**Political Analysis Advance Access published May 17, 2007**

doi:10.1093/pan/mpm010

# A Robust Transformation Procedure for Interpreting Political Text

**Lanny W. Martin**
*Department of Political Science, Rice University,*
*PO Box 1892, MS 24, Houston TX 77251-1892*
*e-mail: lmartin@rice.edu*

**Georg Vanberg**
*Department of Political Science, University of North Carolina,*
*Chapel Hill, NC 27599-3265*
*e-mail: gvanberg@unc.edu (corresponding author)*

In a recent article in the *American Political Science Review*, Laver, Benoit, and Garry (2003, "Extracting policy positions from political texts using words as data," 97:311–331) propose a new method for conducting content analysis. Their *Wordscores* approach, by automating text-coding procedures, represents an advance in content analysis that will potentially have a large long-term impact on research across the discipline. To allow substantive interpretation, the scores produced by the *Wordscores* procedure require transformation. In this note, we address several shortcomings in the transformation procedure introduced in the original program. We demonstrate that the original transformation distorts the metric on which content scores are placed—hindering the ability of scholars to make meaningful comparisons across texts—and that it is very sensitive to the texts that are scored—opening up the possibility that researchers may generate, inadvertently or not, results that depend on the texts they choose to include in their analyses. We propose a transformation procedure that solves these problems.

## 1 Introduction

In a recent article, Laver, Benoit, and Garry (2003) (hereafter, LBG) propose a new method for conducting content analysis. The thrust of their proposal is to estimate the ideological position expressed in a text by treating the individual words in that text as "data" to be scored rather than as words to be understood. This completely automated approach, implemented in their program *Wordscores*, represents a potentially fundamental advance in how to conduct content analysis. In its current form, however, the LBG approach has a significant limitation. The *Wordscores* procedure uses a set of "reference" texts to generate a dictionary of words with ideological scores and then uses this dictionary to generate two important ideological measures for a "virgin" text: a *raw score* and a *transformed score*. The transformed score is central because it permits a substantive interpretation of the results.

*Authors' note:* We would like to thank Ken Benoit, Michael Laver, three anonymous referees, and the editor for comments on earlier versions of this article.

Unfortunately, the particular transformation offered by LBG suffers from several weaknesses. First, the transformed scores are not robust to the set of texts that are scored. This particular problem opens up the danger that researchers may generate, inadvertently or not, results that depend on the set of virgin texts they choose to include in their analysis. Second, the LBG transformation places scores on different metrics, thereby not allowing researchers to make direct comparisons across all the texts in the analysis. We propose a transformation procedure that resolves both these problems.[1] This transformation has been incorporated into the *Wordscores* software and is available as an alternative scoring method.[2]

## 2 Interpreting Virgin Text Scores

The essence of the LBG approach is to use texts with exogenously defined ideological positions (the reference texts) to estimate the ideological positions of texts whose positions are unknown (the virgin texts). To do so, the LBG procedure creates a "dictionary" that assigns each word that appears in the reference texts an ideological score. This score is an average of the exogenous ideological scores assigned to the reference texts, weighted by the relative frequency of the word across the reference texts. The ideological position of any virgin text is then estimated as the frequency-weighted average position of the dictionary words appearing in the virgin text. The raw scores assigned by the LBG procedure are thus uniquely determined by the exogenous scores assigned to the reference texts and the relative frequencies of words across the texts. Importantly, we can apply this scoring procedure to the original reference texts used to create the dictionary (although the current LBG procedure does not do so).

The single most important issue for scholars interested in applications of the *Wordscores* procedure concerns the substantive interpretation of the results. As LBG point out (2003, 316), this is not straightforward. Many words are shared across reference texts, and these words receive a centrist score. The presence of these overlapping words pulls raw scores toward the interior of the interval defined by the reference scores. Because raw scores are dispersed on a much smaller scale, they cannot be directly compared to the exogenous scores attached to the reference texts. Moreover, the "bunching" of raw scores places them on an unintuitive metric that makes interpretation difficult. The following example from LBG (2003) illustrates this. Suppose we use British party manifestos from the 1992 general election, along with exogenous scores (on a 20-point scale) on an "economics" dimension for these manifestos derived from an expert survey (Laver and Hunt 1992), to estimate the positions of British party manifestos during the 1997 election. Following LBG, we give the Labour manifesto in 1992 an expert-assigned reference score of 5.35, the Liberal Democrat manifesto a score of 8.21, and the Conservative manifesto a score of 17.21. Using these texts to score the 1997 election manifestos, *Wordscores* yields raw scores of 10.3954 for Labour, 10.2181 for the Liberal Democrats, and 10.7361 for the Conservatives. Although these raw scores convey some information (e.g., the relative ordering of parties on the left-right dimension), they are clearly hard to interpret, especially with the original 20-point expert scale as the point of reference.

