Performance evaluation of child welfare departments using Data Envelopment Analysis: A comparative study across US states

Sedghi, S., Azizi, S., Canada, K., Charles, V., & Trapp, A. C. (2024). Performance evaluation of child welfare departments using Data Envelopment Analysis: A comparative study across US states. *Socio-Economic Planning Sciences*, Article 102101. Advance online publication. https://doi.org/10.1016/j.seps.2024.102101

Published in:
Socio-Economic Planning Sciences

Document Version:
Peer reviewed version

Performance Evaluation of Child Welfare Departments Using Data Envelopment Analysis: A Comparative Study Across US States

Sepideh Sedghi[a], Shima Azizi[b], Katherine Canada[c], Vincent Charles[d], and Andrew C. Trapp[a, e]

[a]The Business School, Worcester Polytechnic Institute, Worcester, MA 01609
[b]INFICON, East Syracuse, NA USA
[c]Canada Consulting, Chelmsford, MA 01824
[d]Queen's Business School, Queen's University Belfast, BT9 5EE, UK
[e]Data Science Program, Worcester Polytechnic Institute, Worcester, MA 01609

ARTICLE HISTORY
Compiled October 25, 2024

ABSTRACT. Public child welfare agencies play a pivotal role in safeguarding the well-being of children and thus, the future of our society. While the performance of child welfare agencies is of critical importance, limited previous research relying on operations research and advanced analytics appears to exist in the analysis of their performance. We conduct a multi-criteria analysis for benchmarking the performance of the United States child welfare system, using Data Envelopment Analysis (DEA) to evaluate the performance of public child welfare agencies across different US states. We select as outputs various statewide data indicators from the Child and Family Services Review (CFSR), while our inputs include the total annual expenditure by each state on the child welfare system. We use clustering to differentiate agencies based on the presence of the "Alternative Response" policy, which provides for preventive and support options for families, and apply DEA to each homogenous cluster. We identify best-practice agencies and provide benchmarks for the remaining agencies to enhance their performance. Our study offers data-driven directions for child welfare agencies to improve safety and permanency outcomes for children.

KEYWORDS
Performance Analysis, Child Welfare, Data Envelopment Analysis, Slacks-Based DEA Model, Child and Family Services Review (CFSR)

CONTACT Andrew C. Trapp. Email: atrapp@wpi.edu

1. Introduction

The Adoption and Safe Families Act of 1997 established three goals for children served by child welfare in the United States (US): safety, permanency, and well-being. These goals have become paramount in guiding the responsibilities of these agencies. To reach these goals, child welfare agencies try to support families and children in crisis, by preventing abuse and neglect, and when necessary, finding permanent and supportive homes for children. The authority for delivering child welfare services is reserved for the states. More than 4.3M referrals—initial notifications alleging child maltreatment—were made to child welfare agencies nationwide in 2018, a rate of 58.5 per 1,000 children in the general population (Children's Bureau, 2020b). Approximately 2.4M of these referrals were screened-in, meaning they were identified as meeting the criteria for further investigation by the child welfare system, with the national rate of screened-in referrals being 32.5 per 1,000 children (Children's Bureau, 2020b).

Approximately 678,000 victims of child abuse and neglect had at least one substantiated instance of maltreatment among Federal Fiscal Year (FFY) 2018 screened-in referrals. Public child welfare agencies play an important role in supporting these vulnerable children and their families. It is estimated that government spending on the child welfare system totaled $33 billion USD in 2018 (Rosinsky et al., 2021). At times, child welfare services may assist children and families within their own homes; at other times, they may provide assistance upon and post-removal of children from their homes for foster care placement, the purpose of which is to protect removed children by providing a temporary environment while aiming for either reunification with the original family or finding another permanent option. Children who are removed from their homes typically spend an average of 22 months in state care, while approximately 6% have been in the system for over five years (Annie E. Casey Foundation, 2021). In addition, among children discharged from foster care, 8.1% re-entered the system within 12 months (Children's Bureau, 2021). This emphasizes the need to work towards decreasing these occurrences of re-entrance. Evaluating the performance of child welfare agencies provides an opportunity to understand the relative strengths and areas for improvement among state child welfare agencies.

The Child and Family Services Review (CFSR) is the primary mechanism for evaluating child welfare agencies, based on a variety of factors such as organizational structure, policies and procedures, case management practices, and service delivery of the child welfare agency. The CFSR monitors child welfare systems in 50 states plus the District of Columbia and Puerto Rico to determine whether each conforms with federal requirements (Children's Bureau, 2014a). The primary objective of the CFSR assessment is to guide state child welfare programs and practices toward achieving positive outcomes in safety, permanency, and the well-being of children and families (Ahn et al., 2017). The CFSR program is discussed in greater detail in Section 2. However, the focus of this program is predominantly on assessing child welfare outcomes. This is where Data Envelopment Analysis (DEA) can play a pivotal role: DEA provides an avenue to evaluate the performance of child welfare agencies by not only assessing their outputs but in relation to the *levels of input resources*.

Our study makes several contributions to the use of analytics for child welfare systems. We are the first to apply DEA in this domain. In contrast to the historic CFSR, which does not consider performance levels relative to resource allocations, the use of DEA offers the opportunity to assess performance (i.e., outputs) within the context of available resources (i.e., inputs). Another contribution is establishing benchmarking practices through the DEA, which can help aspiring agencies improve their performance by providing reference sets and target variables. We also provide state child welfare agencies with actionable recommendations on how they can reach these target sets that aid in managerial decision-making. Finally, we introduce a public-facing, interactive interface in Tableau that displays our analysis results, to increase the accessibility of our DEA findings for state child welfare stakeholders.

Our paper is structured in the following manner. Section 2 provides a background of the CFSR program and also presents a brief history of DEA modeling, including their use in nonprofit organizations. In Section 3, we introduce the dataset utilized in our analysis and explain the particular DEA methodology employed in the context of child welfare, as well as our clustering approach. Section 4 discusses the results obtained from the application of DEA, along with suggestions for performance enhancement in agencies. The paper concludes with Section 5.

2. Background and Literature Review

In this section, we focus on the CFSR and research studies that have focused on the CFSR. Following this, we introduce the methodology of DEA with a focus on its application in various public and social service sectors. Finally, we highlight the existing gap in the application of quantitative methods, such as DEA, in analyzing child welfare, thereby providing a new perspective on performance assessment in this important field.

2.1. Child and Family Services Review

The CFSR is administered by the federal Children's Bureau, housed within the United States Department of Health and Human Services' Administration for Children and Families (ACF). The CFSR process generally includes three sequential steps: 1) a statewide self-assessment that involves federal and state partners who review data on established indicators; 2) an onsite review to examine randomly selected cases and interviews with key participants; 3) a two-year Program Improvement Plan (PIP) developed by the state to address areas identified as needing improvement as a result of the self-assessment and the findings of the onsite review. Finally, states must show regular advancements toward PIP goals to avoid penalties. Since its establishment in 2001, CFSR has undergone three rounds of revision, including changes to the methodology for calculating statewide data indicators measuring child safety and permanency.

Several research investigations have employed CFSR data in descriptive child welfare examinations, using content analysis (Belanger et al., 2008), assessment of training programs (Amodeo et al., 2009), policy analysis (Huggins-Hoyt et al., 2019), and literature review (Carnochan et al., 2014). Another study compiled data from all three rounds of the federal CFSR conducted between 2001 and 2018 (Ahn et al., 2022).

Previous research using CFSR data focuses primarily on outcomes, with little attention given to the resources or inputs used by states to achieve these outcomes. It may be inequitable to directly compare the outputs of larger states with higher budgets to those of smaller states with limited resources. Such comparisons can lead to misleading conclusions about the performance of child welfare agencies. To address this gap, a more comprehensive approach is needed to assess the performance of the US child welfare system by considering both inputs and outputs of the agencies. The mathematical method of DEA allows for the consideration of both inputs and outputs, thereby providing a more balanced and fair evaluation of state performance in child welfare. DEA enables a number of valuable post-optimality analyses such as establishing benchmark references for each state that is not considered best practice, providing insights into the strengths of each state, identifying areas for improvement, and setting attainable goals for those areas based on the performance of the reference set.

2.2. Data Envelopment Analysis

Data envelopment analysis is a linear programming method for evaluating the technical efficiency of homogenous decision-making units (DMUs). It was introduced by Charnes, Cooper, and Rhodes (1978), and subsequently expanded upon by Banker et al. (1984). DEA uses a vector consisting of a set of inputs and outputs to assign an efficiency score to each DMU. Efficiency refers to how well a unit is able to generate outputs with the least possible amount of resources. This method identifies the best practice DMUs and forms a frontier that envelops other inefficient DMUs. Efficient DMUs are the ones that cannot increase their outputs or decrease their inputs without simultaneously reducing another output or increasing another input. Once the efficient DMUs are determined, the non-frontier DMUs are scored based on their distance from the efficient frontier. There exists a wide variety of DEA models suitable for specific contexts and measurement requirements which will be discussed in Section 3.

