Recalibration of Predicted Probabilities Using the "Logit Shift": Why Does It Work, and When Can It Be Expected to Work Well?

Evan T. R. Rosenman[1], Cory McCartan[2] and Santiago Olivella[3]

[1] Data Science Initiative, Harvard University, Cambridge, MA 02138, USA. E-mail: erosenm@fas.harvard.edu
[2] Department of Statistics, Harvard University, Cambridge, MA 02138, USA. E-mail: cmccartan@g.harvard.edu
[3] Department of Political Science, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA.
E-mail: olivella@unc.edu

Abstract

The output of predictive models is routinely recalibrated by reconciling low-level predictions with known quantities defined at higher levels of aggregation. For example, models predicting vote probabilities at the individual level in U.S. elections can be adjusted so that their aggregation matches the observed vote totals in each county, thus producing better-calibrated predictions. In this research note, we provide theoretical grounding for one of the most commonly used recalibration strategies, known colloquially as the "logit shift." Typically cast as a heuristic adjustment strategy (whereby a constant correction on the logit scale is found, such that aggregated predictions match target totals), we show that the logit shift offers a fast and accurate approximation to a principled, but computationally impractical adjustment strategy: computing the posterior prediction probabilities, conditional on the observed totals. After deriving analytical bounds on the quality of the approximation, we illustrate its accuracy using Monte Carlo simulations. We also discuss scenarios in which the logit shift is less effective at recalibrating predictions: when the target totals are defined only for highly heterogeneous populations, and when the original predictions correctly capture the mean of true individual probabilities, but fail to capture the shape of their distribution.

Keywords: recalibration, Poisson–Binomial distribution, logit shift, election prediction

1 Problem Description

A common problem in predictive modeling is that of calibrating predicted probabilities to observed totals. For example, an analyst may generate individual-level scores $p_i \in (0,1), i = 1,\ldots,N$, to estimate the probability that each of the N registered voters in a particular voting precinct will support the Democratic candidate in an upcoming election. After the election, the analyst can observe the total number of Democratic votes, D, cast among the subset $\mathcal{V} \subset \{1,\ldots,N\}$ of registered voters who cast a ballot. However, she cannot observe individual-level outcomes due to the secret ballot. In the absence of perfect prediction, the analyst will find that $\sum_{i \in \mathcal{V}} p_i \neq D$. She must decide how to compute recalibrated scores, $\tilde{p}_i$, to better reflect the realized electoral outcome.

This practical problem has direct implications for public opinion research. For example, Ghitza and Gelman (2020) recalibrate their Multilevel Regression and Postratification (MRP) estimates of voter support levels after an election to match county-level totals, whereas Schwenzfeier (2019) proposes using the amount by which original predictions are adjusted in the recalibration exercise to estimate nonresponse bias in public opinion polls. The problem is also important to campaign work. Campaigns frequently seek to target voters who are likely to have supported their party in the prior presidential election. Estimates of prior party support may also serve as predictor variables in models estimating support in successive elections. Recalibrating the scores to match known aggregated outcomes is a crucial step to improve the scores' accuracy and bolster future electioneering.

Political Analysis (2023)
vol. 31: 651–661
DOI: 10.1017/pan.2022.31

Published
9 January 2023

Corresponding author
Evan T. R. Rosenman

Edited by
Daniel Hopkins

A common heuristic solution to the recalibration problem is the so-called "logit shift" (e.g., Ghitza and Gelman 2013; Ghitza and Gelman 2020; Hanretty, Lauderdale, and Vivyan 2016; Kuriwaki *et al.* 2022).[1] To motivate this approach, consider a simple scenario in which the p_i are generated from a logistic regression model. The recalibrated scores $\tilde{p}_i$ are then computed by uniformly shifting the model's intercept until the $\tilde{p}_i$ sum to the desired total D, with all other coefficients kept constant.

Explicitly, we define the scalar $\alpha \in [0, \infty)$ such that its log is equal to the intercept shift,

$$\text{logit}(\tilde{p}_i) = \text{logit}(p_i) - \log(\alpha), \tag{1}$$

where $\text{logit}(z) = \log(z/(1-z))$. Denote also the inverse of the logit function as $\sigma(z) = \exp(z)/(1 + \exp(z))$. We next define the summed, recalibrated probabilities as a function of α,

$$h(\alpha) = \sum_{i \in \mathcal{V}} \tilde{p}_i = \sum_{i \in \mathcal{V}} \sigma\left(\text{logit}(p_i) - \log(\alpha)\right), \tag{2}$$

and solve for the value of α that satisfies the equation

$$h(\alpha) = D. \tag{3}$$

The function $h(\cdot)$ is monotonic in α, so Equation (3) can be solved in logarithmic time using binary search. The resulting scores $\tilde{p}_i$ are defined explicitly in Equation (1), and they recalibrate the original predictions so that $\sum_{i \in \mathcal{V}} \tilde{p}_i = D$.

