Evidence Building and Information Accumulation: Using the Bayesian Paradigm to Advance Child Welfare Intervention Research

Ding-Geng Chen *University of North Carolina at Chapel Hill*
Mark F. Testa *University of North Carolina at Chapel Hill*
David Ansong *University of North Carolina at Chapel Hill*
Kanisha C. Brevard *University of North Carolina at Chapel Hill*

ABSTRACT ***Objective*:** Intervention evaluation typically follows a frequentist paradigm: New analyses are conducted for each subsequent study, and findings are then used to improve policy practice. This approach largely ignores data from prior studies, leading to information loss and incomplete or inaccurate conclusions. Unlike the frequentist paradigm, the Bayesian paradigm uses formative data (as priors), which can be updated with the summative data, thus building on existing evidence about an intervention's effectiveness. ***Method*:** This article uses data from the Safe Families for Children randomized controlled trial to illustrate how the Bayesian paradigm incorporates prior evidence at the formative phase with data at the summative phase to provide a more comprehensive analysis. This approach is consistent with the scientific principle of evidence building. We compare the merits of each paradigm on two evaluation criteria: (a) p-values from a chi-square test, and (b) the probability that the intervention is superior to the comparison group on three outcome variables (protective custody, deflection from foster care, and whether repeat victimization occurred). ***Results*:** The Bayesian paradigm consistently outperformed the frequentist paradigm. ***Conclusion*:** The Bayesian paradigm is superior to the frequentist paradigm in demonstrating the effectiveness of an intervention, as evidenced by smaller p-values and a higher probability that the intervention group outperformed the comparison group.

KEYWORDS: intervention research, Bayesian, prior distribution, posterior distribution, posterior probability of treatment effects

doi: 10.1086/711376

Child welfare is facing a credibility deficit. The dearth of replicable evidence for many programs that purport to improve child and family outcomes is eroding public confidence in the child welfare system's capacity to deliver on its promises of child safety, family permanence, and adolescent well-being (Epstein & Klerman, 2012; Testa, 2018). As of March 2019, only 30 of the 456 programs (7%) cataloged by the California Evidence-Based Clearinghouse for Child Welfare (2019) were rated as well supported by research evidence. To receive this rating, a program must have been evaluated in at least two randomized controlled trials (RCTs) conducted in different practice settings, and RCT results must demonstrate that the program is superior to an appropriate comparison program. The number of supported programs increases to 79 (17%) if the reproducibility criterion is dropped and superiority is demonstrated in only one RCT. Still, only 20 of these "evidence-lite" programs are specifically designed or commonly used for children and families served by the child welfare system due to insufficient evidence for other programs.

In 2018, the U.S. Congress codified similar evidence standards for funding child welfare preventive services under the Family First Prevention Services Act (Children's Defense Fund, 2018). Because the Act prioritizes funding for services and programs that meet certain evidence-based standards (i.e., promising, supported, or well-supported evidence), a pressing need exists for child welfare practitioners and researchers to conduct routine, low-cost, rigorous evaluations to expand the supply of evidence-supported child welfare programs in usual practice settings (Testa, 2018). New phase-based approaches to evidence building can readily incorporate recent methodological advances into existing continuous quality improvement efforts (State of Illinois, 2016; Testa & Poertner, 2010). These advances include (a) automation of unbiased assignment mechanisms—such as randomization, alternation, and rotational assignment—as part of routine service delivery (Higgins & Green, 2011; Testa, 2010b, 2018); (b) tracking outcomes using existing administrative data in lieu of more expensive primary data collection (Kum et al., 2015; Permanency Innovations Initiative Evaluation Team, 2016); and (c) application of the Bayesian paradigm as a supplement to conventional frequentist approaches—which are fully based on the probability theory—to testing validity of statistical conclusions (Barboza, 2019; Chen & Ansong, 2019; Chen & Fraser, 2017a, 2017b; Chen et al., 2018; Freisthler & Weiss, 2008; Kaplan, 2014; Williams et al., 2015). The integration of these methods into phase-based approaches to evidence building holds great promise for policy work given the increased ability to generate low-cost, replicable results. Knowing which programs work for whom and under which conditions is critically important with appropriate application of methods following a commitment to monitoring and evaluating programs used in child welfare.

In this article, we focus on the third methodological advancement—the application of the Bayesian paradigm—to demonstrate how incorporating prior evidence with current data is a more comprehensive analytic approach that closely aligns with the scientific principle of phase-based evidence building and information accumulation in intervention research. Using examples from a low-cost RCT called

Safe Families for Children, we provide insight into the Bayesian paradigm's ability to build on or strengthen conclusions about the efficacy of a promising intervention for a targeted population.