Among the variety of available performance assessment techniques such as ratio analysis and regression analysis, DEA has several attractive features (Nyhan & Martin, 1999). First, DEA is a non-parametric approach that makes it flexible enough to accommodate multiple independent inputs and outputs to assess performance. Second, DEA requires no predetermined judgment about the importance of each input and output, as DEA assigns its own optimal weights for each DMU. Finally, DEA provides a single measure of efficiency ranging between 0 and 1 (larger values mean greater efficiency) for each DMU. This simplicity makes the results of DEA more comprehensive for researchers and stakeholders. This last feature is of particular importance to us as our main goal is to provide child welfare agencies with clear and straightforward insights into their performance.

The application domain of DEA has grown widely, having been applied to many real-world cases (Liu et al., 2013). Beyond the use of DEA in private sector applications such as banking and manufacturing, it has also been widely applied in the public, social, nonprofit, and healthcare sectors (Martin, 2002). Healthcare is one such notable sector of application for DEA, dating back to at least the 1980s (Kohl et al., 2019). A DEA assessment has been conducted to evaluate nursing service efficiency in a group of Wisconsin hospitals (Nunamaker, 1983). DEA has also been used to identify relatively inefficient hospitals among a set of teaching hospitals in Massachusetts (Sherman, 1984). Many researchers have since conducted empirical studies analyzing the efficiency of healthcare systems through the use of DEA (Bonasia et al., 2020). DEA applications in healthcare, focusing on hospitals, have also been reviewed (Kohl et al., 2019). The application of DEA to evaluate the quality of care in nursing homes has also been studied (Garavaglia, Lettieri, Agasisti, & Lopez, 2011), analyzing the efficiency of 40 nursing homes in the northwestern area of Lombardy, Italy, from 2005 to 2007. Similarly, DEA has been used for evaluating the efficiency of local councils in England regarding the quality of services delivered for older adults (Iparraguirre & Ma, 2015).

In the education sector, DEA has proven beneficial as well. A notable study by Johnson and Ruggiero (2014) applied DEA to investigate 604 school districts in Ohio, finding that technical progress plays an important role in the performance of these schools.

Contreras and Lozano (2020) also used DEA to evaluate the performance of nine public Spanish universities, examining the impact of the additional resources allocated to these institutions. In addition, DEA applications have been deployed in practice in social service settings as well. Large-scale implementation of DEA was conducted in more than 1,000 humanitarian organizations operating in disaster relief, emergency communications, and life-saving skills training (Medina-Borja et al., 2007). Moreover, the use of DEA in performance measurement in NPOs has been studied (Vakkuri, 2003). The author presented prominent research papers as well as common inputs and outputs used in four areas: higher education, fundraising activities, cultural activities, and healthcare environment. A comprehensive review of the utilization of DEA models in the public sector also exists (Ahn et al., 2018). However, despite various uses of DEA in these areas, to the best of our knowledge, there are no known uses of DEA to assess the performance of child welfare agencies.

3. Methodology

We leverage the framework of the Design Science Research Methodology (DSRM) to improve the structure and clarity of our research objectives and methodology. DSRM is a systematic approach for conducting research that consists of six main activities (Peffers et al., 2007.Charles, Aparicio, & Zhu, 2019).

In Table 1, we present five of these activities[1]. The first column of the table lists the DSRM activities, the second provides a description of each activity, and the third highlights the specific knowledge and resources used in these activities, such as models, methods, and theories.

Table 1. Design Science Research Methodology (DSRM) applied to the current study

DSRM Activities	Activity Description	Knowledge
Problem identification and motivation	Existing gaps in the literature on child welfare performance evaluation for not considering input resources in the evaluation process	Literature review; interview with child welfare practitioner
Define the objectives of a solution	Providing a more comprehensive approach by applying DEA to this domain	Knowledge of DEA analysis
Design and development	Developing a SBM output-oriented model with undesirable outputs	Non-radial models in DEA
Demonstration	Applying the proposed model to 43 US child welfare agencies	Applying DEA model to a real-world case
Evaluation	Comparison of states with and without Alternative Response (AR) policy based on their DEA scores	Understanding of the current child welfare system

We employed DEA to evaluate the performance of child welfare agencies across 43 states in the United States. The purpose was to identify the agencies that perform well in light of DEA scores, as well as to provide a benchmark for underperforming states. This benchmarking would serve as a reference for improving their outcomes.

[1]The sixth DSRM activity, communication with stakeholders, is not applicable in our research context.

3.1. Data Discussions

This section details the various modeling choices we make in our study including DMUs, inputs, and outputs. Our child welfare context is not a traditional production environment where each input directly influences the amount of produced outputs. Rather, the criteria we define as our inputs can influence the child welfare agency performance for each state, while the criteria we define as outputs are a subset of the data indicators used to evaluate the child welfare performance of each state. Our study is a multi-criteria performance analysis for the purpose of benchmarking high-performing states. As we use the methodology of DEA to conduct the performance analysis, we thus have associated inputs and outputs so as to align with DEA terminology, as opposed to a strict production context.

3.1.1. Decision Making Units (DMUs)

While it was our intent to include all 50 US states plus the District of Columbia and Puerto Rico as the DMUs in our analysis, we discovered that some of the chosen inputs and outputs were missing data. This led us to omit a few DMUs from our study including Delaware, Idaho, Kentucky, Maryland, North Carolina, Oregon, Pennsylvania, Puerto Rico, and Wyoming, resulting in 43 final DMUs. Going forward, we refer to all DMUs as states despite the District of Columbia not being a state.

3.1.2. Selected Inputs

We considered two input variables. The first indicator pertains to child welfare spending for FY 2018, encompassing both federal and state/local expenditures (Rosinsky et al., 2021). Funding is an input variable because it represents the resources available for achieving desired performance. To account for the different sizes of states and to avoid inadvertently biasing the analysis to favor states of a particular size, funding was standardized across states. Standardization was achieved by dividing the total funding of each state by the number of children under the age of 18 within each respective state in the year 2018.

The second indicator is the rate of maltreatment in the state (Children's Defense Fund, 2020). The rate of maltreatment is an input variable because it represents the size of the workload for the child welfare agency. This rate is quantified by calculating the total reported cases of child victims of various forms of abuse and neglect (such as medical neglect and physical abuse) per 1,000 children in the state.

3.1.3. Selected Outputs

One primary way to evaluate the performance of child welfare agencies across states is through CFSR statewide data indicators. Consequently, we have utilized a subset of the indicators provided by the third and most recent CFSR process round as our DEA outputs. The statewide data indicators from the CFSR data dictionary (Children's Bureau, 2020a) are, verbatim:

(1) **Maltreatment in Foster Care.** Of all children in foster care during a 12-month period, what was the rate of victimization per 100,000 days of care?

(2) **Recurrence of Maltreatment.** Of all children who were victims of a substantiated or indicated maltreatment report during a 12-month period, what percent were victims of another substantiated or indicated maltreatment report within 12 months of the initial victimization?

(3) **Re-entry to Foster Care in 12 Months.** Of all children who enter care in a 12-month period, who are discharged within 12 months to reunification, live with a relative, or guardianship, what percent re-entered care within 12 months of their discharge?

(4) **Permanency in 12 Months for Children Entering Care.** Of children who enter care in a 12-month period, what percent are discharged to permanency within 12 months of entering care?

These indicators were measured and assessed in FFY 2018, while for the three remaining indicators[2], there was a lag between the year of their measurement and when they were assessed because of the inherent characteristics of these variables. As FFY 2018 had more comprehensive data available for other input indicators, this year was chosen for conducting our DEA analysis. Thus to maintain consistency in our study, we excluded the indicators assessed in subsequent years. These indicators exist in two forms in the CFSR data. The first is observed performance, which describes how a state performed on a given indicator without any adjustments. The second is Risk-Standardized Performance (RSP), which attempts to remove the impact of certain factors that influence indicator performance and are beyond the control of states. For instance, the age of children in care can significantly impact outcomes like achieving permanency, regardless of the quality of care provided by a state. To identify the impact of these factors, in RSP logistic regression is used for binary outcomes and Poisson regression for outcomes with count per unit of time (Children's Bureau, 2015). Because accounting for such factors enables a fairer comparison of each state's performance, we have chosen RSP data as our final output indicator.