This approach does not depend on how the original p_i are estimated—so while it is common for these scores to be obtained via logistic regression, it is possible to implement the logit shift with any model that produces predicted probabilities. An alternative characterization of this approach emerges from information theory: solving Equation (3) is equivalent to finding the set of probabilities $\tilde{p}_i$ which sum to D and minimize the summed Kullback–Leibler divergence (Kullback and Leibler 1951) between the distribution induced by $\tilde{p}_i$ and the distribution induced by the original scores p_i.[2]

Such a recalibration strategy cannot universally be expected to perform well. As the logit shift is rank-preserving, it cannot correct for substantial heterogeneity in the direction of prediction errors in the individual p_i's (e.g., instances in which prediction errors are negative for Black voters but positive for White voters). Furthermore, as it relies on limited information conveyed by the aggregated outcomes to determine the best value for its constant shift, it cannot rectify instances in which these original predictions get the average scores right, but miss the shape of the true score distribution altogether (e.g., when individual predicted probabilities are bell-shaped, but true probabilities are more uniformly distributed).[3]

[1] The procedure is also sometimes referred to as the "logit swing," and it bears resemblance to the commonly used Platt scaling procedure for calibrating support vector machines and other margin classifiers (Platt *et al.* 1999). The logit shift, however, is designed to calibrate to observed totals when individual scores are not observable.

[2] For more details, see the Supplementary Material.

[3] That researchers can generate *worse*-calibrated scores with the logit shift is demonstrated by the following simple example: suppose that there is a set of true individual-level probabilities, p_i^{true}, such that the precinct-level Democratic vote tally is sampled as

$$D = \sum_{i \in \mathcal{V}} \text{Bern}(p_i^{\text{true}}), \tag{4}$$

and consider the extreme case in which the original predicted scores are exactly correct: $p_i^{\text{true}} = p_i$. Under the sampling model in Equation (4), and assuming independence across $i \in \mathcal{V}$, the precinct total D has variance

$$\text{var}(D) = \sum_{i \in \mathcal{V}} p_i^{\text{true}}(1 - p_i^{\text{true}}).$$

https://doi.org/10.1017/pan.2022.31 Published online by Cambridge University Press

In this research note, we provide analytical justification for using the logit shift under circumstances in which it can be expected to generate better-calibrated scores, and illustrate conditions under which this heuristic strategy can fail to improve calibration of a set of predicted probabilities. To do so, we introduce a principled procedure for score updating which computes the updated scores as posterior probabilities, conditional on observed totals. In our running example, this means we treat the original scores p_i as a kind of prior Democratic support probability, whereas the updated scores $\tilde{p}_i$ reflect conditional Democratic voting probabilities given observed aggregate outcomes. Next, we show that this Bayesian take on recalibration is well approximated by the heuristic logit shift in large samples, demonstrating this result both analytically and through a simulation study. Then, we rely on similar simulation exercises to illustrate conditions under which the logit shift can fail as a recalibration strategy. We conclude with a discussion of potential extensions of the logit shift for other recalibration problems.

2 Recalibration as a Posterior Update

To motivate the posterior update approach, we introduce additional notation. We define each voter i's choice as a binary variable $W_i \in \{0,1\}$, where $W_i = 1$ signifies a Democratic vote and $W_i = 0$ signifies a Republican vote.[4] The W_i are modeled as independent Bernoulli random variables, where $W_i \sim \text{Bern}(p_i)$. The $p_i = \mathbb{P}(W_i = 1)$ can be thought of as the prior, unconditional probability of casting a Democratic vote. In this model, scores can straightforwardly be recalibrated by defining a set of updated scores, $\{p_i^\star\}$ (which automatically sum to D over actual voters $i \in \mathcal{V}$) using the following conditional probability:

$$\begin{aligned} p_i^\star &= \mathbb{P}\left(W_i = 1 \,\middle|\, \sum_{j\in\mathcal{V}} W_j = D\right) \\ &= \frac{\mathbb{P}\left(W_i = 1, \sum_{j\in\mathcal{V}} W_j = D\right)}{\mathbb{P}\left(\sum_{j\in\mathcal{V}} W_j = D\right)} \\ &= p_i \times \frac{\mathbb{P}\left(\sum_{j\neq i} W_j = D-1\right)}{\mathbb{P}\left(\sum_{j\in\mathcal{V}} W_j = D\right)}, \end{aligned} \tag{5}$$

where a sum taken over "$j \neq i$" is understood to mean a sum over all voters in $\mathcal{V}$ other than i.

From the final line of Equation (5), we observe that the recalibrated $p_i^\star$ is obtained by multiplying the original p_i by a unit-specific probability ratio. The numerator represents the probability that there are $D-1$ Democratic votes among all voters in $\mathcal{V}$ *except* voter i, whereas the denominator represents the probability that there are D Democratic votes among all voters in $\mathcal{V}$. Given our assumptions about the W_i, computing each of these probabilities requires evaluating the distribution function of a Poisson–Binomial random variable, which emerges as the sum of independent *but not identically distributed* Bernoulli random variables (Chen and Liu 1997).