Evidence-Building and Information Accumulation from a Bayesian Perspective

Phase-Based Evidence Building

One of the contributors to the credibility deficit in child welfare, and across the social sciences in general (Maxwell, 2004), is a tendency to rush to publish selective findings that arise by chance in underpowered studies but fail to demonstrate statistical or practical significance at replication. To remedy this problem, scholars in all disciplines have called for researchers to adhere to a process of evidence building that progresses through successive phases of increasingly generalizable studies (Shadish et al., 2002; Testa, 2010a). As outlined in the *Framework to Design, Test, Spread, and Sustain Effective Practice in Child Welfare* (Framework Workgroup, 2014, p. 7), rigorous evidence building occurs in five phases: identify and explore, develop and test, compare and learn, replicate and adapt, and apply and improve. The second phase, develop and test, is most relevant to addressing the problem of irreproducible findings. In this phase, researchers first confirm program usability and then conduct formative evaluation to assess whether program outputs and primary short-term outcomes are statistically trending in the desired direction before proceeding to the next phase, compare and learn. In the compare-and-learn phase, researchers conduct a summative evaluation to assess whether the program creates practical improvements in primary (preferably preregistered) long-term outcomes that can plausibly be attributed to the causal effects of the intervention.

As originally formulated by Scriven (1991), a formative evaluation is conducted with the intent of identifying weaknesses in the logic and early implementation of a program. Scriven suggested the most useful formative evaluation is "early warning summative," which pilots the same unbiased allocation mechanism that will be used at the later summative evaluation stage. The value of early warning evaluations in child welfare reflects the reality that when rigorously evaluated, only a fraction of promising innovations turn out to be effective or even marginally successful in accomplishing their set aims (Rossi, 1978). As a benchmark, the Laura and John Arnold Foundation (2018) and Manzi (2016) have separately reported data showing that most RCTs conducted in business, education, criminology, political science, and economics that progress through all five phases—from promising innovation to reproducible successes—seldom exceed one improvement for every four unsuccessful attempts. Given these odds, the earlier in the evidence-building process that a warning bell can be sounded, the better. Early warning allows corrections to be made before too much time and effort are wasted on the implementation and evaluation of innovations that are unlikely to show improvements when evaluated using adequately powered samples.

For the handful of interventions that complete the recommended phases, a key analytic question worth addressing is whether Scriven's (1991) advice about setting aside the findings from formative evaluation is the best way to proceed. Is there an alternative or optimal way to statistically evaluate promising innovations that is more in keeping with the phase-based evidence-building principle? Given the push for phase-based evidence building, the principle that evidence building ought to pass through successive "tollgates" of implementation and evaluation is gaining ground in the conceptualization, design, and rollout of new programs. However, the analytic approaches that many evaluations use lag behind the movement toward a phase-based approach.

Many child welfare research initiatives would benefit by shifting from a frequentist paradigm, which separately analyzes formative and summative data, to a Bayesian paradigm in which data from the formative evaluation phase are used as prior evidence that can be updated with the final summative evaluation data, thus building on and adding cumulatively to the evidence about an intervention's effectiveness. Next, we offer a brief review of the theoretical basis, foundational principles, and merits of the Bayesian approach to statistical modeling with the goal of calling attention to its compelling consistency with the phase-based mindset to evidence building.

Evidence Building Through the Lens of Bayes' Theorem

Intervention research and program evaluation studies typically build on prior evidence and then collect new data to update and improve this prior evidence. Over time, this sequential process accumulates compelling evidence that should improve evidence-based decision-making. This evidence-building process can be translated into Bayes' theorem, as represented in the following equation:

$$P(A \mid B) = P(A)P(B \mid A)/P(B). \quad (1)$$

A denotes the prior evidence (e.g., data from the formative phase) with probability $P(A) > 0$, and B denotes the subsequently collected evidence (e.g., data from the summative phase) with probability $P(B) > 0$. Based on the probability theory principle of conditional probability, $P(AB) = P(A)P(B|A)$ as well as $P(AB) = P(B)P(A|B)$. Therefore, $P(A)P(B|A) = P(B)P(A|B)$ because they are all equal to $P(AB)$, which leads to $P(A|B) = P(A)P(B|A)/P(B)$, as seen in Equation 1.

The simple, elegant idea represented in Equation 1 exemplifies Bayes' theorem, which is the updated probability distribution of prior evidence (A) with newly collected evidence (B). Thus, $P(A|B)$ is proportional to the product of the prior probability distribution, $P(A)$, and a conditional probability distribution for new evidence of B given the prior evidence of A or $P(B|A)$. The equation also includes a proportional constant, $1/P(B)$. Far from a new development, Bayes' theorem was developed in the 18th century by Thomas Bayes (Bayes & Price, 1763).