[1] Other authors have recently offered critiques of *Wordscores* on more fundamental grounds, such as its sensitivity to the set of reference texts used to construct the dictionary and to assumptions about the dimensionality of the policy space (e.g., Monroe and Maeda 2004). We are taking no position on these issues here. Rather, our argument is simply that *if* researchers choose to use *Wordscores* to conduct content analysis, then they should use a transformation procedure that is insensitive to the selection of virgin texts and that allows for meaningful comparison of scores across texts.

[2] We thank Ken Benoit for including our proposed transformation in the latest version of *Wordscores*, available from the *Wordscores* Web site. For details and commands, see the documentation provided.

To address this problem, LBG propose transforming raw scores in a way that will enable substantive interpretation. In particular, they aim to place scores on the same metric as the original reference scores, thus allowing scholars to "compare the virgin scores directly with the reference scores" (2003, 316). Their transformation attempts to do so by providing raw scores with the same dispersion as the reference text scores. The transformation centers raw scores around their mean and adjusts the variance of the scores to correspond to the variance of the reference texts. The LBG-transformed score, $P_t^*$, of text $t$ is given by:

$$P_t^* = (P_t - \bar{P}_{\mathrm{v}})\left(\frac{\mathrm{SD}_{\mathrm{r}}}{\mathrm{SD}_{\mathrm{v}}}\right) + \bar{P}_{\mathrm{v}}, \tag{1}$$

where $P_t$ is the raw score assigned to text $t$, $\bar{P}_{\mathrm{v}}$ is the average raw score of the virgin texts and $\mathrm{SD}_{\mathrm{r}}$ and $\mathrm{SD}_{\mathrm{v}}$ are the standard deviations of the reference and virgin text scores, respectively. Applying this transformation to the raw scores in our example, we obtain scores of 9.11 for Labour, 5.00 for the Liberal Democrats, and 17.17 for the Conservatives.

In what follows, we propose an alternative transformation. To understand the advantages of our proposal, it is necessary to discuss several shortcomings of the LBG transformation. The first is that *the transformed scores of virgin texts depend on the particular combination of virgin texts scored*. The second—and more significant—problem is that, contrary to the explicit purpose of transforming scores, the LBG transformation *fails to place scores on the same metric as the original reference scores*. As a result, transformed scores do not allow for meaningful comparisons across reference and virgin texts.

### 2.1 *The LBG-Transformed Scores Depend on the Combination of Virgin Texts Scored*

Suppose that instead of scoring all three 1997 manifestos in the illustration above, one were to vary the set of virgin texts. For example, consider a researcher who might be interested in figuring out whether (or how) the main two parties competing on the center-left—Labour and the Liberal Democrats—ideologically "repositioned" themselves *relative to one another* following the narrow electoral defeat suffered by Labour in 1992. Given his/her question (and certain assumptions about how parties compete), this researcher might feel quite comfortable excluding the 1997 Conservative party manifesto from his/her analysis. Unfortunately, if he/she were to do so using the LBG transformation, he/she would produce radically different ideological positions from those obtained above. Specifically, the LBG transformation in this scenario would assign a score of 5.91 to the Liberal Democratic manifesto and a score of 14.66 to the Labour party. In other words, Labour would now look as though it has moved quite far to the right—in fact, much closer to the LBG-estimated position for the 1997 Conservatives than to the Liberal Democrats. Clearly, this paints a dramatically different picture of British politics for this period than most experts would accept. Other combinations of virgin texts would yield still other scores.