While simple and theoretically elegant, this recalibration approach is highly impractical. Calculation of Poisson–Binomial probabilities is extremely computationally demanding even at moderate sample sizes, despite recent advances in the literature (Junge 2020; Olivella and Shiraito 2017). To compute the recalibrated $p_i^\star$ values, we would need to compute one unique Poisson–Binomial

If even a modest proportion of the p_i^{true} are far from 0 and 1, it is unlikely that any given sample satisfies $D = \mathbb{E}[D] = \sum_{i\in\mathcal{V}} p_i^{\text{true}}$ (and it is impossible if $\mathbb{E}[D]$ is not an integer). That is, even a perfectly calibrated set of predictions may fail to aggregate to the observed D. Under these conditions, conducting the logit shift after observing a value of D will almost always adjust the probabilities *away* from their true values p_i^{true}, worsening the calibration of the initial scores p_i.

[4] Throughout this note, we focus on a binary outcome for simplicity. The same logic applies in cases of multinomial outcomes, however, by using the Poisson–Multinomial distribution (see Lin, Wang, and Hong 2022).

https://doi.org/10.1017/pan.2022.31 Published online by Cambridge University Press

probability for each voter. Hence, if the number of actual voters $|\mathcal{V}|$ were even modestly large, it would be computationally infeasible to obtain these exact posterior probabilities.

2.1 The Logit Shift Approximates the Recalibrating Posterior

2.1.1 *Preliminaries.*

In this section, we show analytically why the logit shift is a good approximation to the posterior update in Equation (5). We begin by defining two terms. In direct analogy to the right-hand side of Equation (2), we define the function

$$f(s,t) = \sigma\left(\text{logit}(s) + \log(t)\right) = \frac{1}{1 + \frac{1-s}{s}(t)}.$$

Next, we define the Poisson–Binomial ratio

$$\phi_i = \frac{\mathbb{P}\left(\sum_{j \neq i} W_j = D\right)}{\mathbb{P}\left(\sum_{j \neq i} W_j = D - 1\right)}.$$

Simple substitution, along with a useful recursive property of the Poisson–Binomial distribution,[5] makes clear that

$$\begin{aligned}
\sum_{i \in \mathcal{V}} f(p_i, \phi_i) &= \sum_{i \in \mathcal{V}} \frac{1}{1 + \frac{1-p_i}{p_i} \phi_i} \\
&= \sum_{i \in \mathcal{V}} \frac{1}{1 + \frac{1-p_i}{p_i} \frac{\mathbb{P}\left(\sum_{j \neq i} W_j = D\right)}{\mathbb{P}\left(\sum_{j \neq i} W_j = D-1\right)}} \\
&= \sum_{i \in \mathcal{V}} \frac{p_i \times \mathbb{P}\left(\sum_{j \neq i} W_j = D - 1\right)}{p_i \times \mathbb{P}\left(\sum_{j \neq i} W_j = D - 1\right) + (1 - p_i) \times \mathbb{P}\left(\sum_{j \neq i} W_j = D\right)} \qquad (6) \\
&= \sum_{i \in \mathcal{V}} \frac{\mathbb{P}\left(W_i = 1, \sum_{i \in \mathcal{V}} W_i = D\right)}{\mathbb{P}\left(\sum_{i \in \mathcal{V}} W_i = D\right)} \\
&= \sum_{i \in \mathcal{V}} p_i^{\star} \\
&= D.
\end{aligned}$$

In words, Equation (6) shows that the unit-specific ϕ_i is precisely the "shift" (in the sense of the second argument to the function f) that turns each p_i into the desired, recalibrated posterior probability $p_i^{\star}$. The logit shift, however, uses a constant α to approximate the vector of recalibrating shifts $\{\phi_i\}_{i \in \mathcal{V}}$. What remains, therefore, is to show that the single value of α that solves Equation (3) is a very good approximation of ϕ_i for all values of i.

To do so, we establish that the value of α is bounded by the range of $\{\phi_i\}_{i \in \mathcal{V}}$, and that each ϕ_i, in turn, has well-defined bounds. This will allow us to find that, in practice, the range of values that the unit-specific shifts ϕ_i can take is very small, and thus that a constant shift α can approximate them very well.

[5] Namely,

$$\mathbb{P}\left(\sum_j W_j = D\right) = p_i \times \mathbb{P}\left(\sum_{j \neq i} W_j = D - 1\right) + (1 - p_i) \times \mathbb{P}\left(\sum_{j \neq i} W_j = D\right).$$

https://doi.org/10.1017/pan.2022.31 Published online by Cambridge University Press

Theorem 1. *The value of α which solves Equation (3) satisfies*

$$\min_i \frac{\mathbb{P}\left(\sum_{j\neq i} W_j = D\right)}{\mathbb{P}\left(\sum_{j\neq i} W_j = D-1\right)} \leq \alpha \leq \max_i \frac{\mathbb{P}\left(\sum_{j\neq i} W_j = D\right)}{\mathbb{P}\left(\sum_{j\neq i} W_j = D-1\right)}.$$