From Bayes' Theorem to Bayesian Modeling

Application of Bayes' theorem to intervention research and evaluation studies offers an intuitive framework for Bayesian modeling. To illustrate this framework, let B represent the observed new data (D) from the summative phase; A represents the hypothetical intervention effect parameter (θ) estimated from the formative phase. In this case, the probability $P(B|A)$ is the *data likelihood function*, $L(\theta) = L(D|\theta)$, and $P(A) = P(\theta)$ is the prior distribution of intervention effect parameter θ at the formative phase. In this setting, Bayes' theorem becomes

$$P(\theta \mid D) = P(\theta)P(D \mid \theta)/P(D), \tag{2}$$

where θ is the intervention effect parameter to be estimated with the current data. $P(\theta)$ is the prior probability distribution of intervention parameter θ before D is observed, which is the prior belief about how likely different parameters are based on the prior evidence (e.g., in the formative phase). $P(D|\theta)$ is the probability of observing D given intervention parameter θ, which is also known as the data likelihood used in classical frequentist statistical modeling. The constant, $P(D)$, is the marginal likelihood of the integration of $P(D|\theta)$, which is the probability of new evidence. Moreover, $P(\theta|D)$ becomes the updated posterior probability distribution of intervention parameter θ after D is observed, which is the probability of intervention effect θ given the observed data D with the incorporation of prior evidence. This *posterior probability function* is central in Bayesian modeling and is used for estimation of the intervention effect after data are collected.

Therefore, the fundamental advantage of a Bayesian paradigm is that it provides a logically cohesive framework to update prior information with new data through the application of Bayes' theorem. Bayesian modeling can be applied iteratively. In other words, after observing some data, the resulting posterior probability distribution can be treated as the next prior probability distribution, and a new posterior probability can be derived from the next new data. This iterative process allows for Bayesian principles to be applied to various kinds of data, whether viewed all at once or over time. The statistical literature on Bayesian modeling is vast (e.g., see Berger, 1985; Carlin & Louis, 2008; Chen et al., 2017; Gelman et al., 2013) and is increasingly being recognized in social work research focused on health intervention research (Chen & Ansong, 2019; Chen & Fraser, 2017a, 2017b; Chen et al., 2018; Kaplan, 2014).

Evaluation Criteria for Assessing Intervention Effectiveness

Alpha Cutoff Criterion With a One-Tailed Test

Scholars who apply Bayesian methods in intervention research have a variety of options for evaluating intervention effectiveness, one of which is the alpha cutoff with an option of using a one-tailed test. Through the null hypothesis statistical testing approach, the use of a chi-square test with a predetermined alpha cutoff allows for a binary (yes/no) confirmation of whether the data support a hypothesis that the intervention is superior to the control. More important, because of insights from theory or prior studies, the ability to formulate a directional hypothesis of intervention superiority allows for the use of a one-tailed statistical test, which leads to more statistical power to detect effects. Because an intervention study is typically designed to determine the effectiveness of an intervention (i.e., treatment, denoted by T) compared with services-as-usual (i.e., control, denoted by C), a directional alternative hypothesis that the intervention (T) is better than the control (C)—$H_a : T > C$—and a one-tailed test are most appropriate.

To illustrate further, the parameter to evaluate the effectiveness of the intervention (θ) is typically denoted by the difference in the probabilities of how participants in the intervention arm (p^T) and the control arm (p^C) responded after treatment delivery, which can be written as $\theta = p^T - p^C$. These probabilities can be estimated from the beta distributions as seen in Equations 3–5. The statistical chi-square test and a predetermined alpha cutoff can be used to assess the significance of intervention effectiveness (Chen et al., 2017, p. 39).

The Probability Criterion

In addition to the simple yes/no confirmation of the superiority of an intervention based on p-values, the Bayesian approach offers a way to assess the extent to which an intervention is superior to the control. In other words, it is possible to determine the probability that the new intervention is better than the control to a certain degree—that is, $P(T > C)$. In this paper, we introduce two approaches to calculating the probability of $P(T > C)$. The first approach is to directly calculate this probability using the theory of probability distributions for the beta distribution in Equations 3–5 (Chen et al., 2017, p. 268):

$$\text{Prior at formative phase:} \quad P(p) = \text{Beta}(y_1 + 1, n_1 - y_1 + 1) \tag{3}$$

$$\text{New data at summative phase:} \quad P(D|p) = \text{Beta}(y_2 + 1, n_2 - y_2 + 1) \tag{4}$$

$$\text{Bayesian posterior:} \quad P(p \mid D) = \text{Beta}(y_1 + y_2 + 2, n_1 + n_2 - y_1 - y_2 + 2) \tag{5}$$

In Equation 3, n_1 represents participants enrolled in the study at the formative phase who were assigned to a specific intervention arm. Among the n_1 participants, y_1 represents those who responded to the intervention and achieved the outcome. Therefore, the binary outcome, y_1, is binomially distributed as $y_1 \sim \text{Bin}(n_1, p)$, where p is the binomial proportion parameter representing the proportion of participants who responded to the intervention arm. This binomial distribution can be reformulated as the beta distribution for Bayesian inference (Chen et al., 2017, p. 245), which is the prior distribution shown in Equation 3. Similarly, in Equation 4, n_2 represents participants enrolled in the study at the summative phase who were assigned to the same specific intervention arm, and y_2 represents those who responded to the intervention and achieved the outcome. Per Bayes' theory, a posterior distribution can be derived as in Equation 5. Therefore, to calculate the $P(T > C)$ directly, let us denote $T \sim$ beta(s_T, t_T) and $C \sim$ beta (s_C, t_T), where s_T, t_T and s_C, t_T are the beta distribution parameters in Equations 3–5. With some mathematical manipulations, we can show:

$$
\begin{aligned}
P(T > C) &= \int_0^1 \frac{x^{s_T-1}(1-x)^{t_T-1}}{B(s_T, t_T)} \left(\int_0^x \frac{y^{s_C-1}(1-y)^{t_C-1}}{B(s_C, t_C)} dy \right) dx \\
&= \frac{B(s_T + s_C, t_T + t_T)}{B(s_T, t_T) B(s_C, t_C) s_T} {}_3F_2 \left(\begin{matrix} s_T + s_C, s_T + t_C, 1 \\ s_T + 1, s_T + t_T + s_C + t_C \end{matrix} \right),
\end{aligned}
\tag{6}
$$

where $B(s, t)$ is the usual beta function and ${}_3F_2$ is the hypergeometric function with upper parameters $(s_T + s_C, s_T + t_C, 1)$ and lower parameters $(s_T + 1, s_T + t_T + s_C + t_C)$. This calculation can be done using the R library *hypergeo* (Chen et al., 2017).

The second approach uses Monte Carlo simulation-based calculations (MCSC; Chen & Chen, 2017), which were also used by Chen and Fraser (2017b) to demonstrate how the Bayesian approach can be used to inform decisions about the adequacy of sample size and whether to continue program development in intervention research. MCSCs are easy to implement with computer simulation techniques and can be a remarkable alternative to the first approach, especially when the integration in Equation 6 does not converge numerically. The implementation of the MCSC approach is a three-step process:

Step 1: Simulate the beta distributions in Equations 3–5 for the treatment arm as $T \sim$ Beta(s_T, t_T) and for the control arm as $C \sim$ Beta(s_C, t_C).

Step 2: Evaluate whether $T > C$ from Step 1. If true, then record 1, otherwise 0.

Step 3: Repeat Steps 1 and 2 many times (typically, $B = 1{,}000{,}000$ times) and count the proportion of $T > C$ from Step 2 among these B simulations to obtain the MCSC estimate of $P(T > C)$.

Application to Child Welfare Data

The phase-based approach to evidence building was used in the development and testing of the Safe Families for Children (SFFC) project, a low-cost RCT. In the SFFC project, the phase-based approach informed the conceptualization, design, and implementation of the intervention, thus producing formative and summative data, which are best suited for Bayesian statistical modeling. In this section, we provide an example of how Bayesian statistical modeling can be applied to real-world data to add value to child welfare evidence building.

Description of the Safe Families for Children Intervention

SFFC is a program in Illinois developed to prevent the recurrence of child maltreatment and the removal of children into state custody. Children are eligible for this program if they have been formally investigated by child protective services (CPS) and voluntarily placed by their parents with host families during a time of need, such as during a CPS investigation. LYDIA Home Association, a nonprofit Christian organization in Illinois, helps match a host family with a child while CPS completes its investigation. The arrangement that SFFC provides is temporary, typically lasts less than 2 months, and the parents retain legal custody of their children (as they do with voluntary kinship care). One of the many benefits of the SFFC program is that it helps reduce the amount of government intrusion in the lives of children while maintaining child safety.

The SFFC program was evaluated to determine whether children referred to SFFC's host families are less likely to enter state protective custody, more likely to be deflected from foster care, and less likely to experience subsequent maltreatment within 60 days of a prior report as compared with children from similar families who received child protective services-as-usual (SAU). The formative phase implemented the early warning summative design recommended by Scriven (1991) and included an unbiased allocation mechanism that alternately assigned families to SFFC or services-as-usual. Families assigned to the intervention were given the option to participate in the treatment or receive services-as-usual with those assigned to the comparison condition. This type of design is referred to as a randomized encouragement design, (Behaghel et al., 2013; Holland, 1988; West et al., 2008). Families preassigned to the intervention group were exposed to all of the usual services plus the choice of participating in a SFFC host family arrangement if appropriate.

Bayesian Modeling of Safe Families for Children Data

The SFFC study focused on whether protective custody, deflection from foster care, and repeat victimization occurred; each binary outcome was coded 1 (outcome occurred) or 0 (outcome did not occur).The variables were defined as follows: "Children not taken into protective custody 2 or more days after assignment" (i.e., *no protective custody* variable), "Deflection of children from foster care 2 or more days after assignment" (i.e., *deflection from foster care* variable), and "No repeat victimization within 2 months of report at assignment" (i.e., *no repeat victimization* variable).

Data Modeling Process

We used a three-step sequential Bayesian modeling process to analyze the SFFC data for the intervention arm (T) and the services-as-usual control arm (C). For simplicity, we describe the three-step process for the intervention arm; the same process can be applied to the service-as-usual control arm.