3.2. DEA Modeling

Several models have been proposed to calculate DEA scores for DMUs and can be categorized as radial, or non-radial. In radial models, it is assumed that inputs and outputs change proportionally. However, in reality, disproportionate changes in inputs and outputs are often expected and thus require estimation of disproportionate changes using non-radial models to provide a more realistic representation of the systems under evaluation (Avkiran & McCrystal, 2012). In this context, several non-radial DEA models have been developed, including the notable Slacks-Based Measure (SBM) introduced by Tone (2001) that has been broadly used in many studies.

We first highlight the importance of the input and output slacks in the context of DEA. Input slack refers to the amount by which a DMU can reduce its input usage without affecting its output levels, while output slack indicates the potential increase in outputs without additional inputs. In DEA models, it is probable that non-zero input and output slacks emerge after determining the efficiency score. These non-zero slack values indicate a certain degree of inefficiency. Therefore, to more accurately measure the performance of DMUs, it is key to also account for these non-zero slacks. The slacks-based model can work under constant returns to scale (CRS) and variable returns to scale (VRS). In the CRS slacks-based model, it is assumed that all DMUs operate under constant returns to scale, meaning that the scale of resource usage remains the same as the DMU moves towards efficiency. In the VRS slacks-based model, it is assumed that DMUs can operate under variable returns to scale, allowing for changes in the scale of resource usage as they move towards efficiency. In the context

[2]These remaining indicators include: Permanency in 12 Months for Children in Care 12-23 Months, Permanency in 12 Months for Children in Care 24 Months or More, and Placement Stability

of child welfare operations, because the change in inputs would not necessarily result in the change of outputs on the same scale, we consider the slacks-based model under the VRS assumption. Moreover, as our primary focus is to improve the CFSR indicators as outputs, we opt for an output-oriented model.

The output-oriented DEA model assumes that it is desirable to produce more outputs relative to fixed input resources. In some situations, a DEA model variant that accommodates undesirable input(s) or output(s) is needed. Koopmans mentioned in his study that the production process may also generate undesirable outputs such as smoke pollution or waste (Koopmans, 1951). In this situation, the undesirable outputs should be reduced to improve DEA scores, suggesting that undesirable and desirable outputs be treated differently (Seiford & Zhu, 2002).

Several methods have been proposed in the DEA literature to account for undesirable outputs and inputs, such as the nonparametric approach by Färe et al. (1989) and classification invariance method by Seiford and Zhu (2002). We adopt a method for including undesirable outputs in the SBM model first proposed by Tone (2004), which has seen numerous SBM-employing studies also use this strategy to explore performance evaluation (Apergis et al., 2015; Chen et al., 2021; Zhang & Choi, 2013). The model used in this study is in the Appendix A.

Our study features both undesirable outputs *and* inputs. These outputs include the maltreatment rate in foster care, the recurrence of maltreatment, and the re-entry rate of children. As these indicators should clearly be lower to raise the DEA scores, we consider them undesirable outputs. Concerning inputs, the maltreatment rate in the state is also considered undesirable, due to its negative impact on agency performance. While increased child welfare spending is considered desirable as it has a logically positive impact on the agency outputs, a higher maltreatment rate requires greater resource allocation to manage the high volume of reports. This often results in fewer resources available for improving the CFSR outputs, thereby negatively affecting the performance of the agency. Therefore, we modeled desirable and undesirable inputs with separate constraints to reflect their distinct effects.

We analyze n DMUs within this context. Each DMU has four factors: desirable inputs, undesirable inputs, desirable outputs, and undesirable outputs as represented by four matrices: $X^g \in \mathbb{R}^{m_1 \times n}$, $X^b \in \mathbb{R}^{m_2 \times n}$, $Y^g \in \mathbb{R}^{s_1 \times n}$ and $Y^b \in \mathbb{R}^{s_2 \times n}$, with the corresponding number of factors denoted by m_1, m_2, s_1, and s_2.
Define the input and output matrices X^g, X^b, Y^g, and Y^b as follows: $X^g = [x^g_{i_g j}] = [x^g_{11}, \ldots, x^g_{m_1 n}]$, $X^b = [x^b_{i_b j}] = [x^b_{11}, \ldots, x^b_{m_2 n}]$, $Y^g = [y^g_{r_g j}] = [y^g_{11}, \ldots, y^g_{s_1 n}]$, and $Y^b = [y^b_{r_b j}] = [y^b_{11}, \ldots, y^b_{s_2 n}]$. Assume that $X^g \geq 0$, $X^b \geq 0$, $Y^g \geq 0$, and $Y^b \geq 0$. Then the operational set O is defined by:

$$O = \left\{ (x^g, x^b, y^g, y^b) \mid x^g \geq X^g \lambda,\ x^b \leq X^b \lambda,\ y^g \leq Y^g \lambda,\ y^b \geq Y^b \lambda \right\},$$

where $\lambda \in \mathbb{R}^n$ is the intensity vector, and the four inequalities in the set O respectively represent the actual desirable input levels that should meet or exceed the frontier desirable input level, the actual undesirable input level that is less than or equal to the frontier undesirable input level, the actual desirable output levels that are no larger than the frontier desirable output level, and the actual undesirable output that is greater than or equal to the frontier of the undesirable output level. Inspired by the theoretical model of Tone (2004), the output-oriented SBM model under VRS conditions and accommodating undesirable inputs and outputs for evaluating a specific DMU, denoted as DMU_0 $(x^g_0, x^b_0, y^g_0, y^b_0)$ is as follows:

$$\frac{1}{\rho^*} = \max\left\{1 + \frac{1}{s_1 + s_2}\left(\sum_{r_g=1}^{s_1}\frac{s^g_{r_g}}{y^g_{r_g\text{o}}} + \sum_{r_b=1}^{s_2}\frac{s^b_{r_b}}{y^b_{r_b\text{o}}}\right)\right\} \tag{1a}$$

$$\text{subject to} \quad x^g_{i_g\text{o}} = \sum_{j=1}^{n} x^g_{i_gj}\lambda_j + t^g_{i_g} \quad i_g = 1, ..., m_1, \tag{1b}$$

$$x^b_{i_b\text{o}} = \sum_{j=1}^{n} x^b_{i_bj}\lambda_j - t^b_{i_b} \quad i_b = 1, ..., m_2, \tag{1c}$$

$$y^g_{r_g\text{o}} = \sum_{j=1}^{n} y^g_{r_gj}\lambda_j - s^g_{r_g} \quad r_g = 1, ..., s_1, \tag{1d}$$

$$y^b_{r_b\text{o}} = \sum_{j=1}^{n} y^b_{r_bj}\lambda_j + s^b_{r_b} \quad r_b = 1, ..., s_2, \tag{1e}$$

$$\sum_{j=1}^{n} \lambda_j = 1, \tag{1f}$$

$$t^g_{i_g}, t^b_{i_b}, s^g_{r_g}, s^b_{r_b}, \lambda_j \geq 0 \quad \forall i_g, i_b, r_g, r_b, j, \tag{1g}$$

where t^g, t^b, s^g, and s^b correspond to the slacks in desirable inputs, undesirable inputs, desirable outputs, and undesirable outputs, respectively. The value of ρ^* is the efficiency value of the DMU_0. The parameters m_1, m_2, s_1, and s_2 denote the number of desirable inputs, undesirable inputs, desirable outputs, and undesirable outputs, respectively. The indices i_g, i_b, r_g, and r_b are used to navigate through these inputs and outputs, allowing for a detailed and structured analysis of each DMU. In the presented SBM model, Equations 1b and 1c describe how the actual levels of desirable and undesirable inputs for a specific DMU can be expressed as a linear combination of the inputs of other DMUs plus any excess or surplus. Similarly, Equations 1d and 1e correspond to the level of desirable and undesirable outputs. Equation 1f describes the condition that ensures the sum of the intensity vectors equals one, signifying a variable return to scale (VRS) condition in the model. Additionally, the model imposes a lower bound condition of non-negativity on the slack variables and the intensity vector as expressed in Equation 1g. Generally, if the model finds that a DMU can increase its level of desirable outputs or decrease its level of undesirable outputs, it would suggest that the DMU is not operating efficiently. Otherwise, if $\rho^* = 1$ and $s^g = s^b = 0$ the DMU is deemed efficient.

In DEA, it is important to ensure that all compared DMUs are relatively homogeneous to ensure a fair efficiency assessment. Homogeneity implies three main aspects: engaging in similar processes, using consistent efficiency metrics based on chosen inputs and outputs, and operating under comparable conditions (Haas & Murphy, 2003).