Proof. The proof can be found in the Supplementary Material. □

Theorem 2. *For any choice of $i \in \mathcal{V}$, we have*

$$\frac{\mathbb{P}\left(\sum_{j\in\mathcal{V}} W_j = D+1\right)}{\mathbb{P}\left(\sum_{j\in\mathcal{V}} W_j = D\right)} \leq \frac{\mathbb{P}\left(\sum_{j\neq i} W_j = D\right)}{\mathbb{P}\left(\sum_{j\neq i} W_j = D-1\right)} \leq \frac{\mathbb{P}\left(\sum_{j\in\mathcal{V}} W_j = D\right)}{\mathbb{P}\left(\sum_{j\in\mathcal{V}} W_j = D-1\right)}.$$

Proof. The proof can be found in the Supplementary Material. □

2.1.2 *Main Results.* The bounds from Theorem 2 apply regardless of the choice of i, so we can combine the two theorems to observe

$$\begin{aligned}\frac{\mathbb{P}\left(\sum_{j\in\mathcal{V}} W_j = D+1\right)}{\mathbb{P}\left(\sum_{j\in\mathcal{V}} W_j = D\right)} &\leq \min_i \frac{\mathbb{P}\left(\sum_{j\neq i} W_j = D\right)}{\mathbb{P}\left(\sum_{j\neq i} W_j = D-1\right)} \leq \alpha \\ &\leq \max_i \frac{\mathbb{P}\left(\sum_{j\neq i} W_j = D\right)}{\mathbb{P}\left(\sum_{j\neq i} W_j = D-1\right)} \leq \frac{\mathbb{P}\left(\sum_{j\in\mathcal{V}} W_j = D\right)}{\mathbb{P}\left(\sum_{j\in\mathcal{V}} W_j = D-1\right)}.\end{aligned} \tag{7}$$

This is useful, because we can now use the outer bounds in Equation (7) to obtain a bound on the approximation error when estimating recalibrated scores $p_i^\star$ (obtained from the posterior update approach) via $\tilde{p}_i$ (obtained from the logit shift).

Theorem 3. *For large sample sizes, we obtain*

$$\tilde{p}_i = p_i^\star + \mathcal{O}\left(\frac{1}{\sum_{j\in\mathcal{V}} p_j(1-p_j)}\right).$$

Proof. The proof can be found in the Supplementary Material. □

Theorem 3 relies on the tightness of the bound in (7). Under the assumption of an independent Bernoulli sampling model for individual vote choices, the upper and lower bounds differ by a factor inversely proportional to the variance of D—the Poisson–Binomial variable representing total votes for the Democratic candidate. Theorem 3 states that the error in using the logit shift to approximate the posterior recalibration update is bounded by a term of the same order.

Thus, Theorem 3 implies that the magnitude of the approximation error is inversely proportional to sample size, and becomes quite small for large enough samples. As the binding bounds in Equation (7) are tight for even moderately large $|\mathcal{V}|$, the approximation can be expected to perform well in most practical settings.

However, Theorem 3 *does not* imply that the logit shift is a universally good recalibration strategy. Rather, it implies that when a strategy like the posterior update is appropriate, the logit shift offers a very close approximation at a low computational cost. Next, we illustrate the approximation's accuracy for even modestly sized electorates, under various possible score distributions, using a simple Monte Carlo simulation.

https://doi.org/10.1017/pan.2022.31 Published online by Cambridge University Press