Step 1. For the SFFC intervention arm (T) at the formative phase, we specified the following prior distribution:

$$P(p_T) = \text{Beta}(y_{1T} + 1, n_{1T} - y_{1T} + 1). \tag{7}$$

In Equation 7, n_{1T} represents participants enrolled in the SFFC study who received the intervention arm. Among the n_{1T} participants, y_{1T} participants achieved and responded to the intervention arm. Therefore, the binary outcome, y_{1T}, is binomially distributed as $y_{1T} \sim \text{Bin}(n_{1T}, p_T)$, where p_T is the binomial proportion parameter representing the proportion of participants who responded to the intervention arm.

Step 2. For new data at the summative phase of the SFFC study, we specified a data likelihood function as follows:

$$P(D \mid p_T) = \text{Beta}(y_{2T} + 1, n_{2T} - y_{2T} + 1). \tag{8}$$

In Equation 8, n_{2T} represents participants enrolled in the intervention arm—that is, SFFC(T)—and y_{2T} represents participants who achieved the intended intervention outcome. Thus, y_{2T} is binomially distributed as $y_{2T} \sim \text{Bin}(n_{2T}, p_T)$, which can be reformulated as the beta distribution for Bayesian inference.

Step 3. The Bayesian paradigm systematically incorporates both prior data (i.e., at the formative phase) in Equation 3 with new data (i.e., at the summative phase) in Equation 4. Here, the new SFFC data is used to update the prior evidence to estimate the intervention parameter (p_T). Chen and colleagues (2017, p. 245) have shown that the posterior distribution is still a conjugate beta distribution. This Bayesian posterior distribution is denoted by

$$P(p_T \mid D) = \text{Beta}(y_{1T} + y_{2T} + 2, n_{1T} + n_{2T} - y_{1T} - y_{2T} + 2). \tag{9}$$

Probability Analysis

We also used the probability criteria and method described in the background section to examine the probability that the intervention was superior to the control regarding rates of reunification, stable permanency, and stable reunification. All calculations were implemented in R software, which can be requested from the authors.

Monte Carlo Calculation

In addition to the χ^2 test with an alpha cutoff, we used the Monte Carlo simulation-based calculation to assess the extent to which the intervention is superior to control $P(T > C)$ on all three binary outcomes: protective custody, deflection from foster care, and repeat victimization.

Results

Results from the SFFC data are presented in Table 1. It can be seen from Table 1 that the observed proportions in the formative phase are 0.974, 0.819, and 1.000 for SFFC, and 0.843, 0.622, and 0.941 for SAU. These proportions would result in the intervention effect of 0.131 for *no protective custody*, 0.197 for *deflection from foster care*, and 0.059 for *no repeat victimization* at the formative phase. Similarly, the observed proportions in the summative phase are 0.937, 0.797, and 0.971 for SFFC, and 0.970, 0.612, and 0.966 for SAU. These would result in the summative phase intervention effect of −0.033, 0.186, and 0.005, respectively.

As shown in the lower part of Table 1, the application of a chi-square test to assess the effectiveness of the SFFC at the summative phase using the frequentist paradigm resulted in p-values of 0.710 for *no protective custody*, 0.011 for *deflection from foster care*, and 0.500 for *no repeat victimization*. Here, the p-values for *no protective custody* and *no repeat victimization* were not statistically significant, suggesting no intervention effect at the .05 significance level. However, when we applied the Bayesian paradigm, the p-values for all three outcomes were reduced (i.e., the statistical significance of the intervention effect improved). As shown in the lower part of Table 1, the p-values changed from a nonsignificant value of .710 to a significant value of .006 for the *no protective custody* outcome. Similarly, in the *deflection from foster care* model, the already significant p-value of .011 improved to a highly significant p-value of $< .001$. Even in the *no repeat victimization* model where both the frequentist and Bayesian model produced nonsignificant p-values, the Bayesian approached reduced the p-value by 63% (i.e., from .500 to .186). Essentially, by incorporating summative data to update the formative data, the Bayesian paradigm gives smaller p-values for all three outcomes.

Because the direct calculation using Equation 6 did not converge, we used the MCSC to evaluate the intervention effectiveness with the criterion of $P(T > C)$. Based on the MCSC, the Bayesian paradigm improved the calculated probabilities for the *no protective custody* outcome by 405% (i.e., from 0.197 to 0.995), by 1% for the *deflection from foster care* outcome (i.e., from 0.993 to 0.999), and by 53% for the *no repeat victimization* outcome (i.e., from 0.559 to 0.858). The improvements driven by the Bayesian paradigm are substantial (53% to 405%), except for the *deflection from foster care* outcome (1%), which already had a high probability value and, therefore, could not go any higher than .999 (surely a favorable intervention effect).

Results from Table 1 numerically illustrate the superiority of the Bayesian paradigm in evidence-based intervention research. The Bayesian approach can provide statistically more powerful analysis with smaller and more reliable p-values for intervention effectiveness. The Bayesian paradigm is a logical methodological approach that allows for formative information to be updated with summative data to provide a more comprehensive analysis, which is consistent with the scientific principle of evidence building. Consequently, the Bayesian paradigm consistently outperformed the frequentist paradigm.