In our context, while the first two criteria are met, the third condition is not satisfied due to fundamental differences in the operational systems among our DMUs. These differences arise from whether a state's child welfare agency has adopted an *Alternative Response (AR)* policy. The presence of AR policy in a state can reduce the number of children entering foster care and their future recurrence in the child welfare system (Children's Bureau, 2020c). One potential strategy to induce homogeneity is grouping DMUs into homogeneous clusters (Haas & Murphy, 2003). Accordingly, we have chosen the presence or absence of an AR policy for clustering our DMUs.

Alternative response policy enables child welfare agencies to customize their response to child abuse and neglect reports. So, rather than initiating an investigation for each screened-in family, the agency employs an alternative response model offering various pathways for addressing reports that have met screening criteria. These distinct pathways are determined by factors such as the type and intensity of the alleged maltreatment, the number of prior reports, and the willingness of the parent to cooperate in resolving the safety issues (Children's Bureau, 2022). Alternative responses encourage community agencies to support families where children are at low risk for abuse or neglect, enabling the public child welfare agency to concentrate on serving families where children are at high risk for abuse or neglect.

An important aspect of AR policy is its effect on the workload of child welfare agencies. In states without AR, agencies handle all screened-in reports, regardless of their risk level, which can result in a high caseload. Conversely, states with AR practices can prioritize their resources and attention on high-risk cases, as lower-risk situations are addressed through alternative pathways. This stratification reduces the overall number of cases requiring full-scale investigation and intervention. Consequently, the presence of an AR policy can substantially affect the performance of a child welfare agency. To assess the association between the adoption of AR practices and the inputs and outputs in our model, we conducted an independent sample t-test analysis. Our findings revealed a statistically significant relationship between one of the input variables and one of the output variables. Specifically, there was a significant relationship between the presence of an AR practice and the input variable of the rate of maltreatment in the state (p-value=0.03) as well as with the output variable of the rate of recurrence of maltreatment (p-value=0.023). We therefore partitioned the DMUs into two groups: 26 DMUs with an AR practice and 17 DMUs without, and compared each DMU only with other DMUs in the same cluster. Figure 1 illustrates the geographical distribution of states based on the adoption of AR policy.[3]

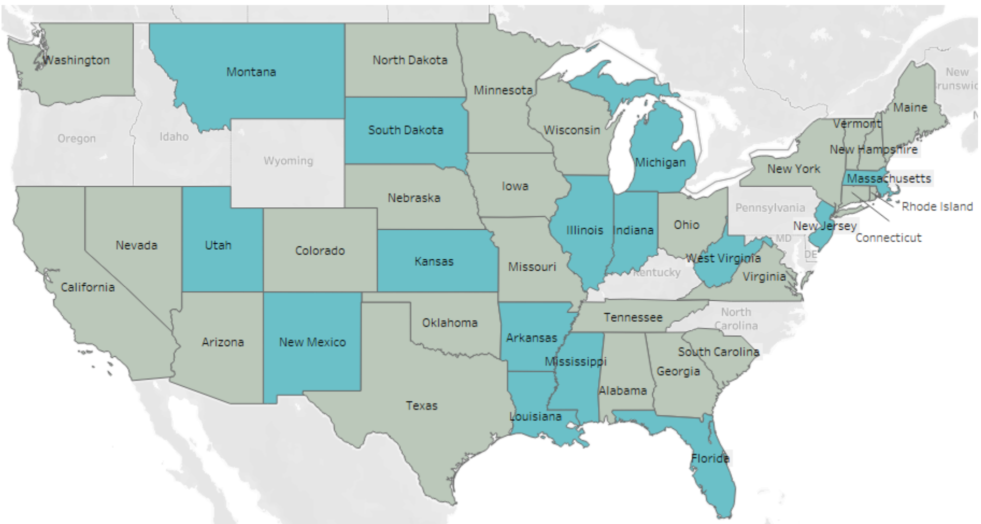

Figure 1. Geographical distribution of states by AR policy status. States with AR policy are colored in darker gray, states without AR policy are colored in blue, while states excluded from this study are colored in lighter gray

[3]It is important to note that some of the states shaded as having adopted AR policy are, in fact, hybrid states. Within these states, there exist counties that do not implement the AR policy.

4. Results and Discussion

We have applied the aforementioned output-oriented VRS slacks-based model to evaluate the performance of child welfare agencies among all states and within the two homogeneous clusters. The DEA scores of all 43 DMUs are shown in Table 2, with each state displayed in the second column. The third column displays the DEA scores of each DMU in comparison to the other 42 DMUs and itself, regardless of their cluster. The fourth and fifth columns report DMU scores under the cluster-wise frontier, accounting for the absence or presence of AR practices, respectively.

Table 2. DEA score of states based on overall and cluster-wise (with and without AR) frontiers.

DMU	State	Overall	Cluster With AR	Cluster Without AR
1	Alabama	1.00	1.00	–
2	Arizona	1.00	1.00	–
3	California	0.74	0.75	–
4	Colorado	0.86	0.86	–
5	Connecticut	0.74	0.75	–
6	Georgia	1.00	1.00	–
7	Hawaii	0.92	0.92	–
8	Iowa	0.70	0.70	–
9	Maine	0.67	0.68	–
10	Minnesota	0.80	0.81	–
11	Missouri	1.00	1.00	–
12	Nebraska	0.93	0.93	–
13	Nevada	1.00	1.00	–
14	New Hampshire	1.00	1.00	–
15	New York	0.64	0.64	–
16	North Dakota	1.00	1.00	–
17	Ohio	0.77	0.79	–
18	Oklahoma	1.00	1.00	–
19	Rhode Island	0.74	1.00	–
20	South Carolina	1.00	1.00	–
21	Tennessee	1.00	1.00	–
22	Texas	1.00	1.00	–
23	Vermont	0.82	0.85	–
24	Virginia	1.00	1.00	–
25	Washington	0.86	0.84	–
26	Wisconsin	0.93	0.96	–
27	Alaska	0.70	–	0.72
28	Arkansas	1.00	–	1.00
29	District of Columbia	0.74	–	0.74
30	Florida	0.81	–	0.82
31	Illinois	0.52	–	0.54
32	Indiana	1.00	–	1.00
33	Kansas	0.81	–	0.83
34	Louisiana	1.00	–	1.00
35	Massachusetts	1.00	–	1.00
36	Michigan	0.70	–	1.00
37	Mississippi	1.00	–	1.00
38	Montana	0.82	–	1.00
39	New Jersey	1.00	–	1.00
40	New Mexico	1.00	–	1.00
41	South Dakota	0.81	–	1.00
42	Utah	1.00	–	1.00
43	West Virginia	1.00	–	1.00

Because each DMU is examined within its cluster here, the performance of each DMU in a cluster with and without AR practice is compared to the other 25 and 16 DMUs, respectively. We note that while clustering has led to fewer DMUs in each group, our partition still satisfies the rule of thumb that the number of DMUs is at least twice the total number of inputs and outputs (Golany & Roll, 1989). Table 3 provides summary statistics of the performance score of each state shown in Table 2. Out of 43 states, 21 are found to be best performing, with an average DEA score of 0.884 under the overall frontier.

Table 3. Summary statistics of DEA scores.

Frontier	# of DMUs	Overall Frontier			Cluster-wise Frontier		
		Mean Score	# of DMUs with Score 1	% of DMUs with Score 1	Mean Score	# of DMUs with Score 1	% of DMUs with Score 1
Cluster 1 (no AR)	17	0.8784	9	52.94	0.9206	12	70.59
Cluster 2 (AR)	26	0.8888	12	46.15	0.9033	13	50.00
Overall	43	0.8847	21	48.83	–	–	–

The percentages of states that are best-performing in clusters with and without AR practice under the overall frontier are 46.15% and 52.94%, respectively. The DEA scores under the cluster-wise frontier are at least as large as those in the overall frontier because smaller reference sets reduce competition and allow DMUs within these clusters to achieve higher scores compared to those assessed against the more diverse and extensive overall frontier.

The number of best-practice DMUs in cluster 1 (DMUs without an AR practice) increases to 12, and an additional three DMUs gain a score of one: Michigan, Montana, and South Dakota. In cluster 2 (DMUs with an AR practice), the only additional best-performance DMU is Rhode Island. The percentages of best-practice DMUs in cluster 1 and cluster 2 are 70.59% and 50%, with average scores of 0.9206 and 0.9033, respectively. Because cluster 1 has fewer DMUs compared to cluster 2, it is expected that the mean DEA score and proportion of best-practice DMUs in this cluster are greater. Figure 2 visually shows the details of the quartiles for the DEA scores in both clusters. Although Figure 2 and the ANOVA results suggest a lack of significant differences in performance scores across groups, supported by statistics such as the F-value of 0.324, a non-significant p-value of 0.861, and a small estimated effect size ($\eta^2 = 0.010$), the subsequent paragraphs provide a more detailed understanding of the performance scores derived from DEA. We further performed a sensitivity analysis by sequentially excluding one output at a time; the results of this analysis are presented in Appendix B.