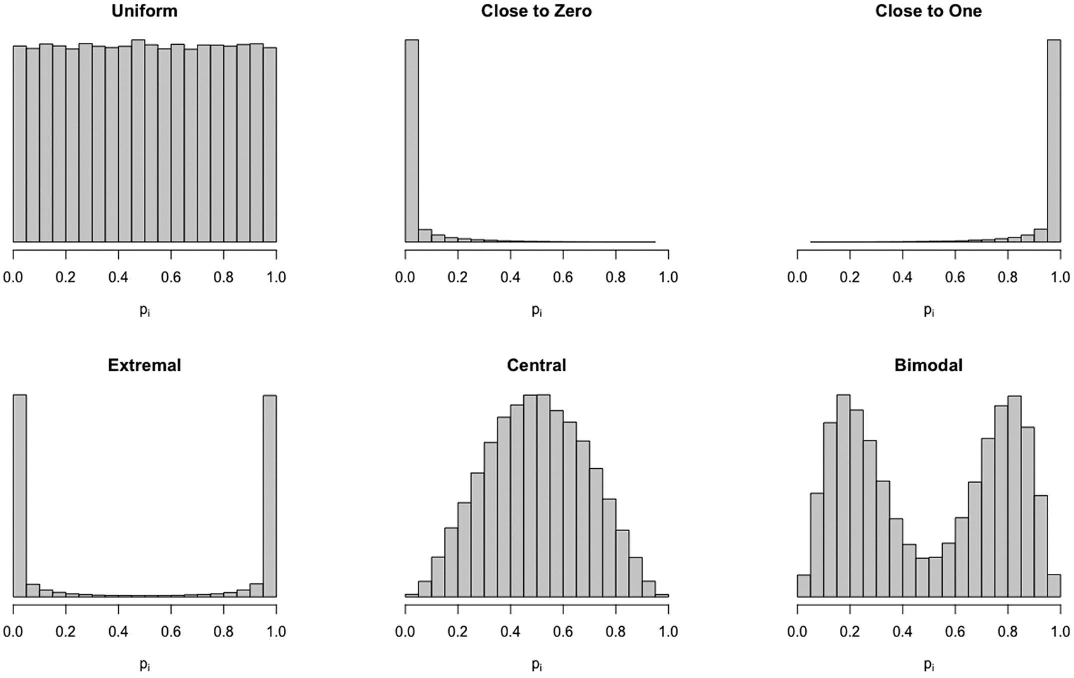

Figure 1. The distributions used in simulations in Section 2.2. We assume that the initial scores p_i and the true probabilities p_i^{true} are drawn from each of these six distributions. These are the same as the distributions provided in Biscarri, Zhao, and Brunner (2018).

Table 1. Discrepancy between the logit shift and the exact Poisson–Binomial probabilities (as measured by RMSE, $1 - R^2$, and KL divergence), under various settings. All results are calculated in the case where the observed D is equal to $0.8 \times \sum_i p_i$.

p_i setting	Sampling distribution	Sample size	RMSE	$1-R^2$	KLD
Uniform	Uniform(0, 1)	100	0.00195	5.81×10^{-5}	1.21×10^{-3}
Uniform	Uniform(0, 1)	1,000	0.00021	5.51×10^{-7}	1.43×10^{-4}
Close to 0	Beta(0.1, 3)	100	0.00772	1.06×10^{-2}	1.68×10^{-2}
Close to 0	Beta(0.1, 3)	1,000	0.00043	3.11×10^{-5}	5.24×10^{-4}
Close to 1	Beta(3, 0.1)	100	0.00369	1.12×10^{-4}	6.38×10^{-3}
Close to 1	Beta(3, 0.1)	1,000	0.00034	1.12×10^{-6}	5.13×10^{-4}
Extremal	0.5*Beta(0.1, 3) + 0.5*Beta(3, 0.1)	100	0.00496	1.16×10^{-6}	1.22×10^{-2}
Extremal	0.5*Beta(0.1, 3) + 0.5*Beta(3, 0.1)	1,000	0.00050	1.19×10^{-6}	1.05×10^{-3}
Central	Beta(3, 3)	100	0.00161	7.66×10^{-5}	7.04×10^{-4}
Central	Beta(3, 3)	1,000	0.00016	7.16×10^{-7}	6.63×10^{-5}
Bimodal	0.5*Beta(3, 10) + 0.5*Beta(10, 3)	100	0.00227	6.72×10^{-5}	1.74×10^{-3}
Bimodal	0.5*Beta(3, 10) + 0.5*Beta(10, 3)	1,000	0.00023	6.77×10^{-7}	1.93×10^{-4}

2.2 Numerical Precision Simulations

To illustrate the precision of the logit shift approximation, we simulate two scenarios: a small-sample case (where $|\mathcal{V}| = 100$) and a more typical, modestly sized sample case (with $|\mathcal{V}| = 1,000$).[6] We draw the initial probabilities p_i according to the six distributions discussed in Biscarri, Zhao, and Brunner (2018). We consider the case in which the observed D is 20% below the expectation, $\sum_i p_i$. The six distributions are visualized in Figure 1.

[6] Code to replicate these and all other simulation results is available at https://doi.org/10.7910/DVN/7MRDUW Rosenman, McCartan, and Olivella 2022.

We compute the exact posterior probabilities using Biscarri's algorithm as implemented in the *PoissonBinomial* package (Junge 2020), and compare it against the estimates obtained using the logit shift heuristic. We report the RMSE, the proportion of variance in the posterior probabilities $p_i^\star$ that is *not* explained by our method, and the summed Kullback-Leibler (KL) divergence. Results are given in Table 1.

These results demonstrate that the logit shift and the posterior probability approach are virtually identical, as expected. Across a wide variety of distributions, we find that the two approaches deviate only nominally—even in small samples of 100 voters. Moreover, as expected, the approximation gets even more accurate as the sample size increases, as illustrated by the reduction across error metrics (sometimes of several orders of magnitude) as we move from 100 to 1,000 voters.

As an approximation, then, the logit shift can be expected to work well when the target total it relies on is based on at least a modest number of voters. We now turn to the question of when we can expect the logit shift to work as a calibration strategy.

3 Why the Logit Shift Can Fail To Generate Well-Calibrated Predictions

The close correspondence between the logit shift and the full posterior update need not mean that the logit shift produces better calibrated scores in all cases. Probabilities updated through the logit shift maintain the same ordering of the original predictions,[7] and can only correct predicted score distributions that misrepresent the location (rather than the overall shape) of the true score distribution. These limitations can prevent the updated scores from improving the calibration of predicted probabilities, and can even exacerbate calibration problems among subsets of voters.

We discuss how these issues can manifest in practice, and illustrate the potential problems using Monte Carlo simulations. We again adopt the independent Bernoulli model of Section 2, associating with each individual a true Democratic support probability p_i^{true} as well as an initial predicted score p_i. We investigate whether logit shifting the p_i scores generates updated predictions $\tilde{p}_i$ that are more closely aligned with p_i^{true} than the original scores p_i.