Table 1

Results From Chi-Square Test and the Probability That the SFFC Intervention is Superior to the Control

		Summary of SFFC Data		
		Outcome Variable: Child Level		Outcome Variable: Household Level
Phase	Treatment	No Protective Custody	Deflection From Foster Care	No Repeat Victimization
Formative	SFFC(T)	$y_{1T}/n_{1T} = 113/116 = 0.974$	$y_{1T}/n_{1T} = 95/116 = 0.819$	$y_{1T}/n_{1T} = 45/45 = 1.000$
	SAU(C)	$y_{1C}/n_{1C} = 107/127 = 0.843$	$y_{1C}/n_{1C} = 79/127 = 0.622$	$y_{1C}/n_{1C} = 48/51 = 0.941$
Summative	SFFC(T)	$y_{2T}/n_{2T} = 74/79 = 0.937$	$y_{2T}/n_{2T} = 63/79 = 0.797$	$y_{2T}/n_{2T} = 33/34 = 0.971$
	SAU(C)	$y_{2C}/n_{2C} = 65/67 = 0.970$	$y_{2C}/n_{2C} = 41/67 = 0.612$	$y_{2C}/n_{2C} = 28/29 = 0.966$
		Evaluation Criteria: Intervention is More Effective than the Control		
Criteria	Method	p-value	p-value	p-value
Chi-square test criterion	Frequentist	0.710	0.011	0.500
	Bayesian	0.006	< 0.001	0.186
		$P(T > C)$	$P(T > C)$	$P(T > C)$
Probability criterion	Frequentist	0.197	0.993	0.559
	Bayesian	0.995	0.999	0.858

Note. SFFC = Safe Families for Children; SAU = services as usual. The rows corresponding to "SFFC" and "SAU" list the observed data as defined by the number of children who achieved the outcome (y), the total number of children (n), and the associated proportions for all three outcomes from the formative to the summative phases. y_{1T} and n_{1T} denote y_1 and n_1 values for SFFC(T), and y_{1C} and n_{1C} denote y_1 and n_1 values for SAU(C) at the formative phase. y_{2T} and n_{2T} denote y_2 and n_2 values for SFFC(T), and y_{2C} and n_{2C} denote y_2 and n_2 values for SAU(C) at the summative phase.

Discussion

In this paper, we have demonstrated the utility and merits of the Bayesian paradigm for decision-making and policymaking. Bayesian methods incorporate prior evidence with current data for a more comprehensive analysis, which aligns with the scientific principle of evidence building and information accumulation in intervention research. We used an example from an RCT to show that when evaluating interventions, Bayesian methods can produce smaller *p*-values and higher probabilities than frequentist methods to help inform decisions about the efficacy of an intervention. These two advantages reduce the level of uncertainty in results.

The Bayesian methodological advancement has practical advantages that may be useful for developing and building on child welfare interventions. Bayesian approaches could foster efficiency in child welfare research, allowing researchers to go beyond merely citing existing data in the literature review and put prior data to good use in statistical modeling. The emerging phase-based approach to child welfare research is an essential step toward advancing evidence-based research in the field. However, such efforts ought to be holistic enough to go beyond the conceptualization, design, and implementation of phase-based intervention research. Thus, it is equally essential that statistical approaches applied to data from different phases of interventions match the sequential nature of phase-based evidence building. Researchers who collect multiple waves of data, such as formative and summative data, can readily incorporate methodological advances such as the Bayesian statistical approaches discussed in this paper. Too often, researchers who rely on the classical frequentist paradigm to test treatment effectiveness fail to take advantage of multiple waves of data available from both the formative and summative phases of their research. The downside to this conventional approach is that statistical conclusions about intervention effectiveness are largely generated in a vacuum because the approach often ignores related information from prior studies. This conventional frequentist approach is in sharp contrast to real-world decision-making and policymaking processes in child welfare and other areas that are generally cumulative and contingent on how well previous interventions, programs, and policies faired.

Considering information from pilot and prior studies offers statistical benefits, including higher statistical power and improved efficiency. Because the Bayesian paradigm relies on prior knowledge to evaluate intervention effectiveness, the ability to consistently use a one-tailed test is an added advantage over the frequentist paradigm. One-tailed tests require less statistical power to detect effects (if effects exist). By acknowledging available information from prior studies, researchers become more certain about the direction of the intervention effect, thus warranting a directional hypothesis and the correspondent one-tailed test. Indeed, Bayesian statistical modeling uses a similar approach as intervention researchers, who often start with some foreknowledge about the direction of intervention effects based either on published studies or pilot and formative research. Ultimately, using one-tailed tests in phase-based studies that use Bayesian statistical modeling would increase the statistical power of these studies to detect an effect, if one exists. A related advantage is that Bayesian approaches might be more efficient and cost-effective because child welfare researchers do not have to render small or old data "unusable"; rather, such data sets can be combined with either prior or future data.

