ChapinHall at the University of Chicago
Policy research that benefits children, families, and their communities

Finding the Return on Investment: A Framework for Monitoring Local Child Welfare Agencies

Fred H. Wulczyn
Britany Orlebeke
Jennifer Haight

2009

child welfare

Finding the Return on Investment: A Framework for Monitoring Local Child Welfare Agencies

Fred H. Wulczyn
Britany Orlebeke
Jennifer Haight

Center for State Foster Care and Adoption Data

Chapin Hall at the University of Chicago

Recommended Citation

Wulczyn, F. H., Orlebeke, B., & Haight , J. (2009). *Finding the return on investment: A framework for monitoring local child welfare agencies.* Chicago: Chapin Hall at the University of Chicago

ISSN: 1097-3125

Chapin Hall
at the University of Chicago
1313 East 60th Street
Chicago, IL 60637

773-753-5900 (phone)
773-753-5940 (fax)

www.chapinhall.org

Acknowledgments

This report was made possible through collaboration with Casey Family Programs, Seattle, Washington. We are very grateful for their support.

We are also grateful to Lijun Chen who helped with the data analysis presented throughout the paper. We also want to recognize the many colleagues who have attended Advanced Analytics, also done in collaboration with Casey Family Programs. Advanced Analytics is a week-long seminar where the ideas presented here have been tested and applied. The feedback offered during those sessions has been particularly valuable. Finally, we want to thank Annie E. Casey Foundation and the state members of the Center for State Foster Care and Adoption Data. Without their ongoing support and commitment, the work of the Center would not be possible.

Table of Contents

Introduction

From year to year, child welfare directors allocate resources in the hope that their efforts will improve children's outcomes. Understandably, it is a difficult task. Most observers acknowledge that there are relatively few evidence-based practices available to the field, so connecting a specific service with a given effect is difficult. Moreover, any child welfare system is subject to external forces (e.g., economic conditions) that drive demand for services but are otherwise beyond the direct control of the system. Small wonder, then, that it is so hard to trace whether and how a given investment makes things better for children, families, and their communities.

Fortunately, with the help of the federal government, states have invested significant resources in the sort of information technology needed to run a smarter, more accountable child welfare system. In addition, science has made real progress when it comes to measuring change in complex systems. There is certainly much more to learn, but for now the main challenge has more to do with making better use of the information we already have so that policymakers, practitioners, advocates, and families themselves have a clearer sense of how the system is doing.[1]

In this paper, we present a framework that state and local child welfare agencies might use to monitor their return on investments in child welfare services. The focus is on outcomes within the traditional child welfare system, which covers child maltreatment and foster care. The goal is to burrow through the complexity that goes along with trying to understand whether system performance is improving and whether the improvements are connected to changes in how resources are invested. In this context, *resources* refer to what workers do (e.g., practice model

[1] The goal here is not to suggest that the information contained in an administrative data system is all the data anyone needs to study to know the child welfare system. Clearly, there is a substantial need for research, including experimental research, beyond what one can hope to accomplish using electronic records. The point here is simply to note what many administrators already suspect: given the investment in child welfare information systems, the field ought to be able to get more knowledge out of the data.

changes) and how resources are deployed (e.g., investment in community-based services or improvements in the quality of services).

The paper is organized into the following sections: First, the fundamental question for any human services system is whether clients—children and families in this case—are better off for having received services. Mission-critical outcomes are the *sine qua non* of any investment strategy. Investments of tax dollars may yield a variety of secondary benefits, but without a strong, unambiguous link to mission-critical outcomes, it will be hard to build a strong case for future investments. For this reason, this framework starts with a statement about the outcomes that are central to the child welfare system.

The second section focuses on the process and quality of care as they relate to agency management in a strategic sense. If outcomes represent the goals toward which agencies are working, then the process and quality of care are the main levers administrators have when it comes to changing what the agency is doing. Management involves development of and support for process and quality standards that are known (or at least thought) to influence outcomes, actions that include the development of organizational structures, human resources, and fiscal resources (or what might be called organizational capacity).

The third and final section focuses on measurement, with specific emphasis on core outcomes and the problem of detecting change over time. Although most states have invested in data-collection strategies, the challenge of converting that data into usable knowledge about what is happening to children and families inside a complex service delivery system remains. On the one hand, the problem is statistical in nature. On the other, it is a question of assembling information in a logical manner so that the past, present, and future can be aligned with the mission of the agency.

Mission-Critical Outcomes

The central or mission-critical outcomes are the centerpiece of any return-on-investment calculation. Outcomes are, or should be, the target of investment and movement in the outcomes (better or worse, up or down), all else being equal, indicates whether the investment is paying off as intended. For the child welfare system, child well-being is what matters most. That said, it is important to be circumspect about what is meant by the term *well-being* and the range of responsibilities that fall to the child welfare system in U.S. context. In *Beyond Common Sense*,

the authors make a case for changing how the child welfare system thinks about well-being, moving away from a frame that juxtaposes safety and permanency along with well-being, to one that embeds safety and permanency within notion of child well-being.[2] The argument presented has two parts. First, well-being, as a developmental construct, is far-reaching in that it encompasses physical, cognitive, social, and behavioral dimensions of how children are doing. Although there is a natural affinity for a broad notion of well-being, policymakers and others guiding the child welfare system have to be careful insofar as, absent safety concerns, the present-day child welfare system does not have a specific mandate to serve children who are not progressing in school, for example. We may hope to one day have such a system, but that day is in the future.

Nevertheless, the child welfare system does have a responsibility for assuring that children are living in safe and stable families, which is the second point of the argument. In developmental terms, the chances a child will do well in school are enhanced when he or she is living in a safe and stable family. If a child has behavioral problems that are being addressed by other professionals, safety and stability within the family are protective with respect to whether interventions will work as well as they should. In short, safety and permanency are integral components of well-being, not two discrete legs of a three-legged stool that treats well-being in a somewhat disconnected way. By treating well-being as the multidimensional construct it is, safety and permanency outcomes are placed alongside a broader set of responsibilities the child welfare agency *shares* with the health care, behavioral health care, and educational systems (to mention a few partner agencies).

If safe and stable (or permanent) families are the child welfare system's contribution to a child's well-being (i.e., the well-being outcomes of prime importance to child welfare administrators), what implications does the reframe have for how we ought to think about outcomes, especially if we want to retain a link between the child welfare system and a more holistic interpretation of the term well-being? The answer to this question depends on a more careful, intentional distinction between outcomes, the process of care, and quality of care. To proceed, the first step is to discuss

---

[2] Wulczyn, Fred, Richard Barth, Ying-Ying Yuan, Brenda Jones Harden, and John Landsverk. (2005). *Beyond Common Sense: Child Welfare, Child Well-Being, and the Evidence for Policy Reform.* New Brunswick, N.J.: Aldine Transaction.

specific outcome measures for safety and permanency and to follow that with a more pointed discussion of the process of care and care quality.

Safety and Permanency Outcomes

The purpose behind this framework is to help public child welfare agencies think about the return-on-investment question. Outcomes improve if investments in policies, programs, and interventions pay off. Fortunately, the outcomes the child welfare system cares most about are relatively straightforward, unlike the measurement and analytical challenges that are the subject of this framework. For safety, the primary question is whether or not the child is a victim of maltreatment.[3] Primary prevention, to the extent a child welfare system makes those investments, should reduce the incidence of maltreatment. If a child has been victimized, the question facing most child welfare administrators is whether services provided reduce the recurrence of maltreatment. Agencies should invest more in services that reduce recurrence and less in services that do not.[4] For permanency, the outcomes have to do with whether the child welfare agencies support families well enough to avoid placement in out-of-home care (i.e., preserve the protective capacity of the family). If not, and placement becomes necessary, then attention shifts to placement stability and to whether the child returns home or leaves foster care to live with relatives or an adoptive family. Because out-of-home placement is a temporary solution, how long it takes a child to achieve permanency is also an important consideration. Finally, because it is a signal that problems within the family persist, reentry to foster care matters too.

In summary, high-level form, the outcomes of interest can be arrayed in the following way:

1. Safety

[3] Safety—measuring maltreatment—is a prime example of why one does not want to trivialize the measurement issues. Official maltreatment reports (i.e., the data collected as part of a state's reporting system) as a source of data for measuring maltreatment or victimization are to some extent limited. However, this is a detection problem as opposed to the kind of measurement problem described here. Given an adequate detection mechanism (i.e., one that reliably identifies incidences of maltreatment), the task of measuring incidence and recurrence rates is fairly straightforward although not without complexity.

[4] "Easier said than done" applies here as well. The *what works* question is tough to answer in controlled settings let alone in the dynamic context of a public child welfare agency.

a. Likelihood of maltreatment (measured as incidence rate per 1,000 children)
    i. In the general population
    ii. During the time in-home services are being provided
    iii. During the time a child is in out-of-home care
b. Likelihood of recurrence

2. Permanency
    a. Likelihood of placement (measured as incidence rates per 1,000 children)
    b. Likelihood of permanency and the timing of the exit
    c. Likelihood of nonpermanent exit and the timing of those exits
    d. Likelihood of reentry and the timing of the return
    e. Placement stability and the timing of the moves relative to entry and exit
3. Health, Education, and Mental Health
    a. Appropriate connections to and follow-up with health care services
    b. Appropriate connections to and follow-up with educational supports
    c. Appropriate connections to and follow-up with behavioral health care

With respect to how a state might monitor its investments in child welfare programs, these outcomes and their measures differ from the outcomes used in the federal Child and Family Service Reviews (CFSR) in several ways. First, the measures proposed here include maltreatment rates (1.a.i) and placement rates (2.a), which are intended to reference per capita measures of how often maltreatment and placement occurs (e.g., the maltreatment rate and the placement rate per 1,000 children). The inclusion of per capita measures is meant to capture the idea that from a public health perspective, investments in child welfare services should at some level pay off in the form of lower rates of maltreatment and placement overall, notwithstanding

the obvious measurement problems.[5] Whether these aims are achieved through primary prevention or some other strategy, knowing the basic rate of maltreatment and placement is an essential component of any state investment strategy.

The permanency measures offered here also differ from the CFSR measures in that nonpermanent exits are included explicitly. Children leave foster care for reasons other than permanency (i.e., reunification, adoption, or guardianship, including subsidized guardianship), with running away or transferring to other child-serving systems among the most common of those reasons. Indeed, for children who enter care as older adolescents (e.g., 15- or 16-year-olds), running away is often the single most common reason they leave placement, so excluding an explicit measure of nonpermanent exits makes little sense when the overall goal is to judge how well the system is serving *all* children.

The list of permanency outcomes also differs in that measures for adoption, reunification, and guardianship are not identified separately. The decision to bundle the different types of permanency into a single measure is based on a particular interpretation of the core mission of the child welfare system. First and foremost, the system should increase the chances any given child will achieve the permanency outcome that *best fits* that child's needs and the family's circumstances. In the aggregate, separate goals for adoption, reunification, or guardianship create worker-level tensions that place different permanency outcomes in competition. More important, having separate permanency goals implies that we know what the optimal mix of permanent exits is for a given group of children in advance of working with their families. Bundling the permanency outcomes into a single measure does not mean that an administrator or anyone else should not know how permanency exits are distributed. It is always possible to monitor permanency exits as distinct outcomes. If the system evolves so that caseworkers are left to make permanency choices based on the best interests of the child relative to the capacity of their family within the context of their community, then the observed mix of permanency outcomes will reflect the optimal mix rather than a predetermined but clinically uninformed expectation.

---

[5] The measurement issues were noted in an earlier footnote. Official reports to a state's abuse and neglect reporting system may not capture the true incidence of maltreatment. However, official reports do convey information about the underlying system.

The last difference to note relative to the CFSR is the explicit distinction made between likelihood and timing. As already noted, one goal of the child welfare system is to make sure children placed in foster care leave placement for a safe and stable family. All else being equal, the child welfare system ought to achieve permanency for all children placed. Moreover, permanency ought to happen quickly, given the circumstances of a particular family—that is, children should not languish in care even if their eventual return home is certain.