3.1 Heterogeneity and Target Aggregation Levels

Perhaps the most serious limitation of the logit shift stems from the fact that it cannot alter the ordering of the original probabilities p_i. This has implications for the ideal grouping level at which we conduct the logit shift. Throughout this note, we have supposed that the logit shift is used within each voting precinct to generate updated scores whose sum is equal to the precinct's vote total. Yet it is entirely plausible (and indeed common in academic research) to conduct the update at higher levels of aggregation, e.g., counties or states.

Executing the update at higher levels of aggregation, however, can imply more heterogeneity in prediction errors—heterogeneity that may not be correctable using the rank-preserving logit shift with a single target total (Kuriwaki *et al.* 2022). To see why, consider an example in which there are two groups of voters: Black voters and White voters. Suppose that for all Black voters, $p_i = 0.7$ and $p_i^{\text{true}} = 0.8$, whereas for all White voters, $p_i = 0.3$ and $p_i^{\text{true}} = 0.2$—i.e., the initial scores underestimate Democratic support among Black voters and overestimate Democratic support among White voters. The logit shift will either increase or decrease all probabilities within a given grouping. Suppose that we choose a high level of aggregation—e.g., the state level—at which to conduct the logit shift, and White voters constitute a large majority of the state's voters. The predicted tally of Democratic votes will significantly overshoot the observed vote count D. Hence, the logit shift will adjust all voters' Democratic support probabilities downward. This will yield

[7] In fact, while this may seem like it is the result of using a constant α to update all scores, it is in fact a property of the Bayesian update strategy, which is itself rank-preserving (despite defining individual "shifts").

https://doi.org/10.1017/pan.2022.31 Published online by Cambridge University Press

Table 2. We consider three racial proportions within the precinct, and report $\frac{\mathrm{cor}(\tilde{p}_i, p_i^{\mathrm{true}}) - \mathrm{cor}(p_i, p_i^{\mathrm{true}})}{\mathrm{cor}(p_i, p_i^{\mathrm{true}})}$ separately for White and Black voters in each pair of columns. Initial scores are drawn from a different distribution in each row. Positive values means the logit shift has improved the correlation with the true probabilities, whereas negative values mean the logit shift has worsened the correlations with the true probabilities.

Initial score dist	W (70%)	B (30%)	All	W (80%)	B (20%)	All	W (90%)	B (10%)	All
Uniform	0.010	−0.017	0.002	0.011	−0.023	0.005	0.012	−0.032	0.007
Close to 0	0.072	−0.062	0.011	0.126	−0.109	0.043	0.249	−0.366	0.160
Close to 1	0.037	−0.071	0.017	0.069	−0.175	0.042	0.065	−0.209	0.050
Extremal	0.043	−0.046	0.015	0.083	−0.117	0.041	0.094	−0.083	0.077
Central	0.005	−0.009	0.001	0.004	−0.009	0.001	0.004	−0.004	0.003
Bimodal	0.004	−0.006	0.001	0.006	−0.011	0.003	0.006	−0.014	0.005

improved predictions for all White voters, but worse predictions for Black voters, whose initial projected support levels were too low rather than too high.

To illustrate the potential issues raised by heterogeneity, we simulate a two-group situation like the one just described. We suppose that there are only White and Black voters present in a precinct of $n = 1{,}000$ individuals, and again suppose that the initial scores are drawn from the same six probability distributions visualized in Figure 1. We assume further that the majority of voters are White, and that their Democratic support probabilities are overestimated by 10 percentage points, whereas a minority of voters are Black, and their Democratic support probabilities are underestimated by 10 percentage points.[8] Crucially, the ordering of p_i and p_i^{true} is *not* the same in this setting.

We consider three racial proportions within the precinct: a 70–30 split of White and Black voters, an 80–20 split, and a 90–10 split. In each case, we sample the initial scores p_i; then compute the true probabilities p_i^{true}; then sample the aggregated outcomes, and conduct the logit shift of p_i. We then report

$$\frac{\mathrm{cor}(\tilde{p}_i, p_i^{\mathrm{true}}) - \mathrm{cor}(p_i, p_i^{\mathrm{true}})}{\mathrm{cor}(p_i, p_i^{\mathrm{true}})}$$

as our success metric. Positive values mean that the logit shift has improved the correlation with the true probabilities, whereas negative values indicate that the logit shift has worsened the correlations with the true probabilities. We compute the quantity separately for Black and White voters, and report results in Table 2.

As expected, in each setting, scores get better for White voters and worse for Black voters. The relative changes are largest when the precinct is 90% White, in which case significant accuracy can be lost for Black voters. The correlation computed over all voters improves in these more homogeneous precincts, as large improvements are achieved for a large proportion of the voters therein (i.e., for White voters).

