's how CFSRs work. And that is only the beginning of what's wrong with them.

Therefore, NCCPR will continue to rely on the wealth of scholarly evidence supporting Michigan's IPFS program, on the detailed, in-depth reviews of an independent, court-appointed

monitor in judging Alabama through 2007[1] and on the history of how foster care panics endanger children in discussing the problems in Arizona. The CFSRs are irrelevant and should be treated as such. The more seriously they are taken, the more dangerous they have the potential to be.

This kind of critique of CFSRs runs the risk of giving "aid and comfort" to lousy child welfare agencies. Those agencies are quick to cite flaws in any process that exposes their poor performance.

But there is an abundance of good, reliable information indicating that most of America's child welfare systems are doing a poor job. Saying that CFSRs are unreliable does *not* mean that agencies are, in fact, doing well.

The critique that follows relies extensively on an outstanding analysis done by the National Center for Youth Law.[1] The opinions expressed in this document by NCCPR should not be taken as reflecting the views of NCYL or any other organization.

Here are some of the problems with CFSRs:

Sample Size. Most state child welfare systems intervene in the lives of tens of thousands of children, either by putting them in foster care or overseeing their families while the children remain in their own homes. But during the first round of reviews, the CFSRs looked at a grand total of only 50 cases in each state. During the second round, the federal government made a "concession" to critics and increased the figure to a whopping 65. Statistically, that means almost nothing.

In contrast, class-action lawsuits against child welfare systems usually use "case readings" of 350 to 500 cases.

And even 65 overestimates the number of cases used in CFSRs to evaluate some measures. The 65 includes cases where children are left in their own homes and foster care cases combined. So when evaluating, for example, "are siblings kept together in foster care?" the actual number of cases examined is 65 - minus the in-home cases, minus the cases where only one child is removed. In some cases, that means the "success" of a state by this criterion was measured using a sample of only ten cases.[2]

CFSRs also are supposed to use overall data supplied by the state itself, but the final, subjective evaluations depend heavily on the 65-case sample. In many cases, states don't even compile the needed data, and when they do, it sometimes is ignored.

For example, NCYL reports that California's statewide data show that the state would not meet the CFSR standard for keeping siblings together while in foster care. But the tiny sample of cases from the nation's largest state did pass. So for California, keeping siblings together was rated as a "strength."[3]

Furthermore, many of the same measurement problems discussed later in this document, such as the reliance on average length of stay, apply equally to the CFSR sample and the statewide data. In addition, a report from the University of Illinois Children and Family Research Center finds that states repeatedly interpret rules for gathering these data differently – and

[1] Such reviews have ended because Alabama has been released from the consent decree that required them. In the future, NCCPR will rely on objective state data and the assessment of the organizations that brought the original lawsuit. A member of NCCPR's board of Directors is Legal Director for one of those organizations, the Bazelon Center for Mental Health Law.

often erroneously[4] – making state-to-state comparison impossible. And NCCPR discovered that one state, Kansas, exploits a loophole in federal regulations (or exploits the unwillingness of the Department of Health and Human Services to enforce those regulations) to keep a huge proportion of its placements out of these data entirely.[5]

The CFSR process is a "pass/fail" test. If performance by the agency in 90 percent of the 65 or fewer cases is deemed acceptable by the reviewers, the state passes in that category. But the tiny sample size means there actually is a huge margin of error.

For example, in public opinion polling, typically the margin of error may be plus or minus three percentage points. So a poll that says, for example, that 40 percent of Americans approve of how the President is doing his job, that really means it is likely that somewhere between 37 and 43 percent like him.

To get this relatively low margin of error, the poll is likely to include at least a thousand respondents.

But if you reduce the sample size to 50, the results are likely to be so unreliable that they can vary by plus or minus at least 12 percentage points. Not 12 percent – *12 percentage points*.[6]

As for the federal government's "concession" in raising the sample size to 65, that reduced the margin of error – to *11 percentage points*.

In other words, if the CFSR says that in 80 percent of the cases the children were kept safe, that really means that anywhere from 69 percent to 91 percent of the children were kept safe. Even that assumes a purely random sample, which, as noted below, was not the case with CFSRs.