Therein lies the distinction between likelihood and timing. It is within the realm of possibility (albeit at the extreme) that although every child who enters foster care leaves to one form of permanency or another, it takes an inordinately long time to do so (e.g., an average length of stay of 6 years). In situations such as this, one could say that although the likelihood of exit to permanency is high (100% likely), exiting takes too long. In contrast, another possibility is that the likelihood of leaving to permanency is relatively low (e.g., only 40 children out of every 100 admitted leave to permanency), but the time needed to achieve permanency is short (e.g., 250 days). The point is, the length of time it takes to achieve permanency is a separate issue from the likelihood that children will reach permanency. It cannot be assumed that if permanency happens *quickly* it is also more *likely* to happen, or that if permanency happens *slowly*, it is *unlikely* to happen. *Unlikely to happen* is often paired with slow (and *likely to happen* is often paired with fast), but the co-occurrence is an empirical question to which the answer differs from jurisdiction to jurisdiction and from population to population within a jurisdiction. Measurement systems that fail to appreciate this distinction offer less value than those that do.[6]

Also on the list of permanency outcomes are those related to health, education, and behavioral health. As broader concerns tied to the overarching construct of well-being, health, education, and behavioral health are a clear and important concern for the child welfare system. How those concerns are operationalized is a significant matter. In a traditional sense, improving outcomes

---

[6] One example of a failure to distinguish properly between timing and likelihood concerns the adoption of African American children. In some jurisdictions, the fact that African American children leave more slowly to adoption than white children has been interpreted to mean that African American children are also less likely to be adopted when in fact they are more likely to be adopted. One needs to know how timing and likelihood combine to generate a given exit pattern. Without that clarity, policy and practice choices are obscured. It is one thing to increase the likelihood of adoption, which one can do without increasing the speed of adoption (i.e., the average time to adoption); it is another thing to increase the speed of adoption without increasing the likelihood; and, it is quite another to try and do both simultaneously.

means changing well-being, as in improving general health or addressing more chronic health conditions so that there is a measurable improvement in the health of an individual. The same holds true for educational outcomes. Improving reading ability and math skills is an important educational objective and outcome.

The question for the child welfare system is how to engage these challenges in the context of accountability. Current practice generally calls for child welfare agencies to make sure children get to the doctor, go to school, and get mental health assessments as needed. As described below, these tasks are really about the process and quality of care, as opposed to specific outcomes that are measurable as a change in health (or educational status or other measures of well-being). Making sure a child goes to school and does well there fits with a broader set of responsibilities, even as the system searches for a way to share responsibility for specific outcomes with the school system. The same tensions arise when dividing responsibility between the child welfare system and another allied service system (e.g., health care).

The Process of Care

The *process of care* refers to the steps followed during the time family members/children receive services. It is conceptualized as a series of activities or events that form the service path through the child welfare system (i.e., trajectories). Although the details that define the process of care may draw on a particular model of practice, a particular intervention, state regulation, or agency practice, any given process has common elements or requirements:

a. Referral, intake, and assessment

The process of care is initialized at the point of first contact (or *inception*, which is a term that will be used later). By necessity, the process consists of a referral mechanism that is used to manage the manner in which a service provider comes to know of a child or family; a procedure for bringing the client into services (i.e., the first contact); and, a procedure for conducting an assessment that is then used to inform what happens next.

b. Treatment planning and the linkage to services and interventions

The referral, intake, and assessment phase leads to a plan that organizes how a service provider will engage the family vis à vis the match between what the child/family needs and what the

provider has to offer. Ultimately, the plan has to link the clients to services/interventions that are designed to address the needs and strengths that were identified during the assessment. At this point, services are provided; that is, the process of care is defined by the requirements of a specific intervention (i.e., cognitive behavioral therapy, multisystemic therapy, home visitation). In general, the better (i.e., more effective) interventions have a specific protocol that is followed in the course of delivering the intervention.

c. Reassessment, discharge planning, follow up, and case closure

In the same way that children and families are brought into the service system, a complementary discharge process governs when services end. Preparation for drawing services to a close begins with a reassessment of needs, at which time the question at hand focuses on resolution—have the treatment objectives been met? If not, the reassessment leads to review of the need/service match. If so, the process of care shifts to discharge planning, clinical follow up, and case closure.

Each step of the process is (or should be) guided by a protocol that outlines the specific activities that reflect a *best practice* or *model of practice*. That is, to the extent that there is a preferred way to conduct an initial assessment, the model of practice would articulate what those steps are. In some cases, the actual requirements that define a particular set of steps are defined in statute or regulation as is the case, for example, with CPS investigations and the requirement that an investigation be completed within a prescribed period of time.

The process of care has an analog at the system level in that a standard process of care means that the necessary structures/capacity have to exist within the service system. These functions are found within the structure of the system as a whole and within the subsystems that make up the system at large. The system as a whole has to demonstrate the capacity to bring clients in, assess their needs, deliver services, and close the case once the issues have been resolved. The same is true for network agencies providing specialty services. The network works best when service connections work smoothly in relation to each other.

Quality of Care

Although the idea of quality care has its own particular resonance, the reality is not quite so straightforward. In everyday language, *quality* refers to how well something is done. One can think of quality as craftsmanship—attention to the details that differentiate the exceptional from

the ordinary—but craftsmanship is a hard thing to measure. In practice, process and quality are closely aligned in that adherence to the process of care is in and of itself an indicator of quality, especially if the underlying process protocols are supported by an evidence base that links the process to outcomes.[7] Quality of care is also inextricably linked to outcomes, but the relationship is complicated because quality improvements need not beget better outcomes. High-quality care that fails to achieve outcome improvements is problematic from a resource perspective because the resources allocated do not produce better outcomes (i.e., a return on investment), notwithstanding the impact quality has on the overall client experience.[8] In short, quality, process, and outcomes are interdependent factors that lend themselves to judgments about how well a system is doing.

With all of that noted, quality care takes shape around the following dimensions, as a starting point:

1. Human resources
    a. Staff members are appropriately trained for the jobs they are asked to do.
    b. Staff members use appropriate tools and other supports to carry out their work in the context of a process of care.
2. Physical plant and equipment
    a. Are buildings family-friendly and well maintained?
    b. Are buildings accessible to families (e.g., when are services offered)?
    c. Are programs based in the community?

[7] As noted earlier, the process of care, as a component of quality, may be outlined in law or regulation. However, in some cases, it is unclear how specific process and quality requirements are related to outcomes.

[8] Quality of care, as it relates to the experience of clients, is an important issue but since the focus here is on outcomes, less time is spent addressing related concerns other than to note that consumer/client feedback is an important source of information about how well an agency is doing. As such, it is one measure in the constellation of measures an agency needs to be fully informed.

d. Do employees have the resources they need to do their jobs (e.g., phones, computers, cars)?

3. Practice protocols

    a. Staff understand and follow best practices, evidence-based interventions, or other practice protocols, as indicated, in their direct work with families (and in accordance with their job description).

    b. Families are engaged in all aspects of their care including decision making.

        i. Families are treated with dignity.

        ii. Cultural awareness, knowledge, attitudes, and skills are applied to work with families and children.

    c. Services/interventions/protocols are developmentally appropriate.

4. Consumer feedback is actively encouraged and sought out.

5. Agency management and clinical supervision are tied to outcomes and practice model fidelity.

Although the listed dimensions are meant to provide concrete rather than abstract characteristics of quality, it is easy to see why quality is so hard to define. Operational definitions of *family friendly* or *appropriate level of training* are hard to come by without reference to some sort of standard. The importance of an external standard is reflected in how the Council on Accreditation defines quality: "the extent to which contemporary and generally recognized standards are met and exceeded, and desirable outcomes achieved."[9] Other dimensions may fit easily within a local context and there is no reason to treat the list above as exhaustive of the possibilities. Perhaps the main point is that a given jurisdiction must be deliberate in its attempt to articulate what quality means, with particular reference to process and outcomes.

[9] From http://www.coaafterschool.org/glossary.php#gq

Monitoring Performance

The return-on-investment strategy links investments in process and quality (otherwise thought of as interventions, programs, and the organizational structures needed to support the underlying functions) to outcomes. Although the implementation issues are by no means trivial, the focus on process, quality, and outcomes does draw the monitoring task into sharper focus. If process and quality are connected to outcomes, then one has to establish fidelity with process and quality requirements, measure outcomes, and test whether process and quality investments produce better outcomes.

In this section, we highlight issues that arise when setting up a system to evaluate performance at the local level. The goal is to explore how one might build a body of evidence that connects process and quality to outcomes.

Monitoring performance starts with data collection, and that is where this discussion begins. The topics covered include sources of data and sample selection, within the context of the question being asked. Beyond making a few relatively simple suggestions, though, the material covered is very basic and intended to guide the collection of data. The next section addresses the issue of nomenclature with particular reference to the distinction between standards and baselines. As discussed, *baselines* seem well suited to monitoring outcomes, whereas *standards* appear better suited to tracking the process and quality of care.

Data Collection

Performance monitoring is a data-intense undertaking that depends on having data that corresponds to the outcomes of interest as well as the process of care, the quality of care, and agency management practices. Sources of data are fairly obvious: administrative data (including electronic clinical records), case records, interviews with service participants (including workers, families, and collaterals), focus groups with stakeholders among others, and shadowing of workers as they carry out their work. Each source of data has strengths and weaknesses relative

to what can be learned about outcomes, the process of care, and quality of care. Combining data from various sources can fill the gaps in any given data source, provided one is careful. Among the pitfalls, sampling and sampling bias are perhaps the most important, so the discussion that follows focuses on sampling issues.

Sampling

Because it is impossible to observe everything taking place in the child welfare system, sampling is an important feature of any data collection plan. To keep matters simple, it is important to have a statistically valid sample of records. Although sampling is a technical subject in its own right, there are several ways to approach sampling that help to simplify matters, at least somewhat.

Performance Window

Because performance occurs in time, a contemporary view of performance means that a performance window has to be selected in which to view what is taking place. The performance window, a simple idea in concept, has rather profound implications for how samples are drawn. First, point-in-time samples are not especially helpful because of the way the sample freezes time at a specific moment. Second, most of the work a child welfare system does in any given year (or month) involves children and families known to the system prior to the start of the window.[10] Relative to children who start services during the window, we can expect substantial differences in the makeup the populations and the expectations as to what will happen during the performance window. From a process-of-care perspective, the notion of the performance window means that entry processes (referrals, assessment, etc.) will only be observed for those entering care during the window. However, it will be harder to observe the exit process for children starting care during the window because their exit is more likely to occur after the window of observation has closed.[11] The opposite is true for the children already in the system—they

[10] The importance of this observation cannot be overstated when it comes to monitoring performance. Take maltreatment data as an example. According to NCANDS (2006), in some states fewer than 50 percent of the maltreatment cases reported during the year were new reports (i.e., children not previously reported). From a performance perspective, to the extent that there are differences that distinguish new cases from others, those differences have to be accounted for in the way performance expectations are developed.

[11] The precise extent to which this is true depends on the process being followed. The investigation process—what happens when a child/family is investigated for maltreatment—typically happens in a relatively short period of time, so

entered the system prior to the start of the performance window. More children/families leaving care during the performance window will come from the population of clients already known to the system. Both situations are prone to selection biases that influence what one can learn about performance and how one generalizes to the population of *all* children served by the child welfare system.

The solution to this problem is to sample admission cases and what might be called *legacy cases* (i.e., children already in care at the start of the performance window) separately, an approach that stratifies the sample, as described below. Each sample is then followed prospectively through the performance window to understand how the process of care unfolds as it unfolds. With respect to the legacy sample, it is important to remember that, as children/families who are still in care relative to the children/families they entered with, the legacy sample is only representative of the children in care on the given date, so the findings cannot be generalized to all children.