4.1. Benchmarking

We use our DEA results to create reference sets that can serve as benchmarks for evaluating the relative score of non-frontier DMUs. These reference sets comprise states that operate under circumstances similar to those of the state being compared and serve as exemplars of best practices. Reference sets provide guidance for making well-informed decisions regarding improvements. Table 4 reports the reference set for each non-frontier DMU under the cluster-wise frontier. We provided reference sets in the

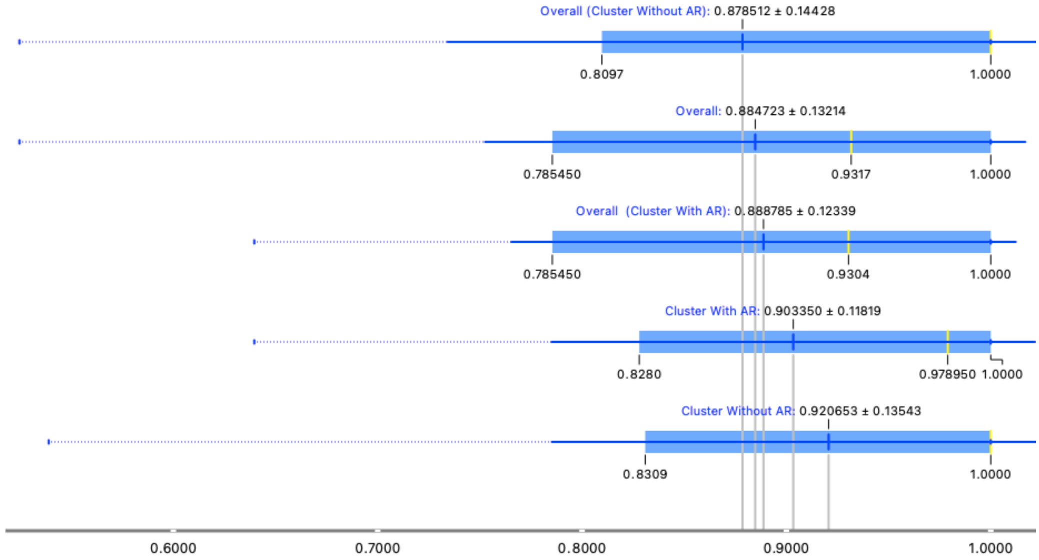

Figure 2. Details of quartiles for the two clusters.

cluster-wise frontier to ensure that each DMU is being compared to the DMUs performing under similar operational conditions, thus offering a relevant and achievable benchmark for evaluation. As observed, Louisiana is the reference for all non-frontier DMUs in the cluster without using the AR policy. West Virginia exhibits a similar pattern, with the exception of not being a reference point for Kansas.

Table 4. Target sets for non-frontier DMUs.

Cluster	DMUs	References	Maltreatment in Foster Care		Recurrence of Maltreatment		Permanency		Re-entry Rate	
			Target Value	Percent Change	Target Value	Percent Change	Target Value	Percent Change	Target Value	Percent Change
1	AK	LA, MS, WV	10.63	-52.31%	11.72	-46.95%	44	37.50%	5.02	-20.17%
	DC	LA, WV	5.51	-27.03%	8.77	-57.79%	41.95	47.19%	6.13	-5.63%
	FL	LA, WV	6.48	-42.63%	8.6	-6.48%	45.71	17.83%	6	-22.04%
	IL	LA, MS, WV	6.02	-67.09%	8.7	-45.60%	43.84	227.22%	6.06	-2.20%
	KS	LA	6.51	-30.08%	8.6	-1.15%	45.8	34.71%	6	-15.49%
2	CA	GA, ND	3.8	-62.50%	6.4	-38.45%	37.63	11.02%	7.66	-20.13%
	CO	GA, AL, SC	10.58	-21.39%	10.6	0.00%	55.4	0.00%	7.18	-46.37%
	CT	GA, ND	4.54	-28.38%	8.48	-26.89%	40.98	59.46%	7.26	-16.55%
	HI	GA, NH	3.58	-13.79%	4.83	-15.26%	37.72	2.78%	12.46	-1.04%
	IA	AR, SC	10.81	-68.54%	11.45	-40.93%	57.46	48.10%	7.03	-14.23%
	ME	ND, SC	7.16	-27.58%	10.32	-31.19%	48.78	84.08%	7.13	-47.51%
	MN	GA, SC	6.68	-42.08%	8.01	-13.83%	45.9	0.00%	7.63	-36.40%
	NE	MO, NV	7.04	-23.88%	7.77	-5.22%	38.2	0.00%	5.5	0.00%
	NY	SC	10.77	-65.48%	11.9	-49.58%	58.8	67.52%	7.2	-42.86%
	OH	GA, NV, SC	7.58	-48.17%	9.67	-26.67%	46.6	0.00%	6.15	-31.64%
	VT	GA, ND	4.07	-18.80%	7.15	-3.28%	38.85	23.34%	7.51	-24.81%
	WA	GA, NV, TX	6.92	-34.24%	8.72	-28.52%	42	11.43%	5.7	0.00%
	WI	GA, AL, NH, SC	4.63	-17.57%	5.4	0.00%	39.5	0.00%	11.4	0.00%

In the cluster with AR policy, Georgia has the highest rate of serving as a reference: 9 out of 13 states. South Carolina follows as the next most frequently referenced state, serving as the reference DMU for 7 states. Figure 3 further illustrates the number of underperforming states for which best-practice states serve as references in the cluster-wise frontier.

After determining the reference sets, we also suggested areas and amounts of improvement that a non-frontier DMU should pursue to reach the frontier line. This could be seen as setting targets for each non-frontier DMU to direct them toward improving their performance. Table 4 reports the target values for each output (*Maltreatment in Foster care*, *Recurrence of Maltreatment*, *Permanency in Foster Care*, and *Re-entry Rate*). It also shows what percentage a DMU needs to change the current level of its outputs to reach those targets. The agencies can identify areas requiring more significant changes and prioritize accordingly.

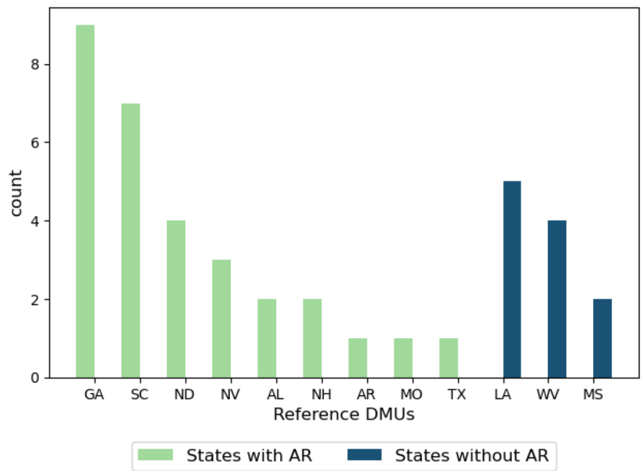

Figure 3. Count of occurrences that best practice DMUs serve as a reference for non-frontier DMUs.

Figure 4 displays a bar chart comparing the average percentage of change between two clusters. It shows that, on average, non-frontier DMUs in the cluster without an AR policy need to adjust their output levels more than those in the cluster with an AR policy, except for the re-entry rate of children.

We further developed and introduced an interactive dashboard in Tableau. This dashboard dynamically showcases the DEA scores, reference sets and target variables for each non-frontier state, offering an intuitive and user-friendly interface for exploring the DEA results. The dashboard enables stakeholders to select a state and immediately view the corresponding reference sets and targets that the state should aim for to improve its performance. This interactive tool helps the stakeholders translate the analytical DEA findings into actionable recommendations. A screenshot of the dashboard in action is illustrated in Figure 5.