Accordingly, using a lower level of aggregation—e.g., voting precincts—will ameliorate the problem only if precincts are more racially homogeneous than the state as a whole. While the errors may increase for the minority in the more homogeneous context, the overall calibration of the updated scores would increase, as a larger proportion of the voters would be accurately

[8] In our simulations, the shifts are themselves induced by first randomly designating voters as White or Black, and then applying the appropriate logit shift to their scores. A Beta(0.1, 0.5) is substituted for the "Close to 0" distribution and a Beta(0.5, 0.1) for the "Close to 1" distribution, so as to allow for these shifts to be plausible.

https://doi.org/10.1017/pan.2022.31 Published online by Cambridge University Press

adjusted. Thankfully, we can generally expect greater homogeneity within smaller aggregation units than what would be observed in the electorate as a whole.

Theorem 4. *Consider two sets of aggregation units, $\mathcal{A}$ and $\mathcal{B}$, with the $\mathcal{A}$ units nested inside the $\mathcal{B}$ units. Then, for each $\mathcal{B}$ unit, the overall proportion of people comprising a minority within their $\mathcal{A}$ unit is at least as small as the proportion of people comprising a minority within the enclosing $\mathcal{B}$ unit.*

Proof. The proof can be found in the Supplementary Material. □

For example, based on the 2020 Census data and census race/ethnicity categories, 39.8% of the voting-age population was a minority nationwide, 38.7% was a minority within their state, 35.3% was a minority within their county, 28.7% was a minority within their tract, and 12.9% was a minority within their Census block (U.S. Census Bureau 2021). This implies that researchers and practitioners would benefit from applying the logit shift at the lowest level of aggregation for which aggregate data are available—provided the number of votes being aggregated is large enough to ensure that the logit shift accurately approximates the posterior update.[9]

3.2 Limits to What Can Be Learned From a Total

While it is clear that the logit shift cannot fully ameliorate errors when the ordering of p_i and p_i^{true} differs, it is also possible to observe little improvement in calibration even in contexts in which the initial ordering is correct. In such instances, we will only see gains from applying the logit shift if the observed target total D differs substantially from the expected total under the set of initial scores.

To see why, recall that the logit shift (and the posterior it approximates) relies only on information about the aggregated outcomes to update individual scores. This total is most informative about the mean of the true individual probabilities, as differently shaped distributions of p_i^{true} that are consistent with the observed total can share the same mean, but distributions with different means will typically result in different observed totals. As a result, even if initial individual scores are ranked correctly, the extent to which the observed total will provide useful information will depend on the extent to which the mean of the initial scores differs from the mean of the true probabilities. This highlights an important weakness of the logit shift: it cannot correct for an incorrect *shape* of the initial score distributions, but only for an incorrect *mean*.

To illustrate this issue through simulation, we use the same six probability distributions used in Table 1, but we alter the setup. Consider each of the 36 possible pairs of distributions. For each pair, we sample 1,000 voters such that the true probabilities p_i^{true} follow the first distribution, and the initial scores p_i follow the second distribution, but the rank of each unit i is identical within each of the two distributions. We sample the outcomes under the Bernoulli model, and conduct the logit shift on the initial scores p_i to generate the updated scores $\tilde{p}_i$. Table 3 reports the results, with each entry again presenting

$$\frac{\text{cor}(\tilde{p}_i, p_i^{\text{true}}) - \text{cor}(p_i, p_i^{\text{true}})}{\text{cor}(p_i, p_i^{\text{true}})}$$

from the simulation involving the corresponding distributions.

Recall that the uniform, extremal, central, and bimodal distributions all have means of 0.5. As expected, if both the true probability distribution and the initial score distribution have the

[9] As the results in Table 1 indicate, the approximation is already highly accurate when aggregating even 100 people. Thus, considering the average number of voters per precinct in the United States is approximately 1,000, the size constraints are not difficult to satisfy.

Table 3. In each cell, we report $\frac{\mathrm{cor}(\tilde{p}_i, p_i^{\mathrm{true}}) - \mathrm{cor}(p_i, p_i^{\mathrm{true}})}{\mathrm{cor}(p_i, p_i^{\mathrm{true}})}$ when the initial scores and true probabilities are drawn from the respective row and column distributions. Positive values means the logit shift has improved the correlation with the true probabilities, whereas negative values mean the logit shift has worsened the correlations with the true probabilities.

		Initial prediction distribution					
		Uniform	**Close to 0**	**Close to 1**	**Extremal**	**Central**	**Bimodal**
True distr.	**Uniform**	0.000	0.653	0.841	0.000	0.000	0.000
	Close to 0	0.648	0.000	4.609	1.681	0.461	0.889
	Close to 1	0.675	2.361	0.000	1.463	0.459	0.914
	Extremal	0.000	1.147	1.125	0.000	0.001	0.000
	Central	0.000	0.548	0.487	0.000	0.000	0.000
	Bimodal	0.000	0.719	0.837	0.000	−0.001	0.000

same mean, the correlation shifts are essentially zero. In contrast, much larger improvements in correlation are seen when the skewed "Close to 0" and "Close to 1" distributions (which have means of 0.032 and 0.968, respectively) are used. The intuition is clear: the observed precinct total is much more informative when it differs drastically from the mean of the initial scores.