The legacy sample may be further stratified by the time spent in care prior to selection into the sample, which increases the probability of observing certain types of exit events. For example, the likelihood of exit by way of adoption is more likely in the next year for children already in care for 2 or more years than it is for children in the legacy sample that have been in care for less than 2 years. The opposite is true for reunification.

Stratification

In contrast to a simple random sample, a stratified sample divides the population into discrete categories (i.e., mutually exclusive and exhaustive) prior to drawing the sample. Stratified samples are especially useful when the subpopulations differ significantly as is often the case

---

one might expect to see both the beginning and end of the process within a single performance window. However, if one wants to study recurrence, one might have to wait up to 2 years before a second investigation is recorded.

within the child welfare population. The benefit of stratified sampling stems from the fact that stratification reduces sampling error.

In the example above, the legacy and admission cases represent one approach to stratification. Age stratification is another useful variable. Case type (e.g., preventive services, protective services, foster care) is yet another useful way to stratify the sample, although care has to be observed when using case type. The main problem has to do with the question being asked and whether the stratification uses an outcome (e.g., children reunified or adopted) or some other type of dependent variable. For example, if placement in foster care is the stratification variable (i.e., the sample includes only children placed in foster care), the sample should not be used to study the benefits of preventive services in relation to placement because there are no children in the sample that have not been placed. Without variation in the dependent variable (i.e., children placed and children not placed) there is no way to understand whether the process of providing preventive services reduced the need for foster care.

To avoid such problems, it is best to follow the process of care from the point a sample is drawn going forward in time rather than backwards. For example, a sample of children who have been reunified might be used to study the process of providing services after reunification and their impact on reentry.[12] A sample of reunified children will be less useful when trying to understand the processes that led to reunification because there is no way to know to what happened with children who were not reunified, which is what one needs to know in order to determine whether what happened during the time a child was in care influenced the likelihood of reunification.

Inception Cohorts

Although stratification by legacy vs. admission cases is an important strategy for managing the sort of case history variability one is likely to encounter when dealing with a performance window, one additional sampling strategy has longer-range importance. The details involve further stratification of the admission group as follows. During a given year—the performance

[12] Even in this example, one has to be careful about sampling bias. In a sample of children who were reunified, one is likely to find a higher proportion of children who stayed a relatively short time in foster care. Judgments about reunification and reentry could reflect the weight of the short stayers, in the event short stays and reentry are related. To overcome the sampling bias, one could stratify reunified children by their prior time in care. That said, one has to always be aware of the selection effects within any given sample.

window—children starting care can be divided into two groups: children with no prior history ever and children with some prior history who were not active cases at the start of the window. The former is known as an *inception cohort*. From a monitoring perspective, inception cohorts are important because they offer an opportunity to monitor the process and quality of care offered to children at the time their service history starts (i.e., the point of initialization), during the same policy and practice context. Children with prior service experiences are members of earlier inception cohorts. The care they received at inception, depending on how long ago care started, would reflect protocols in place at that time, which may not be the current protocols. Inception cohort members who return to care at some later date probably differ from other members of their inception cohort in ways that are clinically important in the present context. In short, the members of this year's admission group who are members of a prior inception cohort (and multiple inception cohorts may be present) are probably different from this year's inception cases. These differences ought to be understood when trying to assess whether the care delivered was appropriate.[13]

Ideally, a review sample would isolate the current inception cases because they provide the clearest opportunity for understanding how the process and quality of care influence outcomes. In the most general sense, this means isolating children without any prior contact and following them prospectively.[14] The difficulty inception cohorts create has to do with how few children proceed further into the system from the point of inception. For example, of the children investigated for abuse for the very first time, most will have no further contact once the investigation has been adjudicated. Even if the allegation that led to the investigation is substantiated, the likelihood that the child will be referred for services is less than 50 percent. The likelihood of subsequent contact (in the form of a second maltreatment investigation), even though services were not offered, is still small. One way to mitigate this difficulty is to select cases based on first-ever investigations that were substantiated or first-ever substantiated cases

[13] Appendix A provides an example of how the group of children with a child welfare event (i.e., an unsubstantiated maltreatment investigation, a substantiated maltreatment investigation, or a placement in foster care) in the year 2000 compare with respect to whether the event in 2000 was the first ever or whether the event in 2000 was preceded in time by at least one other event.

[14] Naturally, inception is something that is identifiable at the child level or the family level, but the meaning is the same.

that were placed into foster care. These samples cannot be used as easily to monitor the process of care further upstream, but they are superior to samples that do not differentiate cases based on their prior case history.

Process Samples vs. Child Samples

One of the main problems in understanding outcomes, the process of care, and the quality of care has to do with how long a family is engaged with the child welfare system. For many children, their connection to the child welfare system is brief. However, for a nontrivial fraction of children, engagement with the system may last years. From a monitoring perspective, it is hard to follow the life of the case and maintain a connection between what is learned from the care provided to a given child and contemporary practice. Again, the issue is tied to when a child starts care. A simple random sample of children in foster care may produce sample members admitted to foster care 2 or more years prior to the review period. If the sample is being used to understand the process/quality of care at intake (or inception), the information gathered about practice may be dated and cannot be used to draw inferences about what is currently true. For the same reason, it is hard to use such samples to understand whether recent changes in practice are being followed and their effect because members of the sample will have entered service prior to the onset of the changes and will have no exposure to the practice modification.

One remedy is to consider *process-based sampling*—that is, rather than follow a sample of children and the care they receive over the life of the case or through the entire performance window, the sampling plan would focus on critical points along the process continuum. For example, if one were interested in fidelity with respect to a new referral model, the sample would be based on children whose entry into the system would have occurred after the model was implemented. For these children, one would ask if the process was followed and the quality of the work was acceptable. However, as for what happens next in the process, a different sample of children would be selected.

There are several advantages to using process-based samples. From the case review perspective, the focus is on a relatively small slice of the service history. When children have been in the system for a long time, case files may have several volumes and a coherent case narrative can be hard to develop. More important, specific elements of the process may be hard to untangle from the other details. A focus on a specific point along the process continuum makes the review process a bit more incisive with respect to all of the details in a case file. Process sampling also

has the advantage of being more contemporary in that new practice changes can be studied in something resembling real time. As noted earlier, one of the main challenges that goes along with monitoring the performance of the child welfare system is understanding what is true today relative to changes that were made yesterday. Very few children who come into contact with the child welfare system are exposed to each point along the process-of-care continuum. For those who are, it may take many months— if not years—to complete a full cycle of care. Seeing how changes early in the process influence what happens later is bound to this natural cycle. A process-based sample would not be as dependent on whether children experience the full service range, which means that elements of the process can be followed contemporaneously.

Of course, looking at discrete steps in the process makes it harder to connect what happens along the chain of events that defines the process of care to outcomes. Because it is important to maintain the connection between process and outcomes, exclusive use of process-based sampling is unwise. Nevertheless, a narrow focus on discrete points along the continuum of care solves certain problems that are critical to understanding whether changes to the process are being implemented faithfully. Use of a process-based sample also depends on having a clear understanding of process and quality, a requirement that may have ancillary benefits if it means developing a deeper appreciation for the local practice model and its relationship to outcomes.

Mind the Gap: Standards and Baselines

Performance monitoring is really an exercise in gap analysis. A practice model (or a set of administrative procedures) sets forth a set of steps (process and quality expectations) that caseworkers should follow in the course of their work with a family. Process and quality requirements may vary with the practice model (or the body of regulations), but the expectations are embedded in the model (or at least should be). Monitoring is used to determine whether the work carried out followed the steps in the manner and order prescribed. If not, the gap between what did and should happen has to be addressed by whatever means are appropriate given the reason(s) why the gap is present. In the case of outcomes (e.g., how long does it take to get children home), the gap analysis follows the same logic. To the extent that a gap exists between the expected and the observed outcome, the gap has to be closed, assuming the gap exists for reasons that are amenable to administrative or clinical action, a point that is emphasized later in the paper.

Gap analysis implies having a sense for what would be true—a *standard*—if the work done on behalf of a client met all expectations. That a gap might exist suggests that there is variation in practice; some steps in the process are not followed, the sequence of steps in the process is not followed as prescribed (i.e., the steps are out of order), or some quality standards are not met. In each case, the issue is *fidelity* with respect to process and quality standards; by extension, the goal of monitoring is to establish variability relative to the standard.

Within this context, standard has two meanings, both of which are important. On the one hand, the standard refers to accepted practice—what the work looks like when it is done correctly. In this context, it refers to a *standard of practice,* or what a client should expect in the course of receiving services. During a general physical exam, a physician takes a patient's blood pressure, heart rate, respiration rate, and temperature (among other things) because the applicable guidelines indicate that this is standard practice.[15] On the other hand, the word *standard* can be taken to mean the level of practice variation one is willing to tolerate. If a process consists of five steps (i.e., the standard of practice), there may be an expectation that the process as prescribed will be followed 95 percent of the time (i.e., the standard of performance), extenuating circumstances notwithstanding. The 95-percent standard is the expected; the *observed* indicates how often the standard of practice is met; the gap is the variation between the observed and the standard of performance. For individuals charged with moving an agency forward with respect to their organization's capacity to improve outcomes, observing variation in the process of care, the quality of care (i.e., the elements of the practice standard), and outcomes are of vital importance. If, for example, patterns of variation in the process and quality of care can be linked to outcomes

[15] There are at least two important additional comments to make here. First, each element of a process has nested within it sub-processes. For example, home visits are a standard practice when a child is placed in foster care. Monitoring would detect whether the home visits are occurring. Home visits consist of other processes including risk and safety assessments, which have their own protocol. In the context of a home visit, a monitor has to ask whether the assessments were completed and whether the assessment was carried out in the manner expected. Second, whether these steps or procedures are in fact the best steps or procedures given the purpose of the visit is, of course, subject to review and revision. Process and quality standards have to be examined periodically so that the practice fits with the intended outcomes. Refreshing practice standards relative to outcomes is an important part of monitoring and evaluation and a key ingredient in the return-on-investment calculus.

such that certain care patterns are routinely associated with positive outcomes, then one has a basis for advancing promising practices.[16]

Standards are often used along with baselines. According to the National Institutes of Health, a *baseline* is the time point just before an intervention when starting measurements are taken (2007).[17] These starting measures are then compared with later time points to establish the pattern of variation relative to the starting point. This use of the term *baseline* has meaning in a clinical context (e.g., the resting heart rate serves as the baseline for comparison with the heart rate following exercise) and in the context of organizational performance improvement. In the latter case, the baseline can be used to describe current performance such that, when compared with performance following some sort of intervention (changing the way services are provided), one has a sense for whether the intervention induced the intended performance changes. Used in conjunction with standards, a baseline may also point to process and quality improvements that are needed in the first place.

With respect to the process of care, the quality of care, and outcomes in the child welfare system, standards and baselines have to be used with some understanding of how each fits in the context of child welfare. Standards, as in standards of practice and standards of performance, have clear applicability to the process and quality of care, provided there is a link to outcomes. In other words, investment in improving the process and quality of care so that process and quality meet applicable standards is critical to the overall sense that the child welfare system is doing the best job it can. Accountability mechanisms geared to knowing whether applicable standards were followed is the essence of performance management because without fidelity to the underlying model of practice, we cannot determine how the process is connected to outcomes. Without a direct connection to outcomes, the logic of performance improvement breaks down almost entirely because the means become the ends.

---

[16] It is important to remember in this context that because the connection between process and quality standards and outcomes is based on observational data, the quality of the evidence behind the promising-practice assertion is relatively weak. It is important, therefore, to further test the connection with more robust methods that use more carefully constructed counterfactuals such as an experiment that uses random assignment. Overall, it is fair to say that too little work connecting process and quality standards to outcomes has been done.

[17] See http://www.niaid.nih.gov/factsheets/Glossary.htm

Outcome standards are another matter. Process and quality are for all intents and purposes on the input side of the input/output equation. By placing process and quality standards on the input side, we are implying that we have a measure of control over adherence to the standards (i.e., fidelity). We can and should expect a certain level of fidelity, particularly if there is a strong relationship between a process/quality standard and an outcome. Indeed, if there is a one-to-one relationship (a given process has a defined benefit, which is rare), then there is no acceptable reason to deviate from the standard of practice.