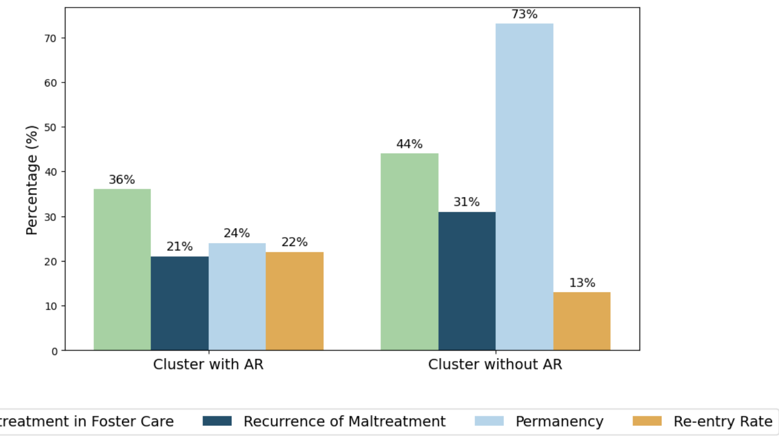

Figure 4. Potential improvement of non-frontier DMUs in each cluster.

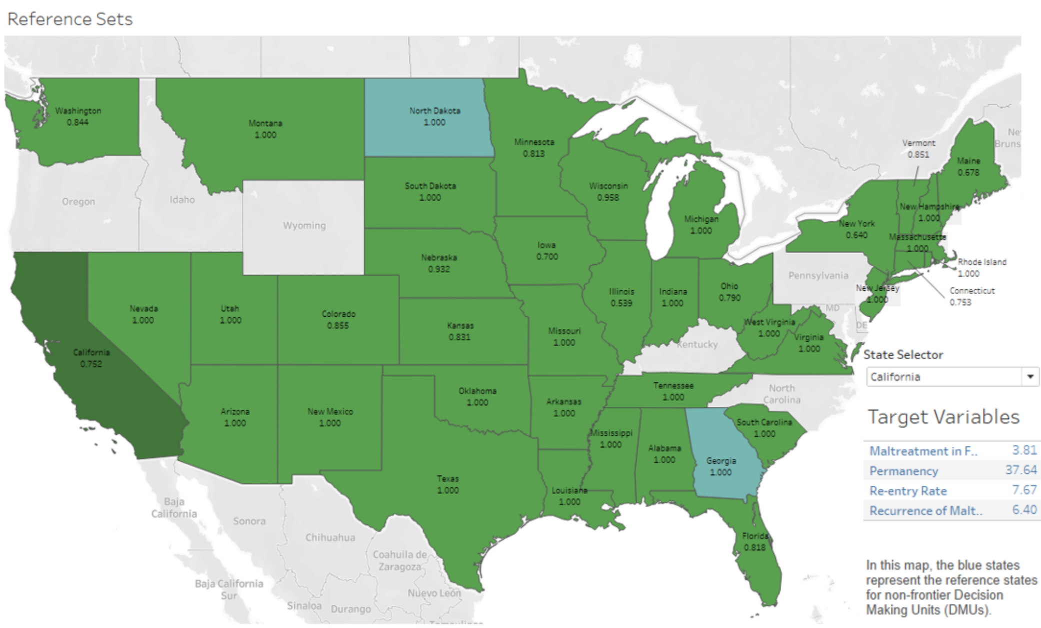

Figure 5. Interactive dashboard visualization in Tableau[4].

In the example provided in Figure 5, the Tableau dashboard is set to display data for California, selected as a non-frontier state using the state selector feature. The reference states for California, which are Georgia and North Dakota, are highlighted in distinct colors on the map. The dashboard also presents a table of target variables of the four outputs for California.

4.2. Societal Implications

This research, which illustrates the applicability of DEA for assessing the performance of child welfare agencies, has implications for improving the practice and outcomes of

[4]A link to the corresponding visualization can be found at: `https://public.tableau.com/app/profile/sepideh.sedghi/viz/reference_sets/Dashboard1?publish=yes`

child welfare and for illustrating the potential of applying advanced analytic techniques to child welfare.

4.2.1. Enhanced Child Safety and Permanency

While recognizing that the CFSR is a federally legislated process for assessing the performance of child welfare agencies, opportunities for applying advanced analytics to child welfare need to be embraced. The private sector benefits from insights obtained through sophisticated mathematical modeling. Public sector child welfare agencies should have the same opportunities to improve practices aimed at increasing safety and permanency for children.

4.2.2. Transparency and Benchmarking

By summarizing multiple inputs and outputs into a single efficiency score, DEA translates performance into an understandable measure, thus providing transparency into the practice of child welfare across states. Such transparency may lead to improved public understanding of the goals of child welfare and to increased confidence in the efficiency of these public agencies.

Both the CFSR methodology for the statewide data indicators and the DEA analysis in this study focus on the relative performance of states, rather than the comparison of states' outputs to a criterion-referenced standard. However, in contrast to the CFSR, which compared all states to each other, this research study grouped states into two categories (i.e., those that use Alternative Response and those that do not) before conducting DEA. This categorization promotes acceptance of the use of benchmarking for continuous quality improvement by reducing concerns about the lack of comparability of peer states.

4.2.3. Resource Allocation

Effective resource allocation is important for enhancing the performance and impact of child welfare agencies. Our findings can inform decisions related to resource allocation within these agencies. In our DEA analysis, it was revealed that certain states, despite having lower levels of spending, achieve better outcomes compared to others with higher expenditures. This observation suggests the potential for benchmarking against those states that achieve the best performance at the best cost. By analyzing the resource allocation strategies of these states, other agencies can gain insights that can enhance both performance and financial stewardship goals.

4.2.4. Actionable Recommendations

We develop and offer actionable recommendations for state child welfare agencies to progress toward their established goals and improve their performance. These recommendations focus on the CFSR indicators considered as outputs.

Strategies for reducing the maltreatment rate in foster care could include: effective implementation of policies that prioritize safe and supportive family time visits between parents and removed children; licensing standards and contracting specifications for foster care settings that prioritize child safety through staffing ratios, training programs, and supervisory standards; and regular announced and unannounced visits from oversight agencies. Recommendations for addressing the high recurrence of maltreatment could include a data-informed approach that assesses the accuracy of an agency's risk assessment method and makes improvements through the use of predictive analytics; reliance on best-practice safety planning techniques; and provision of evidence-based practices that enable families to maintain their children safely at home.

Regarding reducing the re-entry rate to foster care, effective strategies could include comprehensive planning for parent-child reunification that starts soon after a child's home removal. This planning could be most effective when it includes the provision of skill-building and supportive services to parents while children are out of the home and continue after children return home to provide stabilization that supports an enduring reunification. Lastly, strategies for promoting the permanency of children in 12 months after entering foster care could include the use of technology-aided family finding techniques; reliance on supervised family time visits during which coaching in effective parenting techniques is delivered to build caregiver capacity; and consistent use of practices such as permanency reviews that rely on structured protocols for case consultation, collaboration, and timely follow up on action items.

In addition to these recommendations, each non-frontier agency can specifically benefit from benchmarking against the nearest frontier DMU, as shown in Table 4. We provide two examples of this benchmarking process: one for the cluster with the AR policy and one for the cluster without. This analysis serves to illustrate what is possible, leaving the conducting of such analyses for all non-frontier DMUs for future research.

For the cluster with an AR policy, we compared recent progress reporting for the states of Georgia and Connecticut. Georgia was selected for having the highest reference rate, while Connecticut was chosen because it had the lowest DEA score among those referenced to Georgia. The review of the reports from the two states revealed some topics emphasized in the Georgia report that were mentioned in less detail or were absent from the Connecticut report. These topics included creative use of technology, collaborations with schools and community-based programs, and comprehensive training programs for child welfare staff. While not prescriptive, Connecticut could consider Georgia's initiatives, like the Click Safe mobile emergency response tool, as a reference when identifying areas for potential improvement.

For the cluster without an AR policy, we compared recent progress reporting for the states of Louisiana and Illinois. One potential aspect that Illinois could consider is the structure of Louisiana's kinship care program. Louisiana provides specific subsidies and assistance for kinship caregivers to help stabilize placements and prevent re-entry into foster care. Illinois also supports kinship caregivers but might benefit from the more targeted approach used in Louisiana. Another difference between the published policies and practices of Louisiana and Illinois is the emphasis that Louisiana places on conducting extensive public education campaigns on Safe Haven laws for the surrender of newborn infants. Louisiana's thorough Safe Haven public awareness campaigns might serve as an insightful reference for Illinois in the development of their own campaigns.

5. Conclusion

This study represents a pioneering effort in the application of DEA in the domain of child welfare, particularly by contributing to the essential concept of measuring outputs in relation to *inputs*. To our knowledge, historic efforts to evaluate the performance of public child welfare focused exclusively on outcomes without considering that the amount of resources available to achieve those outcomes could affect levels of performance. By analyzing a subset of statewide data indicators from the CFSR and including both rates of maltreatment and funding as inputs, our research introduces a new analytic approach to understanding the performance of child welfare agencies.

