Annex 2 -

SOME STATISTICAL ASPECTS OF ESTIMATION OF SWING AND PREDICTION OF OUTCOMES ON ELECTION NIGHT

Report prepared for the Australian Electoral Commission

July 1986

R B Cunningham, Senior Lecturer (Statistical Consultant)
Australian National University

1. Background

On election night the Electoral Commission provides a comprehensive analysis of the count of the election. As well as providing on-line data for television networks they provide detailed summaries of the count on a tally board and on visual display units linked to their computer network. Amongst the data available for the analyst are cumulative counts for each candidate in each Division, party aggregates on a state-by-state basis and national totals. In addition changes in the percentage vote for each party since last election are provided for analysts to gauge the movement of public support from one party to another.

Since 1980 modern computer technology has provided a means by which reduction and synthesis of large amounts of rapidly changing data can be displayed. Further, the application of statistical methodology allows for a far more sophisticated analysis than in previous years and provides a framework for predicting outcomes from partial counts.

Central to any prediction whether it be by the expert psephologist or by the application of statistical procedures running on a computer, is the choice of suitable measures of electoral swing. Four measures of swing often used for discerning voting patterns are:

a) the rise in the Labour share of first preferences (ALP swing)

b) the rise in the Coalition share of first preferences (Coalition swing)

c) the average of Labor's increase and the Coalition decrease in first preferences (conventional swing)

d) the rise in the two party preferred vote (two-party preferred swing)

Because of the preferential voting system and possible confusion if the vote for the minor parties is significant, the two-party preferred swing d) provides a simple measure of overall trend in public support. It does however suffer from the defect that preference distributions of minor parties have to be nominated prior to the count. Experience suggests that small errors in the a-priori assessment of preference distributions are of little concern.

Probably the most important consideration in modelling voting trends is that votes first counted in many Divisions do not form a representative sample of votes in those Divisions. Thus one of the important steps in any analysis is to quantify the inherent bias in the count in each Division. Appropriate adjustments can then be made before estimates of swing are calculated. Knowledge of bias in past elections provides the only useful information for estimating bias. However this requires the assumption that polling booths will be consistent from one election to the next. Identification of polling booths from which the counts come will overcome many of the current uncertainties.

This report examines some statistical aspects of the effect of the proposal to identify polling places within each Division on estimates of swing. In particular Section 2 examines the effect ~~on~~ of bias and variance ~~of~~ on swing estimates and Section 3 explores methods for predicting final outcomes and hence how the intended additional information may assist. Data from three Queensland Divisions for the 1983 and 1984 elections will be used to provide example calculations.

2. Estimates of Swing

Suppose in any Division we have M polling places and that at a given stage of the count the progress total is an aggregate of m polling places. Let x_2 and n_2 represent the progressive count for one party and the total formal vote, respectively, after m polling places have reported and x_1 and n_1 the (hypothetical) aggregate count corresponding to the same in booths, at the previous election. Without loss of generality we take x to refer to the ALP two party preferred vote. Corresponding results for the ALP swing (a) and Coalition swing (b) will be given in Appendix I.

Let

$$X_1 = \sum_{i=1}^{M} x_{1i} \quad , \quad N_1 = \sum_{i=1}^{M} n_{1i}$$

$$X_2 = \sum_{i=1}^{M} x_{2i} \quad , \quad N_2 = \sum_{i=1}^{M} n_{2i}$$

be the total two-party preferred votes and the total formal votes at the two elections.

If we have no knowledge of polling places from which votes have come, the usual swing estimate is calculated by comparing progress totals to the final vote at the previous election viz.

$$Z = \frac{x_2}{n_2} - \frac{X_1}{N_1}$$

Using the knowledge that booths have been matched an obvious estimator of swing is