Considering outcomes (e.g., the likelihood of and time to reunification) as dependent on a set of inputs implies that getting to the outcome is determined fully by whether the inputs are followed. However, the likelihood of such a direct and unequivocal relationship is low. Even in manufacturing, where processes are highly routinized, one expects variation in the output (i.e., how long a light bulb will last). If that were not true, there would be no need to monitor the output because fidelity with process and quality standards would be enough to know that the desired outcome was achieved.

Child welfare outcomes are much more difficult to achieve. There is a level of uncertainty that mixes with the inputs such that, even if the standard of care is followed to the letter, there is no categorical, unequivocal guarantee that a particular outcome will be observed at either a client or agency/system level. The refrain "we did everything we could, but…" speaks to the real limits of what we can do with the interventions we have. It is not an excuse to ignore the need for fidelity with standards of practice and quality. On the contrary, it is an admission that recognizes the limits of our knowledge and action, a situation that demands more vigilance—not less—in a fair and just system.

How then to think about standards and baselines in practice? Process and quality, it seems, are more amenable to the use of standards. If best practices or regulations stipulate as to a certain process (e.g., maltreatment investigations will be initiated with 24 hours of a report or home visits will occur every 30 days), it seems reasonable to set a performance standard (e.g., the process is followed 95% of the time).[18] Quality standards ought to be treated in a similar fashion. If an

[18] The performance standard ought to be high (90%, 95%, or 100%), given how important the underlying process or quality standard is relative to the outcome. In the case of safety, a safety assessment ought to be done each and every time it is called for because of how important safety is. For the most part, process and quality performance standards

assessment protocol calls for the use of a particular tool by a trained professional, it is not enough to know that an assessment took place (a process requirement). One also needs to know that the assessment was carried out as intended (a quality requirement). Process and quality baselines provide, then, a way of understanding how often the standards are met. In turn, the baseline can be used to determine whether fidelity improves following intervention, as described above.

The use of outcome baselines and standards follows a similar logic, although setting and then enforcing a particular standard is arguably less viable for the reasons set forth above: the uncertainty (i.e., that which is not controllable in a strict sense) is still too great to say what the outcomes for a given population will be. Outcome standards also imply knowledge of what would be true if everything in the system was working according to a well-executed master plan. Knowledge in the field is far from establishing standards with that level of certainty. Instead, the field is more likely to gain from the use of outcome baselines to describe current performance. Then future performance relative to the baseline serves as a way to gauge whether progress is being made toward improved outcomes.

Returning to a point made earlier, caution has to be applied when using standards and baselines in a performance-monitoring and improvement strategy. On the one hand, standards are aspirational in that they reflect what we would accomplish if we could do everything that needs to be done in the way it needs to be done. When compared to a baseline, the standard expresses how far we have to go, assuming there is a difference that places the baseline below the standard. On the other hand, the baseline serves as a reminder of where things stand. Used in conjunction with measures taken in the future, the baseline tells us how far the system has come in the event such a comparison shows improvement. In the end, from the perspective of a continuous quality improvement cycle, the difference between a baseline and observed performance may prove more

---

should be set at very high levels. Put another way, if a process has to be followed just 50 percent of the time, what is the value of the process in relation to the outcome?

useful simply because one does not need a standard to know whether the process is improving or not. Realistic standards indicate whether one might expect or demand more improvement, depending on the circumstances. Whether future improvements are realized will always be understood relative to the past as captured in a baseline.

Case Mix Adjustment

Case mix adjustment refers to the notion that because children or families do not fit a single mold, differences in the clients served have to be actively considered when trying to understand why the performance of one organization differs from that of another or from itself over time.[19] For example, children under the age of 1 at the time they are admitted to foster care are much more likely to be adopted than children admitted after age 15. Whether the differences are too large from a service-quality perspective is an important question, but in all likelihood, differences would remain even in a system that manages to adopt exactly those children for whom adoption was the best outcome. In turn, when agencies (or states or counties) differ in the number of infants served relative to the number of teens, one can expect differences in adoption performance that are due to the population served as opposed to how well the work of adoption is performed.

The methods of case mix adjustment range from the simple to the complex. As a general rule, the more factors one wants to take into account, the more complex the process is because of the need for more sophisticated statistical techniques. Statistical techniques (event history models, hierarchical linear modeling) make it possible to assess outcomes while controlling for many different characteristics of the child, the family, and the service provider. Nevertheless, even simple controls illustrate how case mix adjustment deepens what is known about performance measurement and what can be done to address gaps.

In Table 1 below, two agencies serving two populations are depicted. The average length of stay for each agency and agency/population combination is also displayed. The children served by each agency have the same length of stay; what differs is the mix of children served. In Agency A, 65 percent of the children served are under the age of 5, and 35 percent are over the age of 5. Agency B serves the opposite population: 35 percent of the children are under the age of 5; 65

[19] In this section, risk adjustment and case mix adjustment are used interchangeably.

percent is over the age of 5. Children under the age of 5 have an average length of stay of 500 days as compared to an average length of stay of 250 days for older children. The difference may be due to the fact that the population of younger children is more likely to leave via adoption as opposed to reunification.

Table 1. Case Mix Adjustment When Comparing Two Agencies

	Percent of Caseload		Average Length of Stay		
	Under 5	5 and Over	Under 5	5 and Over	Agency Ave.
Agency A	65%	35%	500	250	412.5
Agency B	35%	65%	500	250	337.5

In this simple example, the unadjusted performance of Agency A, as measured by length of stay, is clearly below that of Agency B. However, the difference is explained entirely by the population served as seen by the fact that the within-population comparison (the so-called apples-to-apples comparison) shows that the agencies performed identically. Agency A, for whatever reason (and the rationale may be entirely deliberate and well-intentioned from a programmatic perspective), serves more young children than the other agency.[20] Failure to take the mix of cases into account increases the likelihood of Type I error (false positive), namely claiming real differences when in fact there are none.[21] To say the least, it would be foolhardy to address changes that do not exist.

The failure to adjust for case mix is equally pernicious when attempting to understand whether an agency is improving relative to itself. In this case, the comparison depends on measuring the outcomes for a single agency at two distinct points in time. Based on the time 2 data, the agency

[20] As for whether differences of this type are likely to be found, consider again the fact that in some states less than one-half of the maltreatment reports in a given year are new reports, whereas in other states the figure is 90 percent. If, as research has shown, the recurrence of maltreatment is related to the number of prior reports, one might expect differences in recurrence rates based simply on the mix of new vs. already recurrent cases. Ultimately, it is an empirical question but case mix differences of this magnitude have to be considered when trying to isolate performance differences from other explanations.

[21] Had the example been constructed in another way—with agency-wide performance equivalent and subpopulation performance differences—it would have demonstrated that Type II errors (false negative) are also possible.

appears to have improved its performance, but the adjusted data indicate that no such improvement occurred. (See Table 2.)

Table 2. Case Mix Adjustment When Comparing a Single Agency at Two Points in Time

	Percent of Caseload		Average Length of Stay		
	Under 5	5 and Over	Under 5	5 and Over	Agency Ave.
Agency A—Time 1	65%	35%	500	250	412.5
Agency A—Time 2	35%	65%	500	250	337.5

Case mix effects are important because false positives (wrongly believing that change did happen) and false negatives (wrongly believing change did not happen) corrode the performance improvement cycle the comparisons are intended to inform. In short, if the process and quality of care are reengineered with the expectation that outcomes will improve, it is essential that alternative explanations for the changes be fully examined. Otherwise, everyone will be left to guess whether their actions helped in the way they imagined.

The use of risk or case mix-adjustment strategies, it should be said, is somewhat more important in the area of outcomes than process and quality because process and quality are generally applied uniformly, regardless of any secondary considerations. Every child/family should have an assessment; every child/family should have a treatment or case plan that addresses any identified needs. The lone consideration with respect to the use of differential process and quality expectations has to do with differential diagnosis and the selection of a particular intervention given assessed needs. If a particular treatment is chosen, then the process and quality standards that apply to that intervention ought to guide the judgments made regarding model fidelity. Within the group of children or families receiving the same set of services, however, it is expected that process and quality expectations would apply without distinction.

In the end, the central question that has to be addressed is this: If there are differences (variation) in outcomes over time based on either the population served or the service or administrative area, is the variation a function of performance or is some other factor in play? The failure to adjust for case mix implies a belief that all or virtually all observed variation is a function of performance that is in turn amenable to process and quality improvements. The reality is that this is extremely unlikely. Even in child welfare systems where the average length of time in foster

care (or some outcome of one's choosing) is low relative to other states or localities, there is considerable variation from one population to the next. Simply stated, not all variation in outcomes is a function of performance, and child welfare has to do a better job of distinguishing one source of variation from another.

Case Mix Adjustment and Baselines

It is important to recall that case mix adjustment strategies provide vital information that can actually accelerate the improvement process because the insight developed can be used to plan service investments. It is unlikely that every agency, public or private, does everything well or everything poorly. It is more likely that providers of child welfare services do some things well and some things less well. It may also be the case that providers work better with some populations than others. Case mix adjustment provides a way to understand performance in a far more textured way so that agency strengths and weaknesses are more readily identified. It makes targeted investments in process and quality more likely and allows scarce financial resources to be invested more wisely and improvements more easily tracked.

The ideas presented in the discussion above are captured in Table 3 below.[22] The data displayed show outcomes for any child admitted to foster care between 2000 and 2006. The data are stratified by age at admission in order to account for case mix variation.[23] Two geographic areas are displayed: the left panel shows the state data; the panel on the right singles out a county within the state. The outcomes are calculated based on the status of the child at the end of the calendar year following the calendar year of admission.

Reading across the rows shows if and how children leave foster care. Children *still in care* means that children had not yet exited at the point the data were summarized. All other children exited somehow, either by permanency (reunified; relatives, which includes guardianship; and adoption) or nonpermanency, which includes exit types not shown separately (listed as *other exits* in Table 3). Reading down the rows shows how the likelihood of a particular exit changed over time.

[22] Unless noted otherwise, all of the data presented from Table 3 through the end of the paper are based on actual data prepared for the purpose of monitoring performance by a state in relation to a local jurisdiction.

[23] This is obviously a simple case mix adjustment. Depending on the data available, other strata could be developed.

Comparing the panels from side to side shows how the county differs from the state with respect to specific outcomes and outcome-specific trends. Reading the panels from top to bottom shows how the outcomes vary for each discrete population. Each set of data is summarized by an outcome-specific average to make general comparisons more readily apparent. Finally, an overall average is displayed at the bottom of the table. The overall average is the unadjusted performance for the state and the county.

Over the 5-year period, on average and without regard for age at admission, about 26 percent of the children admitted to the state's foster care system were still in care. That figure compares to 33 percent for the county represented in Table 3, an indication that the length of stay in the county may be longer. The data also indicate that statewide, children are more likely to be reunified than they are in the comparison county (44% to 27% respectively). That said, county children are more likely to be discharged to a relative (20% to 14% respectively). Both jurisdictions adopted about 4 percent of the children admitted between 2001 and 2005.