In addition to evaluating the performance of these child welfare agencies, our research identified areas where non-frontier agencies may make improvements. The study provides such agencies with a performance improvement roadmap for establishing specific targets and possible improvements. Such guidance provides public child welfare agencies with practical tactics to inform performance improvements.

Our study does come with its limitations, which could provide an opportunity for further research. We recognize the exclusion of certain variables in our DEA study that can impact each DMU's performance. Notably, not all CFSR data indicators were used as outputs and potential inputs; for example, the number of public child welfare agencies in each state (i.e., county-level vs. state-level administration) and the size of each state's child welfare workforce were not considered. Furthermore, our study was constrained by the unavailability of data for certain variables, leading to the exclusion of several states from our analysis. Including these states could have resulted in a more thorough investigation. This limitation highlights opportunities for future research to extend the scope of our study. Even so, we have accomplished the main objective of our study, which was to explore the applicability of DEA as an analytic tool for assessing the performance of public child welfare agencies and for providing a framework for identifying potential areas for improvement.

Acknowledgments

The authors would like to express their gratitude to Worcester Polytechnic Institute for the continuous support it provides to conduct this research.

Declaration of competing interest

The author affirms that there are no conflicts of interest or financial affiliations with any organizations that could benefit from the work submitted.

References

Ahn, H., Afsharian, M., Emrouznejad, A., & Banker, R. (2018). Recent developments on the use of DEA in the public sector. *Socio-Economic Planning Sciences*, *61*, 1–3.

Ahn, H., Carter, L. M., Reiman, S., & Hartzel, S. (2017). Development of a quality assurance and continuous quality improvement (CQI) model in public child welfare systems. *Journal of Public Child Welfare*, *11*(2), 166–189.

Ahn, H., DeLisle, D., & Conway, D. (2022). Child and Family Services Review (CFSR) and child welfare outcomes in the United States. *Journal of Public Child Welfare*, *16*(5), 679–703.

Amodeo, M., Bratiotis, C., & Collins, M. E. (2009). Examining perceptions of the impact of child and family services reviews on training: Reports from state training administrators. *Administration in Social Work*, *33*(4), 423–438.

Annie E. Casey Foundation. (2021). *Foster Care Explained: What it is, how it works and how it can be improved.* `https://www.aecf.org/blog/what-is-foster-care`.

Apergis, N., Aye, G. C., Barros, C. P., Gupta, R., & Wanke, P. (2015). Energy efficiency of selected OECD countries: A slacks based model with undesirable outputs. *Energy Economics*, *51*, 45–53.

Avkiran, N. K., & McCrystal, A. (2012). Sensitivity analysis of network DEA: NSBM versus NRAM. *Applied Mathematics and Computation*, *218*(22), 11226–11239.

Banker, R. D., Charnes, A., & Cooper, W. W. (1984). Some models for estimating technical and scale inefficiencies in Data Envelopment Analysis. *Management Science*, *30*(9), 1078–1092.

Belanger, K., Price-Mayo, B., & Espinosa, D. (2008). The plight of rural child welfare: Meeting standards without services. *Journal of Public Child Welfare*, *1*(4), 1–19.

Bonasia, M., Kounetas, K., & Oreste, N. (2020). Assessment of regional productive performance of European health systems under a metatechnology framework. *Economic Modelling*, *84*, 234–248.

Carnochan, S., Samples, M., Myers, M., & Austin, M. J. (2014). Performance measurement challenges in nonprofit human service organizations. *Nonprofit and Voluntary Sector Quarterly*, *43*(6), 1014–1032.

Charles, V., Aparicio, J., & Zhu, J. (2019). The curse of dimensionality of decision-making units: A simple approach to increase the discriminatory power of data envelopment analysis. *European Journal of Operational Research*, *279*(3), 929–940.

Charnes, A., Cooper, W. W., & Rhodes, E. (1978). Measuring the efficiency of decision making units. *European Journal of Operational Research*, *2*(6), 429–444.

Chen, Y., Ma, L., & Zhu, Z. (2021). The environmental-adjusted energy efficiency of China's construction industry: A three-stage undesirable SBM-DEA model. *Environmental Science and Pollution Research*, *28*, 58442–58455.

Children's Defense Fund. (2020). *The state of America's children 2020.* https://www.childrensdefense.org/wp-content/uploads/2020/02/The-State-Of-Americas-Children-2020.pdf. (Last accessed November 2023)

Children's Bureau. (2014a). *CFSR General Fact Sheet.* `https://www.acf.hhs.gov/sites/default/files/documents/cb/cfsr_general_factsheet.pdf`.

Children's Bureau. (2015). *Child and Family Services Review Technical Bulletin 8.* `https://www.acf.hhs.gov/sites/default/files/documents/cb/`

cfsr_tb8a.pdf.
Children's Bureau. (2020a). *CFSR Round 3 Statewide Data Indicators Data Dictionary.* https://www.acf.hhs.gov/sites/default/files/documents/cb/cfsr3_statewide_data_indicator_data_dictionary2020.pdf.
Children's Bureau. (2020b). *Children Maltreatment 2018.* https://www.acf.hhs.gov/sites/default/files/documents/cb/cm2018.pdf.
Children's Bureau. (2020c). *Differential response: A primer for child welfare professionals.* https://www.childwelfare.gov/pubPDFs/differential_response.pdf.
Children's Bureau. (2021). *CFSR Round 3 Statewide Data Indicators Workbook.* https://www.acf.hhs.gov/cb/report/cfsr-round-3-statewide-data-indicators-workbook.
Children's Bureau. (2022). *Child Maltreatment 2020.* https://www.acf.hhs.gov/sites/default/files/documents/cb/cm2020.pdf.
Contreras, I., & Lozano, S. (2020). Allocating additional resources to public universities. a dea bargaining approach. *Socio-economic planning sciences*, *71*, 100752.
Färe, R., Grosskopf, S., Lovell, C. K., & Pasurka, C. (1989). Multilateral productivity comparisons when some outputs are undesirable: A nonparametric approach. *The Review of Economics and Statistics*, 90–98.
Garavaglia, G., Lettieri, E., Agasisti, T., & Lopez, S. (2011). Efficiency and quality of care in nursing homes: an italian case study. *Health care management science*, *14*, 22–35.
Golany, B., & Roll, Y. (1989). An application procedure for DEA. *Omega*, *17*(3), 237–250.
Haas, D. A., & Murphy, F. H. (2003). Compensating for non-homogeneity in decision-making units in Data Envelopment Analysis. *European Journal of Operational Research*, *144*(3), 530–544.
Huggins-Hoyt, K. Y., Mowbray, O., Briggs, H. E., & Allen, J. L. (2019). Private vs public child welfare systems: A comparative analysis of national safety outcome performance. *Child Abuse & Neglect*, *94*, 104024.
Iparraguirre, J. L., & Ma, R. (2015). Efficiency in the provision of social care for older people. a three-stage data envelopment analysis using self-reported quality of life. *Socio-Economic Planning Sciences*, *49*, 33–46.
Johnson, A. L., & Ruggiero, J. (2014). Nonparametric measurement of productivity and efficiency in education. *Annals of Operations Research*, *221*, 197–210.
Kohl, S., Schoenfelder, J., Fügener, A., & Brunner, J. O. (2019). The use of Data Envelopment Analysis (DEA) in healthcare with a focus on hospitals. *Health Care Management Science*, *22*, 245–286.
Koopmans, T. C. (1951). An analysis of production as an efficient combination of activities. *Activity Analysis of Production and Allocation*.
Liu, J. S., Lu, L. Y., Lu, W.-M., & Lin, B. J. (2013). A survey of DEA applications. *Omega*, *41*(5), 893–902.
Martin, L. L. (2002). Comparing the performance of multiple human service providers using Data Envelopment Analysis. *Administration in Social Work*, *26*(4), 45–60.
Medina-Borja, A., Pasupathy, K. S., & Triantis, K. (2007). Large-scale Data Envelopment Analysis (DEA) implementation: A strategic performance management approach. *Journal of the Operational Rsearch Society*, *58*, 1084–1098.
Nunamaker, T. R. (1983). *Efficiency measurement and medicare reimbursement in nonprofit hospitals: An investigation of the usefulness of Data Envelopment Analysis.* PhD Dissertation, Graduate School of Business, University of Wiscon-

sin.

Nyhan, R. C., & Martin, L. L. (1999). Comparative performance measurement: A primer on Data Envelopment Analysis. *Public Productivity & Management Review*, 348–364.

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of Management Information Systems*, *24*(3), 45–77.