$$Z' = \frac{x_2}{n_2} - \frac{x_1}{n_1}$$

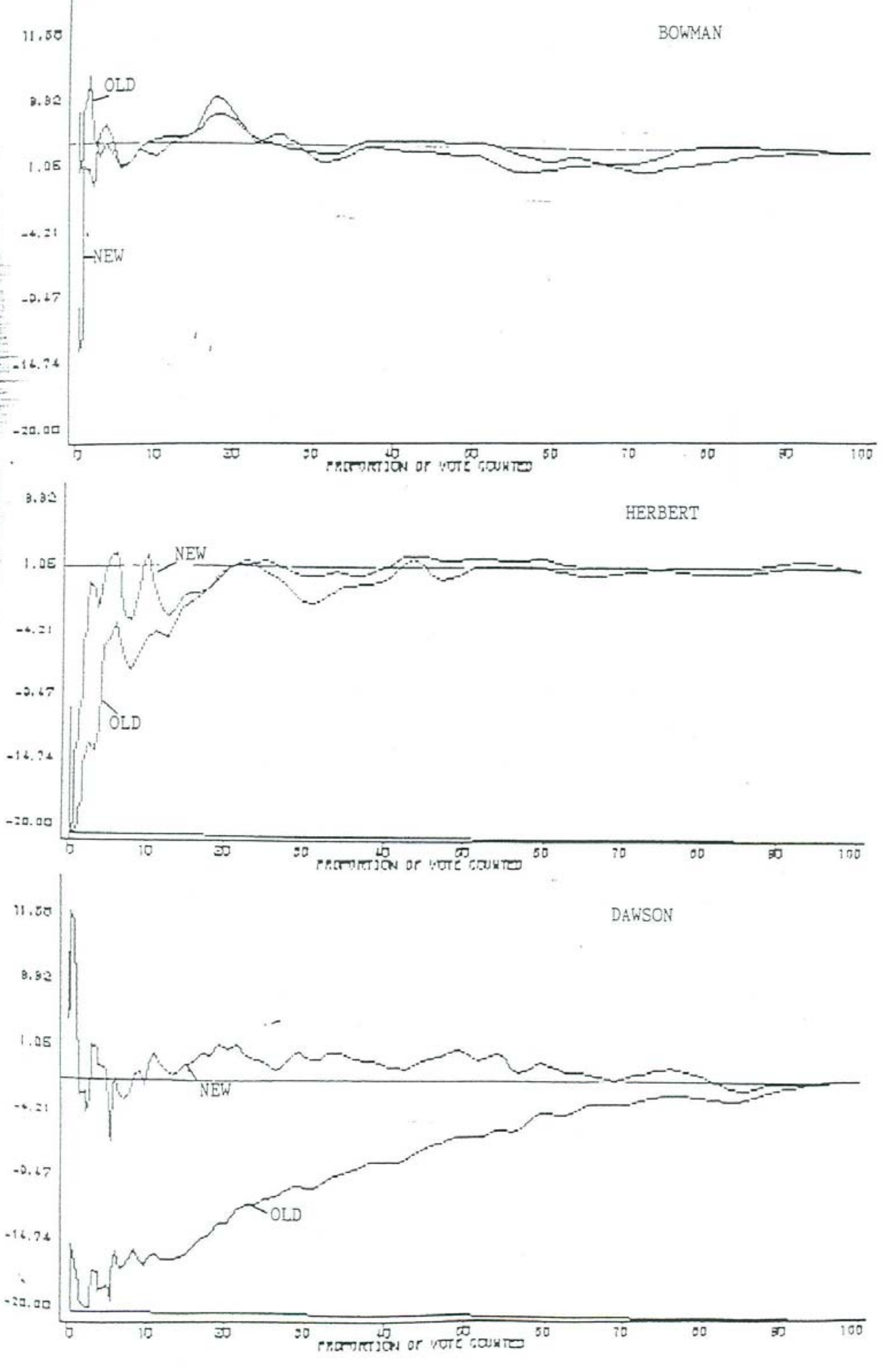
That is, we compare progress totals with totals derived from the same polling places at the previous election. These calculations require a matching of booths from one election to the next and hence the same number M of total booths. In the instance of there not being a match it is suggested that comparisons be made using only those booths for which a match is possible.

a) Bias in Swing Estimates

If booths report at random then both estimates of swing will be unbiased. However we know this to be untrue, and that the order in which booths report cannot be controlled or randomized. Thus it is very difficult to qualify bias other than by using reported progressive counts from previous elections. However we know that booth to booth variance in x_2/n_2 is large, relative to the variation in swing between booths. This will be clearly demonstrated later in this section. Hence the proposal to match booths and then compare progressive totals will substantially reduce bias.