Table 3. Outcomes by Age at Admission, Region, and Year of Admission

Admission Population: Placed Under 1 Year Old

	State					County				
Year	Reunified (%)	Relatives (%)	Adopted (%)	Other Exits (%)	Still In Care (%)	Reunified (%)	Relatives (%)	Adopted (%)	Other Exits (%)	Still In Care (%)
2001	39	15	8	2	36	31	21	14	3	32
2002	37	17	8	1	37	12	24	7	1	56
2003	35	19	10	1	35	24	22	5	3	46
2004	38	21	10	1	30	22	23	6	1	48
2005	35	22	12	1	30	26	29	16	1	28
Average	37	19	10	1	34	23	24	10	2	42

Admission Population: Placed Ages 1 to 13

	State					County				
2001	49	12	1	2	36	32	22	1	9	36
2002	49	12	2	1	36	26	24	1	1	48
2003	47	15	3	1	34	34	18	1	1	47
2004	49	18	3	1	29	34	22	2	2	40
2005	46	21	3	1	28	27	31	2	4	36
Average	48	16	2	1	33	31	23	1	3	41

Admission Population: Placed Ages 14 to 17

	State					County				
2001	52	5	0	30	12	29	13	0	43	15
2002	47	6	0	33	14	27	10	1	47	16
2003	45	7	1	33	14	25	12	1	43	19
2004	44	10	1	34	11	25	13	1	45	16
2005	45	12	1	30	12	26	16	0	42	15
Average	47	8	1	32	13	26	13	1	44	16
Overall average	44	14	4	11	26	27	20	4	16	33

Age differences are important, but the differences are outcome-specific. In both jurisdictions, adoptions are much more likely for children admitted under the age of 1. Above the age of 1, reunification is more likely, given the length of the observation period reflected in the data. For the oldest children, reunification is about as common as it is among the 1 to 13 year olds. However, exits to relatives are much less likely among the older children. All in all, older children are less likely to still be in care, but the difference has more to do with the fact that older children are much more likely to leave for reasons other than permanency. Nonpermanent exits appear more common among older children throughout the state, but the data also indicate that the problem is particularly acute in the county.

With respect to a time trend, the data show general improvements, especially in regard to exits to adoption and relatives (especially if the state had adopted specific policy and practice changes prior to 2003). Statewide, the improvement was in the range of 50 percent when 2001 is compared with 2005. At the county level, the changes are somewhat less remarkable, but in the same general direction. To a certain extent, the more modest gains are attributable to the fact that exits to relatives and adoption were already more common in the county than the rest of the state (on average). Despite the gains in adoption and relative exits, reunification dropped overall but less so for older children (14- to 17-year-olds). The decline in reunification relative to gains in adoption and relative exits speaks to the problem created when separate goals are maintained for different permanency options. The data also reveal the importance of tracking nonpermanent exits as a separate outcome. For the most part, nonpermanent exits (labeled *other exits* in Table 3) are rare among children admitted before age 14. For teens, nonpermanent exits account for between 34 percent and 44 percent of exits, depending on whether one is considering the state or county data.

In sum, the data in Table 3 show why performance has to be disaggregated by population, geography, time, and outcome.[24] Without these data, it is simply not possible to promote specific

[24] It bears mentioning here that the data in Table 3 are concerned with how children leave placement. Safety (e.g., rates of maltreatment, rates of recurrence), entry rates into care, placement stability, and reentry are not covered, although the treatment of these outcomes would easily fit within the basic framework the data in Table 3 are intended to highlight.

process and quality investments with the aim of making improvements where they are most needed. In this example, the county uses relatives as a discharge resource, so other parts of the state may benefit from understanding the practice model that promotes the use of relatives as discharge resources. Going forward, under the terms of a process/quality improvement plan, the state may want to differentiate its expectations to acknowledge current performance.[25] That said, fewer county children are adopted, so county administrators may profit from understanding how adoption rates can be improved. In general, children placed with the county are less likely to achieve permanency, so a general focus on permanency may prove beneficial, notwithstanding the specific differences already mentioned.[26] At both the state and county level, other exits are a primary path out of foster care, particularly in the county. A sole focus on permanency would miss this point. More importantly, reducing other exits (such as runaways) might increase length of stay because children would remain in care, a point that illustrates how different outcomes are interrelated. Positive movement in one direction may have adverse consequences best measured with other outcomes.

Are We Seeing Improvement?

The primary question within the return on investment framework has to do with improvement—are we seeing the improvement we thought we would given new investments or other changes in the process and quality of services provided? The challenge that comes with trying to answer the question, as already discussed, has to do with distinguishing change in performance that is due to the deliberate actions taken from all of the other reasons one might observe changes in performance. From a within-state perspective, the task is further complicated by the fact that changes in performance may be attributable to actions the state takes (e.g., state-level changes in policy and procedure); actions taken at a sub-state level (by a county within a county-

[25] It is also important to point out that simply because use of relatives exceeds the state level, one cannot assume that the process and quality of care are up to standard. Quality, process, and outcomes are interdependent; information about any one indicator without access to information about the others diminishes what can be learned from the information in hand.

[26] County size (the size of the administrative unit under consideration) is important. If the county in Table 3 is a large one, it is hard to imagine how the state's permanency record improves without targeting this particular county. This is another example of how stratification pays off.

administered state system); or both.[27] The same problem exists when trying to understand within-county variation if contract (private) agencies provide the services. Out of all the actions being taken at each of several levels, which ones are the main drivers?

As one might imagine, the answer is difficult to establish, even under ideal conditions. First, one has to establish that a change in the overarching practice context (i.e., process and quality) produced changes in performance net of other factors. That is, if the state passed a law to influence permanency, did the law have the effect legislators imagined it would on the state as a whole? Second, given preexisting differences in county-level performance, such as those shown in Table 3, how did the counties respond individually to the process/quality changes? Put another way, relative to itself, did the county improve its performance, net of other factors including the independent choices a county administrator may have made?

The foregoing suggests that changes in state performance depend on changes at more granular administrative levels (e.g., regions, counties, or private providers). To observe the changes, one has to first establish a baseline performance trend at the state level. As illustrated in Table 3, the trended baseline helps to specify whether in the recent past performance was stable, slipping, or already getting better. Second, it is important to establish a baseline performance trend at the substate administrative level. The choice of units depends on the administrative structure of the state, but the goal is the same: establish the pattern of variation for the units operating in the state. The local view helps to identify which units are already operating at a level that exceeds the state average and those that are not. From a change management perspective, the pattern of with-in state variation is important because one might have different expectations for

[27] The other driver of change is, of course, federal policy. As with state or local policy changes, federal policy changes are implemented in a preexisting performance context. Within that context, it is important to remember that at baseline, states, counties, or other administrative entities may occupy different positions relative to the performance being targeted. Given where a particular county stands in relation to the targeted performance, implementing new rules and regulations may have adverse consequences, particularly if a unit is already performing at a point above the average. This is not to say that administrative units performing above the statistical average will always do more poorly in the context of reform or that there is no room for further improvement. The point is that how units respond to the opportunities changes in policy, practice, or other requirements create may well depend on their starting point. For that reason, it is important to know the path of change from the starting point. Parenthetically, it is another reason to use baselines for outcome measures as opposed to standards.

improvement depending on the level of current performance relative to the state's statistical average.[28]

Each of these interpretive challenges is addressed in Table 4, which reproduces data from a study of provider performance. In this study, the goal was to understand whether individual provider performance improved net of differences in the population of children served and performance changes taking place at the county level. In Model 1, the results show only the system-level time trend. These data show that when compared with 2001, the children placed in 2002 exited to reunification at a rate that was about 6 percent greater (1.059–1= .059), net of child and placement characteristics, which are not shown separately so as to simplify the presentation. The magnitude of the difference is statistically significant. However, for children placed in 2003, the exit rate difference relative to 2001 was negligible from a statistical standpoint (.983–1 = -.017).

By adding a dummy variable for a certain agency (designated as Agency X in the model), the results for Model 2 of Table 4 indicate whether and how the hazard rate of reunification for Agency X differs from the average rate of all other agencies. The hazard ratio of 1.13 for Agency X, which is statistically significant, indicates that children placed with this agency are reunified with their families at a pace 13 percent faster than the rate reported for children placed with the other agencies. Although not shown here, this analysis could be repeated for each agency so as to judge how each agency compares to the average rate of reunification.

[28] A word here about the use of the term *average* is appropriate. The average, or mean, is a statistical term that summarizes a distribution. It is one of several terms (e.g., the median) that serve the same purpose. As such, it is not a term meant to convey a qualitative judgment about a system as in ordinary or just plain average. Were one to locate a child welfare system without fault, performance (e.g., time to permanency) would have a distribution, unless every child served had an identical experience, which would mean that all children were treated the same without regard to their individual needs. The distribution underlying performance would have an average that captures the central tendency in an otherwise excellent system.

Table 4. Proportional Hazard Models for Family Reunification within 2 Years for Children Placed in 2001, 2002, and 2003

	Model 1		Model 2		Model 3	
Variable	Hazard	*P*	Hazard	*P*	Hazard	*P*
System Entry year (2001 as ref.)	1.000		1.000		1.000	
2002	1.059	*	1.060	*	1.050	†
2003	0.983		0.985		0.982	
Agency (all others as ref.)			1.000		1.000	
Agency X			1.130	***	1.090	
Agency/year interaction					1.000	
X-2002					1.084	
X-2003					1.030	
Agency X Performance in 2002					1.18[1]	
Agency X Performance in 2003					1.12[1]	
- 2 Log Likelihood	139,549		139,537		139,536	

Notes: Total agency spells = 13,866; number of reunification events = 7,941.

† $p < 0.10$ * $p < 0.05$ ** $p < 0.01$ *** $p < 0.001$. [1] These figures were derived by multiplying 1.09 (2001) times 1.084 (for 2002) and 1.09 times 1.03 (for 2003). Other control variables in the model are not shown separately.

Because the sample includes children from three entry cohorts, the hazard ratio of 1.13 for Agency X is really the average for the three admission years. To obtain the performance of an agency for each cohort year, an interaction term for the agency with each of the two entry-year variables was added (see Agency/year interaction in Table 4). The results for Agency X, which are found in Model 3, show that the hazard ratio of 1.09 for Agency X now refers to the performance of Agency X in 2001. Although not much smaller than the 3-year average shown in Model 2, the 2001 hazard ratio for Agency X is now not statistically significant, which means that Agency X was about average in its performance of returning children home in 2001.

How then does Agency X compare with all other agencies in 2002 and 2003? The hazard ratios of Agency X for the 2002 and 2003 entry cohorts can be calculated by multiplying Agency X's 2001 hazard ratio (1.09) by the coefficient of the corresponding interaction term. For the 2002

and 2003 entry cohorts, Agency X's performance was 18 and 12 percent higher than the average hazard rate of all other agencies, net of the other factors in the model. Separate analysis not shown here indicates that Agency X's hazard ratios for both entry years are statistically significant.[29] These results suggest that the performance of Agency X did improve slightly relative to the average of the other agencies.

The hazard ratios of 1.05 and 0.982, for the entry years 2002 and 2003 respectively, indicate the change in hazard rate for 2002 and 2003 relative to 2001 for *all* the other agencies except Agency X (see Model 3 in Table 4). These data indicate that after controlling for other factors (case mix) there was a very modest *system-wide* increase in the rate of reunification in 2002 compared to 2001 and no change in 2003 compared to 2001. That the system performance was dependent on the improvements made by Agency X is evident when Model 2 and 3 are compared. By pulling out the increased rate of reunification associated with Agency X, the overall performance (the coefficients associated with the admission year) drops, albeit slightly. The small increment is due to the size of Agency X relative to the whole system and the fact that some other agencies also contributed to the overall performance improvement.

Missing from the analysis so far is an adjustment to the performance of Agency X for the performance of the system as a whole. To calculate the change in the rate of reunification for Agency X from 2001 to 2002 and 2003 relative to itself, the hazard ratios of 1.05 and 0.98 for cohort years 2002 and 2003 are multiplied by the coefficient of the interaction terms of Agency X with entry year 2002 or 2003, which are 1.08 and 1.03, respectively. Presented in Table 5, this calculation shows that for Agency X, the hazard rate of reunification changed by 14 percent for the 2002 entry cohort, after adjusting for the more general system-level changes taking place.[30]

---

[29] This is done by changing the interaction terms between the agency dummy and the two entry year indicators, with 2002 or 2003 as the reference category instead of 2001.

[30] In a separate analysis, the hazard ratio of 1.14 of entry year 2002 versus 2001 for Agency X is shown to be marginally significant.