4 Discussion

In this paper, we have considered the problem of updating voter scores to match observed vote totals from an election. We have shown that the simple "logit shift" algorithm is a very good approximation to computing the exact posterior probability that conditions on the observed total. This is a useful insight for campaign analysts and researchers alike, because the logit shift is significantly more computationally efficient than the calculation of the exact posterior recalibration update, yet the approximation is extremely accurate even in small samples.

We have also discussed limitations of this approach in terms of its ability to recover a true set of individual support probabilities. Crucially, logit-shifted probabilities retain the same ordering as the initial set of scores, which implies that the original scoring model must discriminate positive and negative (but unobservable, in the case of voting) individual cases well. Users of the logit shift can increase the chances of having correctly ranked initial scores by applying the logit shift at low levels of aggregation, where heterogeneity of prediction errors is likely to be low. In turn, users can expect to see little improvements to calibration when their initial scores capture the correct mean of the true unit probabilities, even if the shape of the true and predicted score distributions differ. The limits of what can be learned from a single aggregated outcome about individual probabilities makes this problem hard to address in practice.

While not without pitfalls, the logit shift represents a useful and computationally efficient method of updating individual-level scores to incorporate information from a completed election. Furthermore, recent developments can help correct some of the limitations we have highlighted. For instance, minimizing differences with respect to multiple aggregated targets can help resolve issues raised by heterogeneity in prediction errors among subgroups, and provide more information about the shape of the distribution of true probabilities (e.g., Kuriwaki *et al.* 2022). A fruitful avenue for future research would explore whether these attempts can also be justified as approximations to a posterior update that conditions on multiple totals, highlighting the connections between the logit shift and the problem of ecological inference (e.g., King, Tanner, and Rosen 2004; Rosenman 2019). Establishing those connections represents a promising potential extension of the insights provided in this note.

Data Availability Statement

Replication code for this article is available in Rosenman *et al.* (2022), at https://doi.org/10.7910/DVN/7MRDUW.

Supplementary Material

For supplementary material accompanying this paper, please visit https://doi.org/10.1017/pan.2022.31.

References

Biscarri, W., S. D. Zhao, and R. J. Brunner. 2018. “A Simple and Fast Method for Computing the Poisson Binomial Distribution Function.” *Computational Statistics & Data Analysis* 122: 92–100.

Chen, S. X., and J. S. Liu. 1997. “Statistical Applications of the Poisson-Binomial and Conditional Bernoulli Distributions.” *Statistica Sinica* 7 (4): 875–892.

Ghitza, Y., and A. Gelman. 2013. “Deep Interactions with MRP: Election Turnout and Voting Patterns among Small Electoral Subgroups.” *American Journal of Political Science* 57 (3): 762–776.

Ghitza, Y., and A. Gelman. 2020. “Voter Registration Databases and MRP: Toward the Use of Large-Scale Databases in Public Opinion Research.” *Political Analysis* 28 (4): 507–531.

Hanretty, C., B. Lauderdale, and N. Vivyan. 2016. “Combining National and Constituency Polling for Forecasting.” *Electoral Studies* 41: 239–243.

Junge, F. 2020. “Package ‘PoissonBinomial’.” *Computational Statistics & Data Analysis* 59: 41–51.

King, G., M. A. Tanner, and O. Rosen. 2004. *Ecological Inference: New Methodological Strategies*. New York: Cambridge University Press.

Kullback, S., and R. A. Leibler. 1951. “On Information and Sufficiency.” *The Annals of Mathematical Statistics* 22 (1): 79–86.

Kuriwaki, S., S. Ansolabehere, A. Dagonel, and S. Yamauchi. 2022. “The Geography of Racially Polarized Voting: Calibrating Surveys at the District Level.” *OSF Preprints.* https://doi.org/10.31219/osf.io/mk9e6

Lin, Z., Y. Wang, and Y. Hong. 2022. “The Poisson Multinomial Distribution and its Applications in Voting Theory, Ecological Inference, and Machine Learning.” https://doi.org/10.48550/ARXIV.2201.04237

Olivella, S., and Y. Shiraito. 2017. “poisbinom: A Faster Implementation of the Poisson-Binomial distribution.” *R Package Version 1.0.1.*

Platt, J., et al. 1999. “Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods.” *Advances in Large Margin Classifiers* 10 (3): 61–74.

Rosenman, E. 2019. “Some New Results for Poisson Binomial Models.” https://doi.org/10.48550/ARXIV.1907.09053

Rosenman, E., C. McCartan, and S. Olivella. 2022. “Replication Data for: Recalibration of Predicted Probabilities using the ‘Logit Shift’: Why Does It Work, and When Can It Be Expected to Work Well?” Version V1. https://doi.org/10.7910/DVN/7MRDUW

Schwenzfeier, M. 2019. “Which Non-Responders Drive Non-Response Bias?” In *PolMeth XXXVI*. Cambridge.

U.S. Census Bureau. 2021. *2020 Census*. U.S. Department of Commerce.

https://doi.org/10.1017/pan.2022.31 Published online by Cambridge University Press