Statistically the only way to qualify bias is to model it. This means identifying some systematic components which relate to the order in which booths report and which affect bias. One possibility is that small booths are likely to report early and if booth size correlates with bias we have a possible model for bias. This possibility has been investigated for the three sample Divisions by ordering booths according to size and calculing swing (cumulative) against the proportion of vote counted (Figure 1).

Figure 1. Cumulative swing, when booths are ordered by size, plotted against proportion (cumulative) of vote.



Inspection of the graphs in Figure 1 shows that, although bias is not eliminated by matching, a substantial reduction occurs. This is particularly so in the rural seat of Dawson, where there is a very strong Coalition vote in the small booths. As expected, variation in the vote in the small booths (comprising less than 5% of the total vote) is large.

Past experience has shown bias due to non-random reporting to be large in some Divisions, particularly rural electorates, but this is difficult to quantify. However when one takes the extreme model of accumulation of votes according to booth sizes the analysis of post data from three Divisions presented here, reveals that by invoking the simple principle of matching booths, the effect is largely eliminated.

b)Variance of Swing Estimates

In order to estimate variance we require the assumption that booths report in random order. Exploratory analysis suggests that variance is slightly larger for smaller booths. If this problem is to be addressed a model is required.

We begin by assuming that the party vote and total count are available for the current and past elections for each booth. The consequences of not having this information but having available progressive counts only, will be discussed later.

We use the notation that x_{2i}, n_{2i} and x_{1i} and n_{1i} are the party vote and total vote at each booth for the current and the past elections, respectively; $i = 1,\dots,m$.

Employing the "delta" method for estimating variance of a function of random variables, and appropriate formulae for variance of ratio estimators we obtain the following estimators:

$$\mathrm{Var}(Z) = \frac{M^2}{M}\left(1 - \frac{m}{M}\right)\frac{1}{N_2^2}\left\{\mathrm{Var}(X_2) - 2\,\frac{X_2}{N_2}\,\mathrm{Cov}(X_2,N_2) + \left(\frac{X_2}{N_2}\right)^2\mathrm{Var}(N_2)\right\} \qquad \ldots\ldots (1)$$

$$\begin{aligned}\mathrm{Var}(Z') = \frac{M^2}{M}\left(1 - \frac{m}{M}\right)\Bigg[&\frac{1}{N_2^2}\left\{\mathrm{Var}(X_2) - 2\,\frac{X_2}{N_2}\,\mathrm{Cov}(X_2,N_2) + \left(\frac{X_2}{N_2}\right)^2\mathrm{Var}(N_2)\right\}\\ &+ \frac{1}{N_1^2}\left\{\mathrm{Var}(X_1) - 2\,\frac{X_1}{N_1}\,\mathrm{Cov}(X_1,N_1) + \left(\frac{X_1}{N_1}\right)^2\mathrm{Var}(N_1)\right\}\\ &- 2\,\frac{1}{N_1}\cdot\frac{1}{N_2}\left\{\mathrm{Cov}(X_2,X_1) - \frac{X_2}{N_2}\,\mathrm{Cov}(X_1,N_2) - \frac{X_1}{N_1}\mathrm{Cov}(X_2,N_1)\right.\\ &\left. + \frac{X_1}{N_1}\cdot\frac{X_2}{N_2}\,\mathrm{Cov}(N_1,N_2)\right\}\Bigg] \qquad \ldots\ldots (2)\end{aligned}$$

Table 1 shows these estimates for the three sample Divisions, for varying values of m, the number of booths contributing to progressive totals at different stages.

TABLE 1.

Estimates of variance of swing calculated using unmatched and matched booths methods.

DIVISION

	BOWMAN		HERBERT		DAWSON	
M	34		41		64	
m	Var(Z)	Var(Z^1)	Var(Z)	Var(Z^1)	Var(Z)	Var(Z^1)
5	8.17	1.86	8.51	1.32	26.72	2.73
10	3.38	0.77	3.66	0.57	12.23	1.25
15	1.78	0.41	2.05	0.32	7.40	0.75
20	0.99	0.23	1.24	0.19	4.98	0.51
25	0.51	0.12	0.76	0.12	3.53	0.36
30	0.19	0.04	0.43	0.07	2.57	0.26
35			0.20	0.03	1.88	0.19
40			0.03	0.004	1.36	0.14
45					0.96	0.00
50					0.63	0.06
55					0.37	0.04
60					0.15	0.02

Significant gains in precision of swing estimates are clearly evident when using the proposed method of matching. Reductions of the order of 5 to 10 fold are obtained for the example data. To illustrate further, the confidence interval for swing for m=30 for the Division of Dawson is reduced from ± 3.20 units to ± 1.02 units.