Table 5. Agency to Self Comparison

Year Comparison for Agency X	Coefficients from Table 4/Model 3 Year Term	Interaction Term	Agency-to-Self Comparison
2002 vs. 2001	1.050 times	1.08 equals	1.14
2003 vs. 2001	0.982 times	1.03 equals	1.01

The results sharpen our understanding of performance in that system-level improvements in the process of reunification that may benefit all agencies in some way have been isolated. This is in some sense an example of how one might determine whether a rising tide is lifting all of the boats. The analysis isolates agency-specific changes after adjusting system changes out. In short, the analysis offers a less ambiguous answer to the question: Is the agency's performance getting better or is the agency a beneficiary of changes taking place around it? In practice, the distinction speaks to the difference between improvements to the way an agency does its work versus improvements in the way the system does its work (i.e., court reform), which benefit all agencies in some way.

If the analysis is repeated for each agency (or some other administrative unit), then for each of the agencies one has an estimate of (1) its average hazard rate of reunification relative to its peers for the three entry cohorts, and (2) its change in hazard rate for entry cohort 2002 and 2003 relative to cohort 2001. The latter is used to examine whether an agency's performance in timely reunification of children in its care has improved compared with its own 2001 baseline, after controlling for attributes of the children served and general trends in the system.

In Table 6, agencies are ranked according to the average hazard ratio of reunification for each agency versus all other agencies over the three entry cohorts.[31] The first-ranking agency has an average hazard ratio of 1.5, indicating its average hazard rate is 50 percent higher than that of all other agencies. The lowest-ranked agency (35th) has a hazard rate only half that of all other

[31] Table 6 includes only the agencies with 30 or more admissions and over five reunifications within 2 years for the three entry cohorts.

agencies. Table 6 also reveals the difference between the hazard rate and the likelihood (i.e., proportion of children reunified) as performance measures of family reunification. Some agencies with below-average hazard ratios have returned over 60 percent of children to their parents within 2 years of admission— more than some agencies with above-average hazard ratios. The discrepancy can be explained this way. First, the percentage reunified is unadjusted, that is, differences in the makeup of an agency's caseload are not taken into account. Second, the difference may be connected to the difference between speed and likelihood described earlier. Agencies that get more children home on a percentage bases may do so at a slower rate. The hazard rate reported in Table 6 detects differences in the timing of reunification. On average, agencies with above average hazard rates reunified 60 percent of their children; agencies with below average hazard rates reunified 55 percent of the children in their care.[32]

[32] One way to address the fact that timing and likelihood are different is to monitor both and combine the results. For example, the hazard model used here is useful for measuring timing. A logistic regression can be used to test whether the child was discharged (i.e., the likelihood). The combined rank order can be blended to form an overarching impression of how well an agency (or county) does with respect to reunification.

Table 6. Average Hazard Ratios of Reunification within 2 Years for Contract Agencies and Changes of Hazard Ratios in 2002 and 2003 Relative to 2001

Agency Rank	Percent Reunified	Hazard Ratio	Change in Hazard Rate 2001 vs 2002	Change in Hazard Rate 2001 vs 2003	Agency Rank	Percent Reunified	Hazard Ratio	Change in Hazard Rate 2001 vs 2002	Change in Hazard Rate 2001 vs 2003
1	61	1.50*	Same	Same	19	63	0.96	Same	-
2	62	1.31*	-	Same	20	62	0.95	Same	Same
3	58	1.23	Same	Same	21	59	0.95	Same	Same
4	59	1.18*	Same	+	22	58	0.94	+	Same
5	61	1.13*	+	Same	23	57	0.94	-	Same
6	61	1.12*	Same	Same	24	53	0.92	Same	Same
7	59	1.12	Same	Same	25	58	0.91	Same	Same
8	65	1.11	Same	Same	26	61	0.91	Same	Same
9	67	1.10*	Same	Same	27	56	0.88*	Same	Same
10	67	1.09	Same	Same	28	55	0.88	Same	Same
11	62	1.07	Same	-	29	66	0.87	Same	Same
12	62	1.06	Same	-	30	52	0.87	+	Same
13	64	1.03	+	Same	31	46	0.84*	Same	Same
14	55	1.03	+	Same	32	53	0.77*	+	Same
15	53	1.02	+	Same	33	51	0.77*	+	Same
16	59	1.02	Same	Same	34	53	0.75*	Same	Same
17	61	1.00	+	Same	35	39	0.50*	Same	Same
18	60	0.96	Same	Same					

Notes: Ranking of agencies from 1 to 35 is in descending order of hazard ratios. A "+" means a significant increase in the agency to self comparison; "-" means a significant decrease, and "same" indicates no substantial change in the rate of reunification. The significance level for hazard ratios and change in hazard rates is $p < 0.05$.

The last two columns of Table 6 demonstrate how the hazard rate of reunification for an agency changed in 2002 and 2003 compared with 2001 (the agency-to-self comparison). In 2002, over two-thirds of the agencies maintained more or less the same performance level for the three entry cohorts. Nine agencies significantly increased the hazard rate of reunification for the 2002 entry cohort relative to 2001. The performance of these agencies lifted the system performance to a

level above what it was in 2000. It also suggests that if there were specific system-wide improvements implemented, the benefits were realized by a handful of agencies.

Only two agencies (Nos. 2 and 23) saw their performance slip for cohort 2002 compared with 2001, although their performance did bounce back to the 2001 level in 2003. For the 2003 cohort, all agencies except four maintained the same performance level as in the base year 2001. There is no agency whose performance is consistently above or below the base year for both 2002 and 2003. This scenario of relative stability in performance of most contract agencies is consistent with the finding of Model 1 in Table 2, which indicates that the average hazard rate of family reunification for all contract agencies increased by a mere 6 percent in 2001 and dropped back to the same level as the base year for the 2003 cohort.[33]

In Real Time: Baseline, Target, and Actual

Real-time feedback poses a significant challenge with regard to performance monitoring. Analysis of outcomes—understanding the geographic/administrative variation, population variation, and variation over time—provides an idea of where efforts to improve the system ought to be targeted. Once a change initiative is underway, feedback provides information that describes whether the intended changes are underway. Historically, the information needed for feedback has been slow in coming, sometimes arriving long after it would have been useful to have. Thus, the challenge is to provide useful information at a time when, if midcourse corrections are needed, the opportunity for corrective action is still at hand.

To meet the objective of providing feedback in something approximating real time, it is helpful to have three pieces of information: the baseline, the target, and the actual. The *baseline*, which was discussed previously, is a statement of what would be true if nothing changed—the business-as-usual scenario. The *target* is a statement of what would be true if the process, quality, and other management changes have the impact expected.[34] The *actual* records what did in fact happen

[33] Among other issues, these data illuminate the fact that change often comes slowly to the child welfare system. The data also suggest that performance improvement may be difficult to sustain. All the more reason, however, for being clear about how one looks for improvement.

[34] There are various ways to set targets, some of which are better than others. However, a discussion of those approaches is outside the scope of this paper.

through time. For example, the baseline likelihood of exit to permanency might be 50 percent within 24 months of when a change initiative starts. The target might seek a 20-percent improvement over that period, which would lift exits to permanency up to 60 percent. The actual would indicate whether the goal was accomplished. More importantly, because the target was set over a 24-month period, the actual data at 6-month intervals may be used to assess whether the improvement trajectory observed at 6 months is consistent with what would have to be true if the 20 percent improvement goal is going to be reached. In this way, the time involved in providing feedback is shortened from the end of the 2 years to 6-month intervals. Depending on the circumstances, shorter time intervals (every 3 months, for example) may be used.

The process of developing the feedback process starts with a baseline. There are a number of ways to present baseline data; the approach taken has to fit well with the question being asked and the objectives under consideration. Below in Table 7, the data show the cumulative percentage of children who reach permanency by time interval for each of five admission cohorts. The question being asked in the context of a change initiative might be this: After 6 months in care, what fraction of the cohort has achieved some sort of permanency? After 1 year? After 2 years, and so on? Table 7 shows a slightly different version of the data in Table 3, without the specific reference to type of permanency or other exit types. A similar table, stratified by age at admission can (and should) be produced with this information.

The table highlights a number of important features, many of which have already been discussed. First, after 4 years, 9 percent of the children were still in care. This demonstrates that long-term benefits of any change initiative will take some time to observe in their entirety. There is little that can be done to avoid this basic reality.[35] Second, as one moves closer to the present, the amount of available information regarding performance drops. In this example, only the first 6 months of history were available. Last, the data underscore the fact that although any initiative set to start in 2007 can influence what happens to children from prior cohorts, the timing of that

[35] Often what happens is that observation of outcomes stops after a certain point. For example, when an outcome standard for exits with 12 months is used (as is true with one of the CFSR measures), one has to worry about what happens beyond the 12-month mark. The notch problem refers to this situation and what happens when a specific reference point is used to set a standard against which performance is measured. As the data in Table 7 suggest, it is quite easy to show that both success (e.g., increasing the likelihood of exit to permanency) within the first 12 months and an increase in the overall average length of stay are quite possible because of the timing/likelihood issues raised earlier.

effect is very different depending on whether one is talking about the cohort of children admitted in 2002 or in 2004. For the 2002 cohort, the initiative will only influence permanency for children still in care at the end of 4 years, in which case an impact on adoption is the most likely path of change. For subsequent cohorts, the potential impact will come earlier.

For this particular jurisdiction, the data portray a stable system. With each successive cohort, the cumulative percent discharged changes little over the full trajectory, a fact that frames what one might expect to happen in 2007. These data suggest that within 6 months of admission, about 20 percent of the children will be discharged to permanency. This is a good guess because it has been observed so frequently in the past. Although there are other considerations, such as case mix, type of exit, and so on, the data suggest strongly that history in this system has a tendency to repeat itself. The data also suggest that for earlier cohorts, the stability of the underlying system means that one can take the experience from prior cohorts and fill in the expectation whenever the exits have yet to be observed. For example, notwithstanding changes that might yet influence the outcomes, it appears that after 2 full years of observation, one should expect that 71 percent of the children admitted in 2005 would exit to permanency within 24 months.

The target question is then how these data will change as the result of process, quality, and management improvements. First, it is reasonable to assume that within 6 months, exits to permanency (expressed as the fraction of children admitted who were discharged to permanency) will increase. Similarly, at each subsequent interval, performance improvements ought to raise the fraction of children discharged to permanency. Setting the specific target is difficult insofar as systems typically have little experience projecting what will happen and when at the level of detail implied in Table 7. It is, however, an opportunity to engage stakeholders throughout the child welfare system. The target becomes the goal and the stakeholders at the table have a hand in defining the path to success. The question is, how much improvement and by what means?

An important point here, and one that needs to be considered carefully, has to do with other outcomes. An increase in permanency exits is important provided there are no adverse consequences, such as an increase in reentry rates. When selecting targets, one has to anticipate these other outcomes as well. A state, county, or agency with an admirable record of getting children home has to be assessed in light of its reentry rate. Outcomes are interdependent, and assessment of baseline performance has to take these interdependencies into account. When setting targets, it is helpful to consider other outcomes. The idea of a balanced scorecard comes

to mind when contemplating how one develops a holistic sense of performance. It is also important to remember that process and quality are important components of the overall assessment as well.

Table 7. Historical Data Used to Set the Baseline

Entry Cohort	In 6 Months (%)	In 1 Year (%)	In 2 Years (%)	In 3 Years (%)	In 4 Years (%)	In 5 Years (%)
Historical data						
2002	20	41	70	84	91	?
2003	20	41	71	84	?	?
2004	19	40	71	?	?	?
2005	20	41	?	?	?	?
2006	19	?	?	?	?	?
Baseline Data						
2007	?	?	?	?	?	?

Once the goals for system improvement have been set, initiatives that embody the theory of change are set in motion. Reform initiatives operate at the systems level and involve changes in agency structure, policy, staffing, and how funding is used. Initiatives also involve new process and quality standards. From a monitoring perspective the critical issue is knowing when the initiative starts so that monitoring reports reflect the timing of those initiatives relative to when along a child's service trajectory change can be expected to happen.[36] From that point forward, it is possible to look for and find the impact of those investments.