It is easy to see why transformed scores under the LBG procedure depend so heavily on the specific set of texts scored. To adjust the dispersion of the raw scores, the transformation relies on the standard deviation of the virgin text raw scores, and the standard deviation, of course, depends on the particular set of virgin texts that are analyzed. Put simply, the LBG-transformed scores are inherently nonrobust to the selection of virgin texts. In some circumstances, the set of virgin texts may be defined in a natural way, for

**Table 1** British party scores

| | *Expert-assigned reference score* | *Raw score* | *Relative distance ratio* | *LBG transformation* | *MV transformation* |
|---|---|---|---|---|---|
| Liberal democrats 1992 | 8.21 | 9.98 | 0.26 | n/a | 8.5 |
| Liberal democrats 1997 | n/a | 10.22 | 0.40 | 5.00 | 10.11 |
| Labour 1992 | 5.35 | 9.51 | 0.00 | n/a | 5.35 |
| Labour 1997 | n/a | 10.40 | 0.50 | 9.17 | 11.31 |
| Conservatives 1992 | 17.21 | 11.28 | 1.00 | n/a | 17.21 |
| Conservatives 1997 | n/a | 10.74 | 0.69 | 17.18 | 13.59 |

example, the manifestos for all the parties running in an election campaign (although all these are rarely available to scholars, even in the British elections examined by LBG). In many other applications, however, it will *not* be obvious what the appropriate set of virgin texts to score is. For example, for someone interested in analyzing the content of parliamentary speeches, newspaper stories, or judicial opinions, it may not be apparent which subset of cases to include in the analysis when there are perhaps tens of thousands of possibilities. Because of the sensitivity of scores to the set of texts, the choice of which texts to include or exclude could—inadvertently or not—have significant effects on the position attached to any particular text. To increase confidence in an analysis of transformed scores, it is essential to develop a transformation that makes scores *independent* of the particular set of texts scored.

### 2.2 *The LBG Transformation Fails to Place the Virgin Texts on the Same Metric as the Reference Texts*

The primary purpose of transforming raw scores is to allow scholars to "compare the virgin scores directly with the reference scores" (LBG 2003, 316). For example, after comparing the 1992 British reference scores to the 1997 transformed scores, LBG point out the move of Labour toward a more "centrist" position (2003, 320). To make such comparisons valid, it is clearly necessary that the transformed scores be placed on the same metric as the exogenously assigned reference text scores. Unfortunately, as we shall now show, the LBG scores are *not* on the same metric. Consequently, the comparison of LBG-transformed scores to the reference scores has the potential to result in seriously misleading conclusions. Continuing with the British case, suppose we again take the three 1992 manifestos as the reference texts to construct our word dictionary. We then use *Wordscores* to obtain raw scores for the 1997 virgin manifestos. In addition, we also obtain raw scores for the reference texts (the 1992 manifestos) by scoring the reference texts using the same dictionary used to score the virgin texts. Because they are all generated by a single dictionary, these scores tell us how the word usage across these texts differs as evaluated by the same dictionary. As a result, we can directly compare raw scores across the virgin and reference texts. We present these scores in Table 1.

As already noted, raw scores are difficult to interpret in themselves, although they do convey some information directly, such as the relative ordering of parties in a policy space. One useful way to distill the positional information contained in these scores in a more intuitive way (but without transforming them to correspond to an exogenously defined scale) is to consider their *relative distance ratios*. To calculate the relative distance ratios

for the raw scores, we choose two texts as "anchors" and express the placement of all other texts in relation to this "standard unit":

$$\text{relative distance ratio for text } i = \frac{P_i - P_1}{\mid P_1 - P_2 \mid}. \tag{2}$$

Substituting the 1992 Labour manifesto for $P_1$ and the 1992 Conservative manifesto for $P_2$, we see that the relative distance ratios for the 1997 manifestos reveal several interesting features. For example, the 1997 Labour manifesto is placed halfway between the 1992 Labour and Conservative manifestos, indicating a clear ideological shift to the center by the Labour party. Relative to the distance between 1992 Labour and Conservatives, the Liberal Democrats are placed at 0.264 in 1992 and at 0.400 in 1997, indicating that the Liberal Democrats also shifted toward the center between 1992 and 1997. However, although they were positioned to the right of Labour in 1992, this relative position is reversed in 1997. Thus, the relative distance ratios suggest that both Labour and the Liberal Democrats moderated their position, but that Labour moved more radically toward the center, reversing the relative position of the two parties. Finally, the Conservative manifesto in 1997 is placed at 0.693, indicating that the Conservatives also moderated their position between the two elections, although they remained the rightmost party in 1997.