Some algebraic manipulation of equation (1) and (2) shows that Var(Z') will be less than Var(Z) if the correlation between x_1 and x_2 is greater than 0.5 and the correlations between n_1 and n_2 is greater than 0.5. For the three example electorates estimates of these correlation are as follows:

Correlation Coefficients

Division	$r_{x_1x_2}$	$r_{n_1n_2}$
Bowman	0.86	0.88
Herbert	0.98	0.98
Dawson	0.98	0.99

These conditions are clearly met and it would seem likely on this evidence, that these conditions would always be true. An extremely poor match of booths from one election to the next would have to occur before matching would not be beneficial.

It will be noted that all component variance and covariance terms in equation (1) and (2) are population estimates. As only progressive counts are available an on election night these formulae cannot be applied. If individual booth data were to be available, substitution of sample estimates for population estimates would provide the required variance estimates. Again the intention is, for obvious practical reasons, not to supply such detailed information. Only a list of booths reporting together with cumulative counts is to be available.

Estimates of component variance and covariance terms should not change too much from one election to the next. We can therefore regard these as fixed and rewrite equations (1) and (2) to provide approximate sample estimates as follows:

$$\mathrm{Var}(Z) = \tilde{M}(1 - \frac{\tilde{M}}{M}) \frac{1}{n_2^2} \mathrm{Var}(\frac{X_2}{N_2}) \qquad \ldots\ldots (3)$$

$$\mathrm{Var}(Z') = m(1 - \frac{m}{M}) \frac{1}{n_2^2} \mathrm{Var}(\frac{X_2}{N_2}) + \frac{1}{n_1^2} \mathrm{Var}(\frac{X_1}{N_1}) - 2 \frac{1}{n_1} . \frac{1}{n_2} \mathrm{Cov}(\frac{X_1}{N_1}, \frac{X_2}{N_2}) \qquad \ldots\ldots (4)$$

where the terms $\mathrm{Var}(\frac{X_2}{N_2})$, $\mathrm{Var}(\frac{X_1}{N_1})$ and $\mathrm{Cov}(\frac{X_1}{N_1}, \frac{X_2}{N_2})$ are population estimates obtained from data for the two previous elections. A requirement here is that these estimates have to be obtained for each of the 148 Divisions. It is likely that for groups of

Divisions these terms will be similar and so considerably reduce the effort required in computing them. Identification of such groups (eg socio-demographic classes) is beyond the scope of this report.

For illustration purposes a pooled estimate of each has been obtained from the three sample Division and some simulated estimates of variance obtained. The component variance and covariance terms were:

$$\begin{array}{ll}
\mathrm{Var}(X_1) = 203906 & \mathrm{Cov}(X_1,X_2) = 199564 \\
\mathrm{Var}(X_2) = 215397 & \mathrm{Cov}(X_1,N_1) = 354536 \\
\mathrm{Var}(N_1) = 640142 & \mathrm{Cov}(X_1,N_2) = 351436 \\
\mathrm{Var}(N_2) = 679277 & \mathrm{Cov}(X_2,N_1) = 346794 \\
 & \mathrm{Cov}(X_2,N_2) = 375075 \\
 & \mathrm{Cov}(N_1,N_2) = 633530
\end{array}$$

For $\frac{X_1}{N_1}$ we use the known value from the previous election and for $\frac{X_2}{N_2}$ we substitute $\frac{x_2}{n_2}$, the sample estimate.

The simulations consisted of 20 random selections of m booths. The following table summarizes the results for several selected values of m, for the three chosen Divisions.

Table 2: Summary statistics of 20 simulated selections of booths. Var(z) and Var(z') calculated using equations (3) and (4) respectively.