Rosinsky, K., Williams, S. C., Fischer, M., & Haas, M. (2021). *Child Welfare Financing SFY 2018.* https://cms.childtrends.org/wp-content/uploads/2021/03/ChildWelfareFinancingReport_ChildTrends_March2021.pdf.

Seiford, L. M., & Zhu, J. (2002). Modeling undesirable factors in efficiency evaluation. *European Journal of Operational Research*, *142*(1), 16–20.

Sherman, H. D. (1984). Hospital efficiency measurement and evaluation: Empirical test of a new technique. *Medical Care*, *22*(10), 922–938.

Tone, K. (2001). A slacks-based measure of efficiency in data envelopment analysis. *European Journal of Operational Research*, *130*(3), 498–509.

Tone, K. (2004). Dealing with undesirable outputs in DEA: A Slacks-Based Measure (SBM) approach. *Nippon. Opereshonzu Risachi Gakkai Shunki Kenkyu Happyokai Abus*, 44–45.

Vakkuri, J. (2003). Research techniques and their use in managing non–profit organisations–an illustration of DEA analysis in NPO environments. *Financial Accountability & Management*, *19*(3), 243–263.

Zhang, N., & Choi, Y. (2013). Environmental energy efficiency of China's regional economies: A non-oriented Slacks-Based Measure Analysis. *The Social Science Journal*, *50*(2), 225–234.

Appendix A. The original slacks-based model (SBM)

Introduced by Tone (2004) and considering undesirable outputs, this model is as follows:

$$\rho^* = \min \left\{ \frac{1 - \frac{1}{m}\sum_{i=1}^{m} \frac{s_i^-}{x_{i0}}}{1 + \frac{1}{s_1+s_2}\left(\sum_{r=1}^{s_1} \frac{s_r^g}{y_{r0}^g} + \sum_{r=1}^{s_2} \frac{s_r^b}{y_{r0}^b}\right)} \right\} \tag{2a}$$

subject to

$$x_0 = X\lambda + s^-, \tag{2b}$$

$$y_0^g = Y^g\lambda - s^g, \tag{2c}$$

$$y_0^b = Y^b\lambda + s^b, \tag{2d}$$

$$s^- \geq 0, \quad s^g \geq 0, \quad s^b \geq 0, \quad \lambda \geq 0. \tag{2e}$$

The vectors s^- and s^b correspond to excesses in inputs and bad outputs, respectively, while $\mathbf{s}^g$ expresses shortages in good outputs. In addition, $\lambda \in \mathbb{R}^n$ is the intensity vector, and the parameters m, s_1, and s_2 denote the number of inputs, desirable outputs, and undesirable outputs, respectively. Notations for x_0, y_0^g, y_0^b, X, Y^g, and Y^b are as denoted in section 3.2

Appendix B. Sensitivity Analysis

For sensitivity analysis, we ran additional experiments to check the robustness of our final selection of outputs. Our model initially included four outputs and two inputs. To explore the impact of each output on the DEA scores, we omitted one output at a time from the DEA model. This approach resulted in four alternative models:

- Model 2 excludes the output "Maltreatment in Foster Care"
- Model 3 excludes the output "Recurrence of Maltreatment"
- Model 4 excludes the output "Permanency for Children Entering Care"
- Model 5 excludes the output "Re-entry to Foster Care in 12 Months"

Each of these models was compared against the original model (Model 1), which included all specified outputs. This analysis allowed us to determine the sensitivity of our results to the inclusion of specific outputs and assess the robustness of the DEA scores under different output selections. As can be seen in Table 5, while there are variations in the average DEA scores and the number of frontier DMUs across models, there is no significant difference among the average scores, suggesting robustness to excluding individual outputs. However, the number of frontier DMUs varies considerably across models, indicating differing levels of discriminatory power. For example, the exclusion of "Re-entry to Foster Care in 12 Months" (Model 5) significantly reduces the number of frontier DMUs from 21 to 10, indicating the ability of this model to discriminate among DMUs. This sensitivity analysis provides deeper insights into why the exclusion of specific outputs leads to changes in the status of certain DMUs from frontier to non-frontier.

Model 2 (Excluding "Maltreatment in Foster Care"): Four DMUs, including Nevada, New Jersey, North Dakota, and Virginia, became non-frontier. These states performed well in "Maltreatment in Foster Care" yet their scores decreased when this output was excluded, with Alabama and Minnesota serving as reference DMUs. The fact that Alabama and Minnesota had better overall scores in other outputs, even with worse maltreatment rates, may explain why they were chosen as reference DMUs when this output was excluded.

Model 3 (Excludes “Recurrence of Maltreatment”): Five DMUs, including Arizona, Missouri, and Alabama, became non-frontier DMUs. For these states, Louisiana, Nevada, and South Dakota served as references, which all had less desirable performance in this output. It highlights the dependency of the former states on this specific output for achieving a DEA score of 1.

Model 4 (Exludes “Permanency for Children Entering Care”): Six DMUs, including Massachusetts, Indiana, Tennessee, Arizona, North Dakota, and New Jersey, became non-frontier. It shows the significant impact of this output on the DEA score of the affected states. Georgia and West Virginia, which were not as dependent on this output, became the references of these DMUs.

Model 5 (Excludes ”Re-entry to Foster Care in 12 Months”): Eleven DMUs, namely Louisiana, Texas, Mississippi, Oklahoma, Indiana, Arizona, Missouri, Tennessee, Nevada, New Jersey, and Virginia, became non-frontier DMUs. This shift was primarily due to the contrasting performance of New Hampshire and South Carolina, which stood out in the other three outputs and built a dominant frontier for these eleven states in Model 5. However, their relatively poorer performance on the re-entry rate enabled the other mentioned DMUs to depend on this output to maintain a DEA score of 1.

These findings demonstrate the varied impact of specific outputs on the score of DMUs and demonstrate how certain states rely more heavily on particular outputs than others.

Table 5. Sensitivity Analysis

State	Model 1	Model 2	Model 3	Model 4	Model 5
Alabama	1.00	1.00	0.85	1.00	1.00
Alaska	0.70	0.72	0.71	0.67	0.62
Arizona	1.00	1.00	0.94	0.79	0.88
Arkansas	1.00	1.00	1.00	1.00	1.00
California	0.74	0.75	0.73	0.69	0.70
Colorado	0.85	0.86	0.79	0.65	0.90
Connecticut	0.73	0.67	0.73	0.74	0.69
District of Columbia	0.74	0.68	0.78	0.73	0.67
Florida	0.81	0.83	0.78	0.74	0.75
Georgia	1.00	1.00	1.00	1.00	1.00
Hawaii	0.92	0.88	0.89	0.88	0.90
Illinois	0.52	0.47	0.46	0.70	0.42
Indiana	1.00	1.00	1.00	0.87	0.79
Iowa	0.69	0.74	0.69	0.66	0.65
Kansas	0.81	0.82	0.78	0.76	0.71
Louisiana	1.00	1.00	1.00	1.00	0.93
Maine	0.66	0.61	0.65	0.65	0.66
Massachusetts	1.00	1.00	1.00	0.61	1.00
Michigan	0.70	0.69	0.66	0.69	0.65
Minnesota	0.79	0.81	0.76	0.66	0.79
Mississippi	1.00	1.00	1.00	1.00	0.78
Missouri	1.00	1.00	0.86	1.00	0.78
Montana	0.82	0.86	0.84	0.78	0.75
Nebraska	0.92	0.91	0.88	0.84	0.74
Nevada	1.00	0.94	1.00	1.00	0.82
New Hampshire	1.00	1.00	1.00	1.00	1.00
New Jersey	1.00	0.92	0.88	0.78	0.87
New Mexico	1.00	1.00	1.00	1.00	1.00
New York	0.63	0.65	0.63	0.60	0.62
North Dakota	1.00	0.89	1.00	0.87	1.00
Ohio	0.77	0.80	0.77	0.67	0.74
Oklahoma	1.00	1.00	1.00	1.00	0.72
Rhode Island	0.73	0.76	0.72	0.69	0.72
South Carolina	1.00	1.00	1.00	1.00	1.00
South Dakota	0.80	0.80	0.81	0.72	0.76
Tennessee	1.00	1.00	0.77	0.77	0.84
Texas	1.00	1.00	1.00	1.00	0.80
Utah	1.00	1.00	1.00	1.00	1.00
Vermont	0.81	0.77	0.81	0.79	0.81
Virginia	1.00	0.84	1 .00	1.00	0.79
Washington	0.83	0.83	0.84	0.76	0.69
West Virginia	1.00	1.00	1.00	1.00	1.00
Wisconsin	0.95	0.90	0.83	0.77	0.88
Average	0.884	0.872	0.859	0.829	0.813
N. of frontier DMUs	21	17	16	15	10