The phase-based approach to evidence-based policymaking in child welfare aspires to the rigorous evidence standards that medical research is sometimes able to attain. More often, practical realities in child welfare research require making pragmatic trade-offs that can weaken the generalized validity of empirical findings in the human services. The inability to assume uniform behavioral responses, the impossibility of mounting double-blinded experiments in most cases, and the reactivity of program effects to specific contexts require the use of alternative approaches to conventional designs and analytic methods in medical research. This paper presents one of these alternative analytic methods of phase-based evidence building: Bayesian statistical modeling. The utility of the Bayesian paradigm lies in its ability to add precision to summative findings using formative results. Long term, we expect that more child welfare intervention researchers will consider the Bayesian approach as well as other alternative methods, including preconsent, unbiased allocation designs, and outcomes tracking with administrative data. Future studies should continue to explore and demonstrate the utility of these methods. The Bayesian approach and emerging methodological advancements will gradually shape the future direction of social and health intervention studies. As more researchers become familiar with Bayesian methods through methodological training and demonstration papers such as this, we are hopeful that more child welfare programs will meet the rigorous evidence standards required under the Family First Prevention Services Act (Testa 2010a, 2010b; Testa & Poertner, 2010).

Author Notes

Ding-Geng Chen, PhD, is the Wallace H. Kuralt Distinguished Professor in the University of North Carolina at Chapel Hill School of Social Work and professor of biostatistics, Department of Biostatistics, Gillings School of Global Public Health, University of North Carolina at Chapel Hill.

Mark F. Testa, PhD, is the Sandra Reeves Spears and John B. Turner Distinguished Professor in the University of North Carolina at Chapel Hill School of Social Work.

David Ansong, PhD, is an associate professor in the University of North Carolina at Chapel Hill School of Social Work.

Kanisha C. Brevard, PhD, is a research associate in the University of North Carolina at Chapel Hill School of Social Work.

Correspondence regarding this article should be directed to Ding-Geng Chen, 325 Pittsboro St., Campus Box 3550, Chapel Hill, NC 27599 or via e-mail to dinchen@email.unc.edu.

Acknowledgments

We would like to thank the three anonymous reviewers and former *JSSWR* Editor-in-Chief Jeffrey M. Jenson for their comments and suggestions, which substantially improved the quality of this paper. Our thanks also go to Ms. Diane Wyant and Chelsey Baker-Hauck for their professional editing of this manuscript.

References

Barboza, G. E. (2019). The geography of child maltreatment: A spatiotemporal analysis using Bayesian hierarchical analysis with integrated nested Laplace approximation. *Journal of Interpersonal Violence*, *34*(1), 50–80. https://doi.org/10.1177/0886260516639583

Bayes, T., & Price, R. (1763). An essay towards solving a problem in the doctrine of chance. By the late Rev. Mr. Bayes, communicated by Mr. Price, in a letter to John Canton, A. M. F. R. S. *Philosophical Transactions of the Royal Society of London*, *53*, 370–418. https://doi.org/10.1098/rstl.1763.0053

Behaghel, L., Crépon, B., & Gurgand, M. (2013). *Robustness of the encouragement design in a two-treatment randomized control trial* (IZA DP No. 7447). Institute for the Study of Labor.

Berger, J. O. (1985). *Statistical decision theory and Bayesian analysis* (2nd ed.). Springer-Verlag Press. https://doi.org/10.1007/978-1-4757-4286-2

California Evidence-Based Clearinghouse for Child Welfare. (2019). [Database]. http://www.cebc4cw.org

Carlin, B. P., & Louis, T. A. (2008). *Bayesian methods for data analysis* (3rd ed.). CRC Press.

Chen, D. G., & Ansong, D. (2019). Bayesian spatial-temporal modeling of space and time dynamics: A practical demonstration in social and health research. *Journal of the Society for Social Work and Research*, *10*(2), 275–299. https://doi.org/10.1086/703444

Chen, D. G., & Chen, J. D. (2017). *Monte-Carlo simulation-based statistical modeling*. Springer/ICSA Book Series in Statistics. https://doi.org/10.1007/978-981-10-3307-0

Chen, D. G., & Fraser, M. (2017a). A Bayesian approach to sample size estimation and the decision to continue program development in intervention research. *Journal of the Society for Social Work and Research*, *8*(3), 457–470. https://doi.org/10.1086/693433

Chen, D. G., & Fraser, M. (2017b). A Bayesian perspective on intervention research: Using prior information in the sequential development of social and health programs. *Journal of the Society for Social Work and Research*, *8*(3), 441–456. https://doi.org/10.1086/693432

Chen, D. G., Fraser, M., & Cuddeback, G. (2018). Assurance in intervention research: A Bayesian perspective on statistical power. *Journal of the Society for Social Work and Research*, *9*(1), 158–173. https://doi.org/10.1086/696239

Chen, D. G., Peace, K. E., & Zhang, P. (2017). *Clinical trial data analysis using R and SAS*. CRC Press.