		Swing			Var(z)			Var(z')			% counted		
Division	m	min	median	max	min	median	max	min	median	max	min	median	max
BOWMAN swing = 2.2	5	-0.7	2.0	3.8	3.1	8.9	18.8	0.8	2.1	9.0	11	14	23
	10	-0.3	2.1	3.3	2.0	3.0	6.5	0.5	0.8	2.0	21	30	39
	20	1.2	2.0	3.0	0.7	0.9	1.5	0.2	0.2	0.4	46	58	68
HERBERT swing (=0.9)	5	-0.7	0.3	2.5	5.5	8.7	51.6	1.2	2.9	34.8	7	14	19
	15	-0.4	0.7	1.6	1.6	2.5	5.2	0.3	0.6	2.0	27	41	48
	30	0.4	0.9	1.2	0.4	0.5	0.7	0.1	0.1	0.1	68	80	86
DAWSON swing =-2.0	10	-3.9	-2.2	-0.0	3.8	14.8	38.7	0.8	2.9	9.4	12	18	31
	25	-2.9	-1.9	-0.9	2.0	4.1	9.3	0.4	0.7	1.8	31	44	59
	50	-2.6	-2.0	-1.4	0.6	0.8	1.0	0.1	0.1	0.2	78	88	98

3. Prediction of Final Outcomes

The analysis in Section 2 addressed the problem of estimation of swing from data obtained within a Division. The problem of predicting the final outcomes of an election is primarily concerned with an inter-Divisional analysis of variation in swing, rather than intra-Division analyses.

From the results of Section 2 we can obtain reasonably precise estimates of swing in Division which has have reported a count. On election night, for a given stage in the count, these progress counts are available for only a sample of Divisions; later in the night most Divisions will have reported a count. Typically the order in which counts become available are first the Divisions in the Eastern States, followed by those of South Australia and Northern Territory, and finally Western Australia. The order obviously reflects differences in time zones operative at the time of the election.

The essential problem of forecasting the final outcome is to predict results for those Divisions for which no count is available. A suggested method, based on regression analysis of swing is as follows:

Analysis of variation in swing between Divisions has shown that state-to-state variation together with a (socio-demographic) classification of seats into inner and outer metropolitan, provincial and rural seats are two important factors. Variables such as the magnitude of swing in previous elections and a measure of how safe the seat is, also provide useful predictions of swing.

Given progressive counts from K Divisions a possible regression model for swing is therefore

$$Z'_k = \underset{\sim}{a}_k^T \underset{\sim}{\beta} + \delta_k + \varepsilon_k$$

where $\underset{\sim}{a}_k$ are the explanatory variables including State, demographic class etc. terms,

$\underset{\sim}{\beta}$ is the vector of unknown parameters,

δ_k is the intra-Division error component, and

ε_k is the inter-Division error component; $k = 1,\ldots,K$.

Now

$$\mathrm{Var}(\delta_k) = \hat{v}_k \quad (= \mathrm{Var}(Z_k') \text{ given in Section 2)}$$

$$\mathrm{Var}(\varepsilon_k) = \sigma^2 \text{ (constant)}$$

so

$$\mathrm{Var}(Z_k') = \hat{v} + \sigma^2 = w_k$$

Estimation of $\underset{\sim}{\beta}$ is by weighted least squares, where the weights are $1/w_k$. For computational convenience, as well as possible gains in stability of predictions, it may be advantageous to use ridge regression.

For all Divisions, including those for which no count is available, we have a regression estimate of swing which is

$a_k^T \hat{\underset{\sim}{\beta}}$, where $\hat{\underset{\sim}{\beta}}$ are the estimates of the unknown parameters $\underset{\sim}{\beta}$, with variance

$$a_k^T V a_k = \tilde{v}_k, \text{ say.}$$

V is the estimated variance-covariance matrix of the parameters, $\hat{\underset{\sim}{\beta}}$.

Let $\hat{p}_k$ be the intra-Division estimate of party vote for Division k, and $\tilde{p}_k$ the regression estimate (inter-Division). These are obtained by adding the respective swing estimates to the party vote at the last election. That is

$$\hat{p}_k = \frac{X_1}{N_1} + Z'_k$$

and

$$\tilde{p}_k = \frac{X_1}{N_1} + a_k^T \hat{\underset{\sim}{\beta}}$$

Treating $\hat{p}_k$ and $\tilde{p}_k$ as independent, which is approximately true, we obtain a combined prediction

$$p_k^* = (\tilde{v}_k\hat{p}_k + \hat{v}_k\tilde{p}_k)/(\tilde{v}_k + \hat{v}_k)$$

and

$$\text{Var}(p_k^*) = \left(\frac{1}{\tilde{v}_k} + \frac{1}{\hat{v}_k}\right)^{-1} = v_k^*$$

The approximate distribution of p_k^* is $N(p_k, v_k^*)$.