We now contrast these results with the conclusions we would draw by applying the LBG procedure of comparing their transformed virgin text scores to the exogenously assigned reference text scores. Recall that the LBG procedure transforms only the 1997 virgin raw scores, and *assumes* that these transformed scores are comparable to the exogenously assigned reference scores for 1992. Their transformed score for Labour 1997 is 9.17, compared to the expert-assigned reference score for Labour 1992 of 5.35, indicating Labour's shift to the center, just as the relative distance ratios do. However, the procedure *understates* the magnitude of the shift, placing Labour 1997 about one-third of the distance between the reference scores for Labour and Conservatives in 1992 (instead of the midway point between these scores as indicated by the raw scores derived from actual word usage). For the Conservatives, the LBG transformation suggests hardly any ideological movement at all: a transformed score of 17.18 in 1997 compared to an exogenous score of 17.21 in 1992. The word usage as reflected in the raw scores, however, indicates a substantial movement of the Conservatives toward the center that the LBG transformation does not reflect. Most disturbing of all are the conclusions we would draw about the Liberal Democrats. Applying the LBG procedure and comparing their transformed score to the reference text score suggests a clear move by the Liberal Democrats toward the *left*, from 8.21 in 1992 to 5.00 in 1997. In contrast, the raw scores and relative distance ratios—which are derived from the word usage in both texts as judged by the same dictionary—clearly indicate a move to the *right*.

The conclusion from this exercise is that in order to draw meaningful comparisons across virgin and reference texts, we must have scores for both that are derived from the same underlying scoring procedure. The raw scores (and their relative distance ratios) provide such a common metric. As we have demonstrated above, however, the LBG procedure of comparing transformed virgin text scores to the exogenous reference scores does not accurately recover the relative distance ratios generated by the raw scores. In other words, LBG-transformed virgin text scores and the original reference text scores are on different metrics. Thus, scholars who compare LBG-transformed virgin text scores to the exogenous reference scores do so at their own peril. Doing so can lead to misleading

conclusions about relative ideological positions and even the *direction* of ideological movement as expressed in the texts being analyzed.

## 3 A Robust Transformation Procedure

Scholars wishing to avoid these difficulties may be tempted not to transform scores at all, instead focusing their analysis on the raw scores. The difficulty in this strategy is that raw scores are simply too unintuitive to make comparisons across them meaningful (as an example, consider the raw scores listed in column 3 of Table 1). Some kind of transformation that allows a more intuitive grasp of the information is necessary. Our relative distance ratios represent one possible transformation that faithfully reflects the information contained in the raw scores as a function of the distance between two anchor texts. Moreover, this minimal transformation also avoids the first difficulty confronted by the LBG approach. Each text's relative position is determined uniquely by the words it uses as scored by the dictionary; consequently, varying the set of newly scored texts has no impact on relative placement.

Although relative distance ratios provide all the information necessary for comparisons across texts, it is sometimes desirable to engage in more complicated transformations. In assigning the exogenous reference scores, scholars are typically making use of a scale that has some intuitive meaning (e.g., 10-point ideological scales, ADA scores, etc.). Whenever such an exogenous scale is being used, it is natural to want to place newly scored texts on the same scale. It is precisely the desire to place scores on the same metric as the exogenous reference scores that motivates the LBG transformation. But is it possible to construct a transformation that achieves this purpose—placing scores on the same scale as the exogenous reference scores while faithfully preserving the information contained in the raw text scores?

In what follows, we offer a transformation that can do so (with one caveat). The intuition behind our transformation is straightforward: We take the raw scores and stretch them to match the original scale of the reference scores. Given raw score $P_t$, an assigned score $A_r$ for reference text $r$, and reference texts $R1$ and $R2$ (where $R1$ is located to the left of $R2$), the transformed score $\hat{P}_t$ is given by:

$$\hat{P}_t = \left\{ (P_t - P_{R1}) \frac{A_{R2} - A_{R1}}{P_{R2} - P_{R1}} \right\} + A_{R1}. \tag{3}$$

To demonstrate the properties of this transformation, we begin with the case in which only two reference texts are used to create the dictionary. In this case, inspection of equation (3) reveals immediately that the newly transformed score of reference text $R1$ will be equal to the original reference score assigned to the text ($A_{R1}$). Similarly, the transformed score of reference text $R2$ is given by its exogenously assigned reference score. Thus, our transformation recovers the exogenously assigned reference text scores, and because it is an affine transformation, it also places any virgin text in the proper relative position vis-à-vis these texts as determined by the raw score relative distance ratios. These two features ensure that all scores are now on the same metric as the exogenous reference scale, enabling direct comparison.