In Table 8, baseline, target, and actual data are presented side by side to illustrate how the data can be used to make interim judgments about progress. The data refer to a single state and the service jurisdictions within the state. The population under review is children in care on January

[36] This is especially true for the children already in care when the initiative starts. This is another reason for focusing on the children in care (or known to the system) at the time a new program is launched. The differential exposure to the benefit of a particular intervention will be influenced by where those children are along their service trajectory and why.

1, 2005 and the performance window is 1/1/2005 through 12/31/2005.[37] The performance under review is *exits to permanency* by the end of the performance window. For the state as a whole, the baseline (which was developed using historical data not unlike the data presented in Table 7) suggests that 59 percent of the children in care on January 1, 2005 (column 2) will achieve permanency by 12/31/2006. The performance target was 10 percent, which means that 65 percent of the children would achieve permanency by 12/31/2006 (column 3) if the process, quality, and management innovations were successful.[38] From the long-term target (2 years out), it is possible to suggest what would occur if a successful change process were underway. In column 4, the estimate shows what would be true 1 year into the innovation cycle if the state were on target to meet its 2-year goal. Finally, the data in column 5 show actual exits to permanency, which is what was achieved during the first year (1/1/2005 through 12/31/2005). Comparison of these data with the interim target offers a chance to assess progress along the way (column 6).

Table 8 suggests that 1 year into the improvement cycle, the state was ahead of the interim target (39% to 35%, respectively). Although 1 year into the project is still too soon to say with certainty whether the ultimate goal will be reached, the interim (i.e., nearly real-time) feedback is positive. Two other features of Table 8 are also worth noting. With respect to the substate jurisdictions, interim progress is uneven. Some of the jurisdictions have a larger target-actual gap; one jurisdiction is behind halfway into the cycle. The variation in the performance should be viewed from the perspective of the starting point. In some cases, the lower rate of improvement is found in jurisdictions where the performance is already above average. Other times, the larger rate of improvement is in places that have lower average performance at baseline. The question of how a specific baseline (e.g., at the county level) compares to the overall average is therefore an important consideration.

---

[37] The data presented in this example are for illustrative purposes only. In a real-life setting, one would use the most currently available data. Table 7 does illustrate one important point, which is that the closer one moves to the present, the more incomplete information one has to manage.

[38] It is important to be somewhat circumspect when judging whether, because goals are accomplished, an intervention worked. Although a structured review of data improves acuity during performance monitoring, it is difficult to rule out all of the plausible competing hypotheses. By taking a structured approach (reducing selection biases, stratifying the sample, etc.), one reduces the likelihood of drawing the wrong conclusion but does not altogether eliminate it.

Table 8. The Baseline, Target, and Actual Comparison

Statewide/ Region	(1) Population as of 1/1/05	(2) Two Year Baseline (as of 12/31/06)	(3) Two Year Target (as of 12/31/2006)	(4) Target as of 12/31/2005	(5) Actual as of 12/31/2005	(6) Difference, Actual-Target
			Exits to Permanency as Percent (%) of Admissions			
Statewide	5,500	59	65	35	39	4
A	533	51	56	30	46	16
B	943	63	69	38	37	-1
C	226	41	45	24	33	9
D	362	45	50	26	33	7
E	661	57	63	34	36	2
F	578	64	70	37	43	6
G	247	71	78	42	48	6
H	537	45	50	27	31	4
I	325	64	70	38	42	4
J	319	63	69	38	49	11
K	302	66	73	39	47	8
L	467	62	68	36	36	0
			Exits to Permanency as Number of Care Days Used			
Statewide	5,500	1,622,570	1,541,442	679,872	709,736	29,864
A	533	156,079	148,275	66,204	56,247	-9,957
B	943	247,763	235,375	103,499	120,605	17,106
C	226	77,187	73,328	32,360	33,217	857
D	362	117,650	111,768	48,542	50,313	1,771
E	661	202,594	192,464	84,819	98,529	13,710
F	578	162,632	154,500	67,840	69,072	1,232
G	247	62,844	59,702	26,768	29,828	3,060
H	537	173,050	164,398	72,893	71,472	-1,421
I	325	90,273	85,759	38,162	39,991	1,829
J	319	95,329	90,563	39,934	37,876	-2,058
K	302	88,421	84,000	37,012	35,708	-1,304
L	467	149,421	141,950	62,018	66,878	4,860

The lower panel also shows the baseline, target, and actual care day counts. The separate presentation refers to the distinction between timing and likelihood made earlier. Here, the data show that although jurisdiction D has successfully increased the fraction of children reaching permanency by 6 percent, the total care days used increased by 1,771 days. In fact, in quite a few of the within-state jurisdictions and in the state as whole, there was an increase in care days used to go along with the increase in exits to permanency.[39] What this means is that the total amount of care provided was actually above the target, which translates into higher costs because the costs in a typical foster care system are tied to days used. The example shows quite nicely why timing and likelihood have to be judged separately within a return-on-investment framework.

As with the other examples, the data in Table 8 should be developed for various strata within the population. Doing so sharpens the diagnostic value of the reports insofar as success is then isolated in certain parts of a state, certain populations within the state, or certain populations within certain jurisdictions. One goal of the reporting process is to increase the total amount of usable information available.

Summary

Performance monitoring is not a single act conducted in isolation. Rather, it is a continuous process that *engages* all the stakeholders in a set of structured activities whereby the stakeholders develop a real feel for what the agency does well and what the agency could do better. From there, the goal is to develop hypotheses that connect process, quality, and management improvements to a theory of change that is then implemented through a series of deliberate investments. Monitoring is the deliberate act of making sure that the relevant actors maintain fidelity with process, quality, and management standards in light of outcomes. In the end, the agency should know whether its investments have paid the hoped-for dividends.

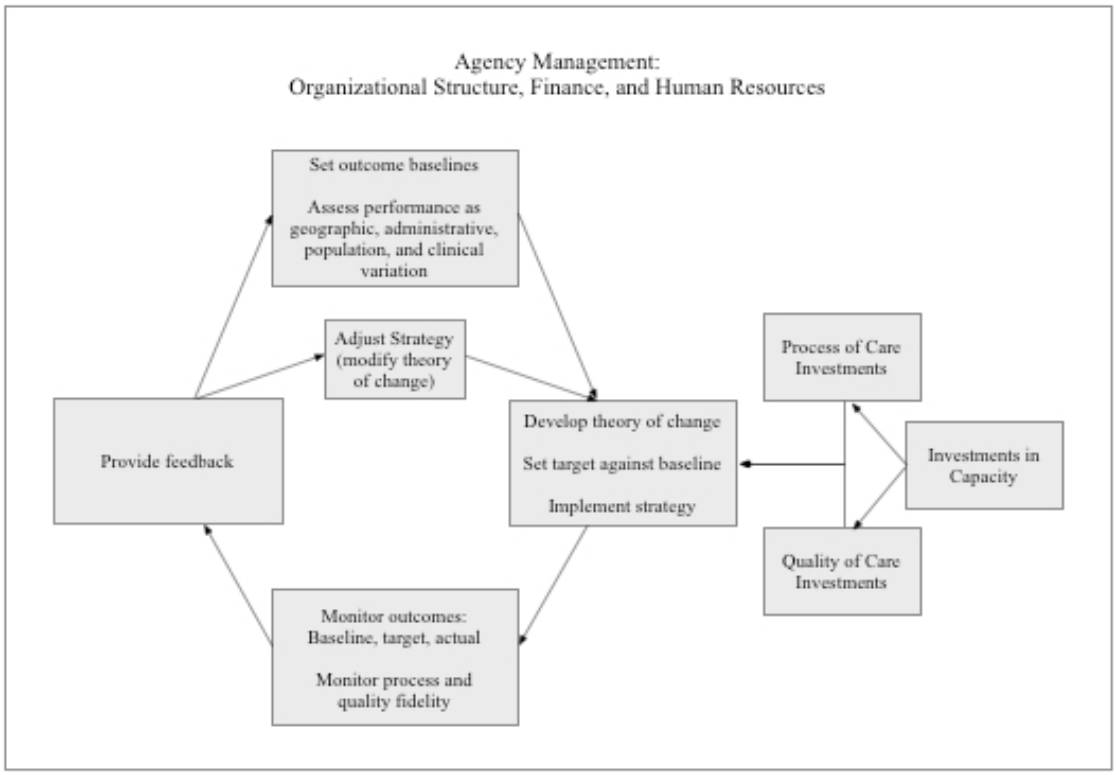
The framework presented here embeds performance monitoring in a routine process of strategic agency management. The process itself is portrayed in Figure 1. The overarching context is agency management, which speaks to how organizational structure, finance, and human resources

[39] This often happens when exits cluster during a particular month near the end of the performance window. How these seemingly anomalous findings should be interpreted depends to a very large extent on how well one understands the practice implications (i.e., process and quality) behind the theory of change.

are deployed in relation to the agency's core mission and values. The main inputs are the process and quality investments an organization makes, given a theory of change that emerges from the agency's assessment of its performance relative to its mission. Within this cycle, monitoring is the mechanism by which an agency tracks whether outcomes are moving in the intended direction and whether process and quality standards are being met relative to those outcomes.

Performance monitoring starts with a clear understanding of the outcomes that are central to the agency's mission and values. In the case of the child welfare system, the core mission takes shape around the construct of well-being with particular emphasis on securing safe and permanent families so that children are able to flourish cognitively, behaviorally, and socially. Toward this end, the process and quality of care represent the main tools administrators have at their disposal when trying to influence the what and how of work with families.

Figure 1: Performance Monitoring in an Organizational Context



Specific steps—or the targeting of efforts—are a function of how one comes to understand the pattern of variation in a given state or locality, best thought of as the history of performance. Baseline performance, couched in terms of geographic, demographic, administrative, or clinical variation, provides the basis on which to frame the question of what needs to get better, and for whom. From that point, the goal is to state in clear and unambiguous terms how changing process and quality requirements will promote desired outcomes. Properly constructed, the baseline provides a way to define the improvement target, to track progress, and to make midcourse adjustments as needed, all with an eye toward shrinking the gap between expected outcomes and reality.

Simple in concept, performance monitoring in complex systems poses unique and often significant challenges. Chief among those challenges is drawing a clear and concise distinction between process, quality, and outcomes when in reality they are closely related. The idea of

process without quality, quality without process, or process and quality without a link to outcomes underscores the interdependencies. Yet, for measurement purposes, process, quality, and outcomes have to be separated. The distinction rests in the difference between what one does (the process), how well ones does what one does (quality), and whether one achieves what one sets out to do (outcomes).

The conceptual challenge of separating process and quality from outcomes are matched by the difficulties in measuring performance. To simplify the problem, the discussion was divided into several discrete issues. Sampling dominates the data-collection issues, in part because monitoring depends on a realistic sample of all that is taking place in the system without blurring the picture one gets back. The performance window, stratification, inception, and process samples were offered as ways to reduce ambiguity in light of underlying difficulties. The performance window establishes the timeframe during which performance is measured, whereas stratification reinforces the idea that children have experiences that differ for reasons that have less to do with performance and more to do with the underlying needs of children and families that have to be respected. Inception, one form of stratification, underscores the fact that too little specific attention is paid to what happens when children and families *first* touch the system. Too often children (and families) are grouped together without distinction, as though prior history is not connected to what happens next. Last, process samples offer an opportunity to study process fidelity in a more selective manner, given that observing all process elements within the history of a single case is relatively unlikely (e.g., children reunified cannot be used to understand fidelity with the adoption process).

As a means of understanding performance improvement, there is an important distinction to be made between standards of practice—fidelity to process requirements or protocols—and performance standards that speak to the level of acceptable variation. We argue that performance standards are more applicable to process and quality in part because one expects to have greater control over fidelity. Outcomes are more amenable to baselines in part because strict adherence to practice standard does not guarantee outcomes. Baselines offer a way to categorize expectations, with an understanding that there are real limits to how much we control.