So if I_k is an indicator random variable taking a value 1 if the seat is won by the 'party' and zero otherwise, it is estimated by

$$\hat{I}_k = \Phi\left(\frac{p_k^* - 0.5}{\sqrt{v_k^*}}\right)$$

Therefore

$$S = \sum_{k=1}^{148} I_k \text{ is estimated by}$$

$$\hat{S} = \sum_{k=1}^{148} \hat{I}_k \text{ is the estimated total number of seats for the}$$

nominated party.

Also

$$\text{Var}(S) \approx \sum_{k=1}^{148} \hat{I}_k(1-\hat{I}_k).$$

Assuming Normality of S, a 95% confidence interval is obtained as

$$\hat{S} \pm 1.96 \sqrt{\text{Var}(\hat{S})}$$

4. Summary

The preceding analysis develops formulae for estimates of variance of swing calculated by the present method and the proposed method of comparing progress figures with figures derived from the same polling booths at the previous election. The two estimators are compared using data from the 1983 and 1984 elections for three Queensland Divisions.

Two clear advantages of the proposed method over the present method emerge. Firstly, an order of magnitude reduction in variance occurs resulting in significant reduction in confidence intervals for swing estimates. Secondly, although difficult to quantify, a substantial reduction in bias due to the usual non-random reporting, is likely.

Approximations to the variance formulae are given which allow calculations to be readily performed on election night. A simulation study using existing data show these approximation to be very good.

Section 3 presents an outline of a possible model for predicting results in Divisions where no counts have been recorded and hence the final outcome of the election. Substantial resources in the form of computing and statistical expertise are required for the implementation and testing of such a model.

Appendix I

Table I (a) Estimates of variance of ALP swing calculated using unmatched and matched booths methods

DIVISION

	BOWMAN		HERBERT		DAWSON	
M	34		41		64	
m	Var(Z)	Var(Z')	Var(Z)	Var(Z')	Var(Z)	Var(Z')
5	10.15	1.94	10.06	1.26	26.50	2.87
10	4.20	0.80	4.33	0.54	12.12	1.32
15	2.22	0.42	2.42	0.30	7.33	0.80
20	1.22	0.23	1.47	0.18	4.93	0.54
25	0.63	0.12	0.89	0.11	3.50	0.38
30	0.23	0.04	0.51	0.06	2.54	0.28
35			0.24	0.03	1.86	0.20
40			0.03	0.004	1.35	0.15
45					0.95	0.10
50					0.63	0.07
55					0.37	0.04
60					0.15	0.01

Table I (b) Estimates of variance of Liberal party swing using unmatched and matched booths methods

DIVISION

	BOWMAN		HERBERT		DAWSON	
M	34		41		64	
m	$\mathrm{Var}(Z)$	$\mathrm{Var}(Z')$	$\mathrm{Var}(Z)$	$\mathrm{Var}(Z')$	$\mathrm{Var}(Z)$	$\mathrm{Var}(Z')$
5	1.04	5.16	3.73	5.22	27.06	2.60
10	0.43	2.14	1.60	2.25	12.38	1.19
15	0.23	1.12	0.90	1.26	7.49	0.72
20	0.13	0.62	0.54	0.76	5.04	0.48
25	0.06	0.32	0.33	0.46	3.58	0.34
30	0.02	0.12	0.19	0.27	2.60	0.25
35			0.09	0.12	1.90	0.18
40			0.01	0.02	1.38	0.13
45					0.97	0.09
50					0.64	0.06
55					0.38	0.04
60					0.15	0.01

Note: Liberal swing in seats contested by both the Liberal and National Parties is not a useful measure. This is reflected by the variance estimates and, in particular $\mathrm{Var}(Z')$ for Dawson is larger than $\mathrm{Var}(Z)$. Maybe coalition swing (Liberal plus National Party) would provide a better measure.