With two exogenous reference texts, the transformation has all the properties we would want a transformation to possess. It produces scores that do not depend on the set of virgin texts, thus eliminating the sensitivity of scores to the choice of virgin texts. Second, our

transformed scores correctly reflect the relative positions of texts as indicated by the relative distance ratios. Third, the transformation recovers the exogenously assigned reference scores exactly. Importantly, however, relying on only two reference texts has a cost. Since only two texts can be used as inputs, the information that is used to construct the word dictionary is limited. Sometimes, scholars may wish to make use of more than two reference texts in order to generate a more finely grained dictionary. If three or more reference texts are used, our transformation retains its two most important properties: Scores remain independent of the set of virgin texts scored, and the relative placement of transformed scores corresponds to the relative distance ratios. However, as soon as we make use of more than *two* reference texts as "inputs," it is no longer possible to recover the original exogenous scores of *all* reference texts exactly. Instead, scholars must choose two reference texts as the anchor points that will be used to stretch the raw scores onto the original metric. Typically, it will be sensible to use the two extreme reference texts as anchors. Denoting the largest and smallest exogenous reference scores by $A_{\max}$ and $A_{\min}$, respectively, the transformed score for text $t$ is:

$$\hat{P_t} = \left\{ (P_t - P_{\min}) \frac{A_{\max} - A_{\min}}{P_{\max} - P_{\min}} \right\} + A_{\min}. \tag{4}$$

This transformation recovers the exogenous scores assigned to the anchor texts. All other reference texts, like the virgin texts, are placed on the same scale at the appropriate relative distance from $A_{\max}$ and $A_{\min}$.[3] However, the transformed scores of these reference texts will (usually) deviate from their exogenous scores. Importantly, this is true of *any* transformation that aims to retain the proper relative distance ratios. In general, (1) stretching the raw scores to the original metric while (2) preserving the proper relative distance between scores is only possible with two reference scores. As soon as three or more scores are involved, the fact that the relative distances of raw scores generally do not correspond to the relative distance of reference scores makes a transformation that achieves both (1) and (2) impossible. In other words, scholars who use more than two reference texts face a trade-off. Any increased accuracy in the word dictionary that is gained by adding reference texts must be purchased at the expense of some degree of internal consistency.[4]

Although our transformation will always retain the right relative distance ratios of all scores, it can no longer reproduce the exogenous scores of all reference texts. In a certain sense, this is unwelcome. After all, the dictionary was constructed on the assumption that the reference texts have a certain relation to one another. In another sense, these deviations between exogenous reference scores and the transformed scores assigned to the reference texts provide valuable information. They indicate how well the exogenously *assumed* relative distances between the positions of the reference texts are reflected in the *actual* word usage of the texts. Given this, these deviations can be used as a partial check on the validity of the expert judgments used to assign ideological positions to the original reference texts.

[3]The last column of Table 1 lists our transformed scores for the British example. Note that these scores, mirroring the relative distance ratios, pick up the move by all parties toward the center in 1997 as well as the switch in relative position by Labour and the Liberal Democrats. Interestingly, although the changes in our transformed scores across elections are inconsistent with the LBG scores (which, again, suggest a movement of the Liberal Democrats to the left), they are consistent with party left-right scores from the independent hand coding of these manifestos by the Comparative Manifestos Project (Budge et al. 2001).

[4]One way to evaluate the costs of this trade-off is to calculate the percentage deviation of each reference text's transformed score from its exogenously assigned score.

More importantly, our transformation retains its key advantages over the LBG approach. It produces scores that are not sensitive to the set of virgin texts analyzed and that accurately reflect the ideological positions of the texts as indicated by their word usage.

## References

Budge, Ian, Hans-Dieter Klingemann, Andrea Volkens, Judith Bara, and Eric Tanenbaum. 2001. *Mapping policy preferences: Estimates for parties, electors, and governments 1945–1998.* Oxford: Oxford University Press.

Laver, Michael, Kenneth Benoit, and John Garry. 2003. Extracting policy positions from political texts using words as data. *American Political Science Review* 97:311–31.

Laver, Michael, and W. Ben Hunt. 1992. *Policy and party competition.* New York: Routledge.

Monroe, Burt L., and Ko Maeda. 2004. Talk's cheap: Text-based estimation of rhetorical ideal-points. Working paper. Michigan State University.