Baselines also have greater applicability in the context of case mix adjustment. Case mix adjustment captures the idea that child and family differences often dictate what happens. By providing different baselines for distinct groups of children, as opposed to a single outcome

standard, one is actually conveying clinically useful information to the individuals charged with serving children and families. That information, one could argue, improves the process of designing services that work.

The last step in the performance-monitoring cycle calls for feedback on whether the goals are being accomplished and for adjustment of the strategy/theory of change. Without feedback, there is no way to communicate whether the process, quality, and management requirements put in place are having their intended impact. To be effective, feedback has to be constructed around the gap between baseline performance, target performance, and actual performance. Properly assembled, the baseline, target, and actual performance offer the possibility of real-time feedback that makes midcourse correction more viable.

As noted at the outset, investments in child welfare services may yield secondary benefits, but without a strong, unambiguous link to mission-critical outcomes, it will be hard to build a strong case for future investments. The framework offered here attempts to grapple with the issues a rigorous system of accountability has to address. Above all, it is important to remember that performance monitoring is not a static undertaking. On the contrary, the process is as dynamic as our ability to test new ideas.

Appendix

Inception cases are children (or families) whose contact in a given year is their *first-ever* contact with the child welfare system. In a given system, the number of children who fit this description will likely vary. For example, the 2006 maltreatment report published by the U.S. Department of Health and Human Services shows that states differ substantially with respect to the fraction of all victims in a year who are first-time victims.[40] Because recurrence is affected by whether or not there are prior reports, it stands to reason that the recurrence rate in a given state will be lower if the population of victims this year consists of more first-time victims.

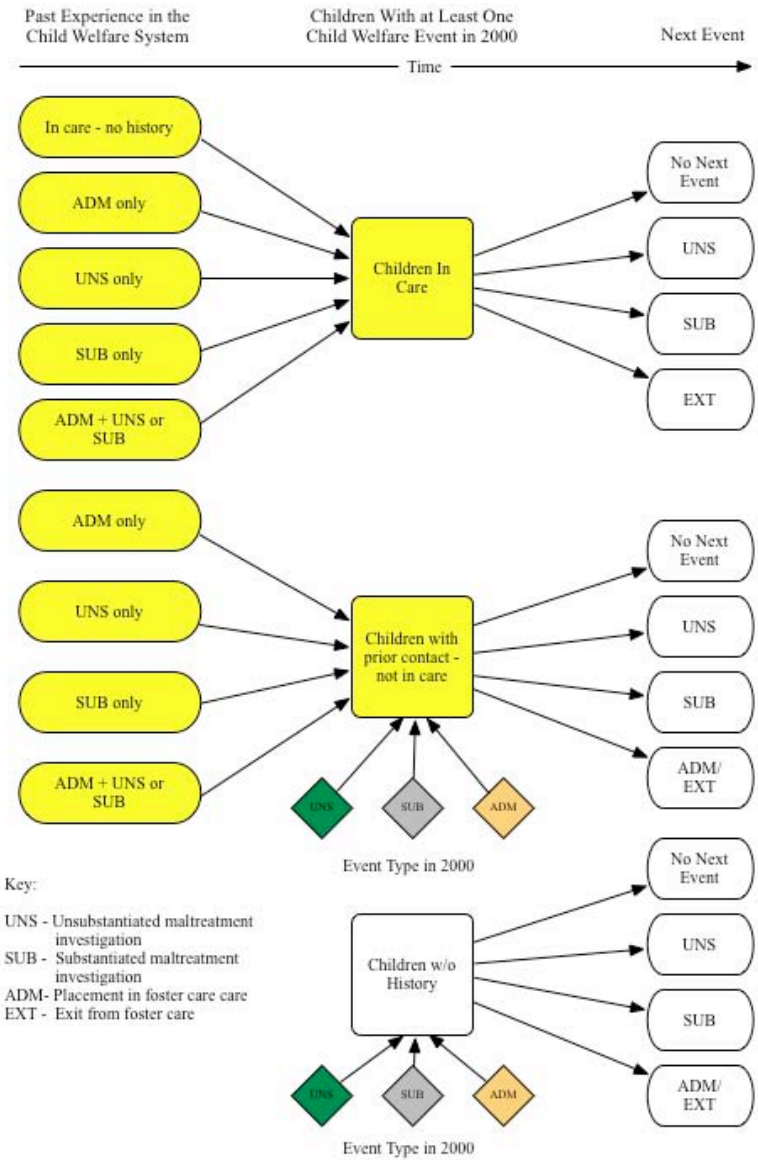
Appendix Figure 1 illustrates the many pathways children may take on their way to being served in a given calendar year (in this example the year is 2000). There are three primary populations depicted. The first is children in care at the start of the year that is being studied (the top, yellow-shaded square in the middle of the diagram). By definition, children in this group were known to the system in a prior year. The second population consists of children who were not in care at the start of the year, but who were known to the system (the second, yellow-shaded square in the middle of the diagram). The diamond shapes connected to the bottom of the square show the options for how the child became known to the system: as the subject of an unsubstantiated maltreatment report (UNS), as the subject of a substantiated maltreatment report (SUB), or via admission to foster care (ADM). The third group of children consists of inception cases or children without any prior history.

To the left of the children with some prior history (shaded in yellow) are the pathways they may have taken prior to their contact in 2000. The options are in-care no history, which refers to children in placement who were never the subject of a maltreatment reports but were placed

[40] It is an imperfect example because the label, *first-time victims*, is not limited to children whose first-ever report was substantiated. For example, first substantiated investigations and substantiated first investigations refer to two different populations. In the first instance, the first substantiated allegation could have been preceded by any number of unsubstantiated investigations. In the second, the first investigation is also the first substantiated investigation.

nevertheless (this is typically a small number of children); children whose only history was one or more prior placements (again a small number of children); children in placement who were the subject of one or more unsubstantiated maltreatment reports only; children in placement who were the subject of one or more substantiated maltreatment reports only; or children with some combination of all the event types. The same combination of events describes the range of possibilities for what may have happened to children with prior history who were not in care. For the inception cases, there was no prior history.

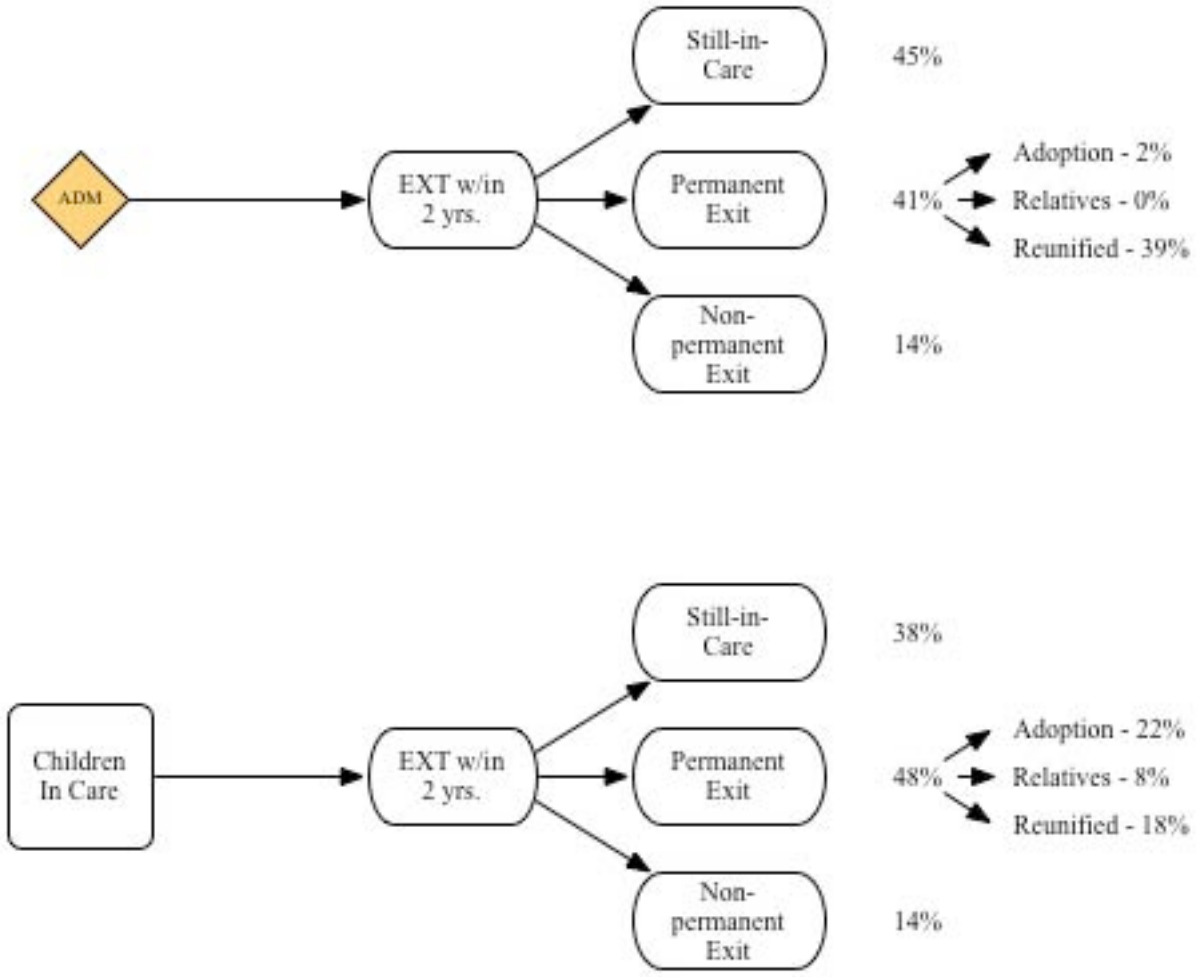
Figure A-1: Child Welfare Trajectories for Children Admitted to Care and Children Already in Care at the Start of the Year



Appendix Figure 1 also shows what might happen after the initial event of the year. There may be no next event, that is, the first event is the only event ever, including subsequent years as in the case of a child who is the subject of one and only one unsubstantiated maltreatment report. Alternatively, the child may have a subsequent maltreatment report (either substantiated or not) or the child may be placed, provided the child was not placed as the first event (in which case the next event may be a discharge from care). The diagram is meant to convey the idea that each combination of successive events is possible (e.g., SUB/SUB, SUB/UNSUB, etc.) and may be followed by yet another event.

Appendix Figure 1 is meant to convey the possibility that what happens next may be affected by what has already happened. Appendix Figure 2 illustrates the point more succinctly by zooming in on two populations: children admitted to care and children in care at the start of the year. The question asked is: where were these children 2 years later (still in care, exit to permanency, or non-permanent exit), and what type of permanent exit was observed for those children who exited to permanency? The data reveal clear differences. The children admitted, as opposed to the children who started the year in care, were more likely to still be in care 2 years later (45% to 38% respectively). Conversely, the children in care were more likely to exit to permanency (48% to 41% respectively). Of the permanency exits, adoptions were more common during the window of observation if the child was already in care at the start, whereas reunification was more common among children in the admission group.

Figure A-2: Child Welfare Trajectories – Exits from Care by Prior Status



As one might expect, these data will differ from state to state and from jurisdiction to jurisdiction within a state. One is also likely to encounter population differences depending on the age of the child and the type of placement, among other factors that influence what happens and when (as seen in Table 3). The challenge is using these data to better understand what will likely happen next, given the sequence of prior events, if any. If one wants to measure whether a child welfare

agency improves permanency outcomes going forward, the baseline has to account for the kind of differences shown in Appendix Figure 2.

About Chapin Hall

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.

1313 East 60th Street
Chicago, IL 60637

T: 773.256.5100
F: 773.753.5940

www.chapinhall.org

casey family programs

fostering families. fostering change.

Casey Family Programs
1300 Dexter Avenue North, Floor 3
Seattle WA 98109
www.Casey.org

ChapinHall at the University of Chicago

Policy research that benefits children, families, and their communities