Children's Defense Fund. (2018). *The Family First Prevention Services Act: Historic reforms to the child welfare system will improve outcomes for vulnerable children*. https://www.childrensdefense.org/wp-content/uploads/2018/08/family-first-detailed-summary.pdf

Epstein, D., & Klerman, J. A. (2012). When is a program ready for rigorous impact evaluation? The role of a falsifiable logic model. *Evaluation Review*, *36*(5), 375–401. https://doi.org/10.1177/0193841X12474275

Framework Workgroup. (2014). *A framework to design, test, spread, and sustain effective practice in child welfare*. Children's Bureau, Administration for Children and Families, U.S. Department of Health and Human Services. https://www.acf.hhs.gov/sites/default/files/cb/pii_ttap_framework.pdf

Freisthler, B., & Weiss, R. E. (2008). Using Bayesian space-time models to understand the substance use environment and risk for being referred to child protective services. *Substance Use & Misuse*, *43*(2), 239–251. https://doi.org/10.1080/10826080701690649

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis* (3rd ed.). CRC Press.

Higgins, J. P. T., & Green, S. (2011). *Cochrane handbook for systematic reviews of interventions version 5.1.0*. Cochrane Collaboration. https://handbook-5-1.cochrane.org/index.htm#part_2_general_methods_for_cochrane_reviews.htm

Holland, P. W. (1988). Causal inference, path analysis, and recursive structural equations models. *Sociological Methodology*, *18*, 449–484. https://doi.org/10.1002/j.2330-8516.1988.tb00270.x

Kaplan, D. (2014). *Bayesian statistics for the social sciences*. Guilford Press.

Kum, H. C., Stewart, C. J., Rose, R. A., & Duncan, D. F. (2015). Using big data for evidence based governance in child welfare. *Children and Youth Services Review*, *58*, 127–136. https://doi.org/10.1016/j.childyouth.2015.09.014

Laura and John Arnold Foundation. (2018, March 1). *How to solve U.S. social problems when most rigorous program evaluations find disappointing effects (Part One in a series)*. https://www.straighttalkonevidence.org/2018/03/21/how-to-solve-u-s-social-problems-when-most-rigorous-program-evaluations-find-disappointing-effects-part-one-in-a-series/

Manzi, J. (2016, June 2). *Rapid cycle testing: Lessons from the private sector*. Presentation at the 2016 Research and Evaluation Conference on Self-Sufficiency, Washington, DC. Retrieved from https://youtu.be/VbVrvHp5xTI

Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: Causes, consequences, and remedies. *Psychological Methods*, *9*(2), 147–163. https://doi.org/10.1037/1082-989X.9.2.147

Permanency Innovations Initiative Evaluation Team. (2016). *Using child welfare administrative data in the Permanency Innovations Initiative evaluation* (OPRE Report 2016-47). U.S. Department of Health and Human Services, Administration for Children and Families, Children's Bureau, and Office of Planning, Research and Evaluation. https://www.acf.hhs.gov/sites/default/files/opre/using_child_welfare_administrative_data_in_pii_compliant.pdf

Rossi, P. H. (1978). Issues in the evaluation of human services delivery. *Evaluation Quarterly*, *2*(4), 573–599. https://doi.org/10.1177/0193841X7800200404

Scriven, M. (1991). *Evaluation thesaurus* (4th ed.). SAGE Publications.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton, Mifflin and Company.

State of Illinois. (2016). *Illinois Birth through Three Waiver: Child and family intervention*. https://www2.illinois.gov/dcfs/aboutus/newsandreports/Documents/IL_IB3_Interim_Evaluation_Report_Mar_2016_Executive_Summary_6-26-17.pdf

Testa, M. F. (2010a). Introduction. In M. F. Testa & J. Poertner (Eds.), *Fostering accountability: Using evidence to guide and improve child welfare policy* (pp. 3–34). Oxford University Press.

Testa, M. F. (2010b). Logic of results-oriented accountability: Populations, interventions, comparisons, and outcomes. In M. F. Testa & J. Poertner (Eds.), *Fostering accountability: Using evidence to guide and improve child welfare policy* (pp. 75–100). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195321302.003.0003

Testa, M. F. (2018, October 31). Child welfare's looming evidence problem. *Chronicle of Social Change*. Retrieved from https://chronicleofsocialchange.org/analysis/child-welfares-looming-evidence-problem/32586

Testa, M. F., & Poertner, J. (2010). *Fostering accountability: Using evidence to guide and improve child welfare policy*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195321302.001.0001

West, S. G., Duan, N., Pequegnat, W., Gaist, P., Des Jarlais, D. C., Holtgrave, D., Szapocznik, J., Fishbein, M., Rapkin, B., Clatts, M., & Mullen, P. D. (2008). Alternatives to the randomized controlled trial. *American Journal of Public Health*, *98*(8), 1359–1366. https://doi.org/10.2105/AJPH.2007.124446

Williams, J. R., Merkel-Holguin, L., Allan, H., Maher, E. J., Fluke, J., & Hollinshead, D. (2015). Factors associated with staff perceptions of the effectiveness of family group conferences. *Journal of the Society for Social Work and Research*, *6*(3), 343–366. https://doi.org/10.1086/682730

Manuscript submitted: May 26, 2019
First revision submitted: June 13, 2019
Accepted: June 14, 2019
Electronically published: September 30, 2020