Reduce the sample further, to 25 cases, as is true for some CFSR measures, and the margin of error rises to plus or minus at least 18 percentage points.

At a minimum, it should be clear that the margin of error is so wide that comparing states is effectively impossible, and even within a state, the results mean very little. Chances are excellent that, if a different case or two were picked at random, states now listed as passing would fail and vice versa.

The sample size problem is only the beginning. Here are some others:

Sample choice: The sample isn't entirely random. The Department of Health and Human Services (HHS) identifies a total of 300 cases. But the state agency being reviewed then gets to choose the final 65 from among that 300. Furthermore, the final 65 are identified weeks before the actual federal review. According to NCYL, one state hired its own consultants to look over the files before the federal reviewers arrived.[7]

Low Standards: Though the reviews are widely described as "rigorous," the evidence suggests otherwise. For example, in measuring whether states are successful at keeping siblings together, a case is rated as a "strength" if any two siblings are kept together. If the case involves a family with four children, the children can be divided among three foster homes, yet the CFSR process considers this a success.[8]

The rigor is further lowered by the choice of reviewers. Some come from the agency being reviewed. The rest currently work or previously worked for other child welfare agencies.[9] Thus, the reviewers may have an excess of sympathy for the problems and pressures faced by child welfare agencies.

Failure to follow Department of Health and Human Services rules. In addition to reviewing files, reviewers are supposed to interview most of the key people involved with each of the 65 cases. Other interviews are done with key "stakeholders" who offer an overview of a state's child welfare system.

NCYL notes that, when interviewing parties to individual cases, children, birth parents and foster parents are supposed to be interviewed in the home, and "in order to facilitate candid and honest disclosures, interviews are supposed to be conducted in private without the caseworker or other agency representative present." However, "One reviewer who participated in CFS Reviews in three states reported that all interviews occurred in the agency office. Another reported going out to the homes in about half of the cases reviewed. In some situations, brief telephone conversations were substituted for the in-person interviews."[10]

We would add one further point: Most birth parents know that child welfare agencies can be vengeful. Birth parents are likely to be terrified that candid comments would get back to the agency and be used against them.

Subjectivity: Different teams of evaluators examine different states. In child welfare, it is easy to find "experts" with identical qualifications who, if given 65 cases, are likely to come to radically different conclusions. And, much as in college courses, some evaluators are "easy graders" and some are "tough graders."

And there is plenty of room for such disagreement. For example, states are rated on their ability to prevent children from bouncing from foster home to foster home. But if the reviewers think a particular move was in a child's best interests, the move doesn't count against the state. As NCYL notes, in deciding if a move serves a child's best interests, "much discretion appears left to the individual reviewers."[11]

Inadvertent bias against systems that emphasize family preservation. It wasn't done on purpose, but the CFSR outcome measures can bias results against states succeeding at safely keeping families together. That's because CFSRs measure things like average length of stay in foster care and average time to reunification, but not efforts to prevent foster care in the first place.

As a result, a state in which caseworkers remove children at the drop of a hat, then realize they made a big mistake and quickly move to return a lot of those children (much the worse for the experience) will have a low average length of stay, and a brief average time to reunification, so they will be "rewarded" in the CFSR process. In contrast, a state that truly removes children only as a last resort may well have to keep those children in foster care longer, so those states will look bad in the CFSR process.

And, as noted above, this is a problem both with the tiny sample used specifically for the CFSR and the statewide data.

Of course, length of stay is an important measure and should be part of the CFSR process. And, of course, in many systems, long length of stay is a result of those systems' failures,

not their ability to keep children out of foster care in the first place. But as long as CFSRs measure length of stay but don't measure safe, successful prevention of foster care, the few systems that do a good job in this area are penalized.

Inadvertent bias in selection of locations measured. Under the CFSR process, the 65 cases are selected from three communities, and states are at least encouraged and possibly required to include their largest metropolitan area. In many states, that largest metropolitan area will be a big city with a relatively large population of poor children - and therefore, quite likely, one of the state's more troubled child welfare systems. Thus, California's CFSR must include Los Angeles, for example. But it so happens that the largest metropolitan area in Virginia also is one of the most affluent places in the entire nation - Fairfax County. Among America's more than 3,100 counties, fewer than 75 have lower child poverty rates than Fairfax. This almost guarantees that Virginia will look better than it deserves on its CFSR.

Similarly, at the time of the first round of CFSRs, one of the most troubled systems in Ohio was Cleveland/Cuyahoga County. According to news accounts, one of that county's biggest problems involved sending foster children far from their homes and communities; indeed, all over the state. But the largest metropolitan area in Ohio happens to be Columbus/Franklin County. So the Ohio CFSR didn't even look at Cuyahoga. And when the results came in, placement of children close to their own homes was rated a "strength" in Ohio.[12]

Other bias in selection of localities measured. The choice of the other areas to evaluate is made jointly by the federal government and the state under examination. At the time of Colorado's first CFSR, El Paso County, then considered a national model of good child welfare practice, made it into the sample. And in Florida, the one county known to be slightly less atrocious than the rest of the state at that time, Sarasota, wound up among the chosen three.

A top federal official at the time acknowledges that HHS *knew* Sarasota County was unrepresentative of Florida's typical performance. Susan Orr, who was then Associate Commissioner of HHS' Administration on Children, Youth and Families, told the *Orlando Sentinel*: "We were interested in Sarasota because they are the poster child of good case practice … We wanted to see if everyone was doing what everyone says they are doing."[13]

An extra level of irrelevance for county-run systems. In at least 11 states, individual counties run their own child welfare systems. They tend to have a high degree of autonomy, and there often are sharp differences among them. So it's hard to see how even an accurate CFSR says anything meaningful about these states.

Better alternatives. Faced with all the problems inherent in the CFSR process, defenders argue that at least they're better than previous reviews, which just checked if states' paperwork was in order.

Not necessarily. At least everyone *knew* the old reviews were worthless and didn't take them seriously.

CFSRs may penalize the few places that truly are doing comparatively well, while letting states off the hook that do poorly. They may allow states that either are lucky or can manipulate the sampling process to fend off legitimate criticism.

A better approach would be to reduce the total number of variables measured, but add a means to measure placement prevention. Then increase the sample size to 300, which would re-

duce the margin of error on many measures to roughly plus or minus five percentage points.

At the same time, states should be required to further improve and standardize their data-gathering for *all* cases, in order to reduce reliance on sampling.

---

[1] Bill Grimm and Isabelle Hurtubise, "Child and Family Services Reviews: An Ongoing Series, Part One: A Background to the Review Process" and "Part Two: An Examination of Placement and Visitation," Youth Law News, Vol. XXIV, No. 1, January-March, 2003.
[2] Ibid.
[3] Ibid.
[4] Mark Testa, et. Al, *Can AFCARS be Rescued?* (Children and Family Research Center, School of Social Work, University of Illinois at Urbana-Champaign, March, 2008). Available online at http://cfrcwww.social.uiuc.edu/pubs/Pdf.files/CAN_AFCARS_BE_RESCUED_final.pdf
[5] For details, see NCCPR's report on Kansas child welfare, available here: https://nccpr.org/wp-content/uploads/2025/04/kansas.pdf
[6] *Polling, A Guide for the Statistically perplexed*. See also, *So How Come a Survey of 1,600 People Can Tell Me What 250 Million Are Thinking* These sites (no longer available online) actually suggest the margin for error is even worse, but they involve polls, where some people don't respond. In theory, there is a 100 percent "response rate" in a casereading (though in fact, some case files are so inadequate they sometimes can't be used).
[7] Grimm and Hurtubise, note 1, supra.
[8] Ibid.
[9] Ibid.
[10] Ibid.
[11] Ibid.
[12] Encarnacion Pyle, "Child-welfare system found lacking," *Columbus Dispatch*, February 27, 2003.
[13] Stephanie Erickson, "Child Advocates Brush Off Report," *Orlando Sentinel*, August 24, 2003.