Advance Access publication August 13, 2007

Political Analysis (2008) 16:101–111
doi:10.1093/pan/mpm020

Compared to What? A Comment on "A Robust Transformation Procedure for Interpreting Political Text" by Martin and Vanberg

Kenneth Benoit
Department of Political Science, Trinity College, Dublin 2, Ireland
e-mail: kbenoit@tcd.ie (corresponding author)

Michael Laver
Wilf Family Department of Politics, New York University,
726 Broadway, Room 756, New York, NY 10003-9580
e-mail: michael.laver@nyu.edu

1 Introduction

In "A Robust Transformation Procedure," Martin and Vanberg (2007, hereafter MV) propose a new method for rescaling the raw virgin text scores produced by the "Wordscores" procedure of Laver, Benoit, and Garry (2003, hereafter LBG). Their alternative method addresses two deficiencies they argue exist with the transformation of virgin text scores proposed by LBG: First, that the LBG transformation is sensitive to the selection of virgin texts, and second, that it distorts the reference metric by failing to recover the original reference scores when reference texts are scored and transformed as if they were virgin texts. Their proposed alternative is "robust" in the sense that it avoids both shortcomings. Not only is MV's transformation a welcome contribution to the Wordscores project but also the critical analysis on which it is based brings to light a number of assumptions and choices that face the analyst seeking to estimate actors' policy positions using statistical analyses of the texts they generate. When first describing the possibility of rescaling the raw virgin text estimates, we emphasized that our

> particular approach to rescaling is not fundamental to our word-scoring technique but, rather, is a matter of substantive research design unrelated to the validity of the raw virgin text scores... Other transformations are of course possible. (LBG, 316)

To explore more fully into the assumptions and choices behind alternative transformations and the research designs which motivate them, we offer the following comments.

2 The Need to Rescale "Raw" Virgin Text Scores

The issue of rescaling or "transforming" the raw virgin text estimates produced by Wordscores arises because overlapping words cause the untransformed (or "raw") virgin text scores tend to cluster around the mean of the reference values. The more words are shared

Authors' note: We thank Georg Vanberg and Lanny Martin for comments and discussions during the drafting of this Comment.

between reference texts, the more virgin text scores will experience this bunching around the reference values' mean. To take a simple example, consider the common "and," "but," and "the" articulators found in every natural text. Whereas other words might discriminate between texts—as a matter of empirical fact, the word "drugs" may be more commonly found in right-wing texts and "contributions" more in left-wing texts—common text articulators tend not to contain positional information since they are generally found with equal relative frequencies in all texts. As a consequence, these noninformative words tend to be scored at the reference text means. Virgin texts that contain these words will have these nondiscriminating word scores averaged into their overall text score, and hence these virgin text scores will also be drawn to the reference score means. In natural language texts, furthermore, not only are nondiscriminating words the most common but also these words tend to exist in fixed proportions in natural language texts, meaning that longer texts will tend to overlap more through common words. The same longer reference texts providing additional information that Wordscores uses for distinguishing positions, therefore, also add additional noninformative content whose word scores will lie at the center of the reference values.

When virgin texts also contain the noninformative words, as all real-world texts will, then Wordscores will produce raw text scores that are considerably more bunched than the original reference score values. In the LBG example of British party manifestos from 1997, for instance, the original reference range of 5.35–17.21 is reduced to a raw virgin score range of 10.2181–10.7361. For Wordscores users wanting to compare the raw scores directly to the reference metric—although this is by no means always necessary—the bunching of raw virgin text scores may leave them feeling somewhat cheated by the "simply too unintuitive" (MV) numeric values that are produced as outputs.

In part, the obvious difference in metrics arises from the relative transparency of the Wordscores technique. By contrast, many other widely used scaling techniques used to estimate dimensional positions—such as factor analysis methods (Gabel and Huber 2000), NOMINATE (Poole and Rosenthal 1997), Bayesian methods based on item-response theory (e.g., Clinton, Jackman, and Rivers 2004), multinomial or Poisson-based methods (e.g., Monroe and Maeda 2004; Slapin and Proksch 2007) for estimating positions from text—produce estimates whose values are not only different from any sort of reference metric but also possibly uniquely different from *any* other metric. Nonetheless, Wordscores purports to extract dimensional information that is (in LBG's terminology) fundamentally *a priori* in nature, meaning that both the dimensions and their metrics are defined in advance, and the reference values represent known positions on these dimensions using the known metric. Since the reference dimension and its metric are known in advance, this suggests that we should be able to compare input and output scores directly. This need arose for LBG, e.g., because their chief objective was to validate what was then a new method by comparing Wordscores results directly to independent, external estimates of the same quantities scaled on the reference metric. This objective motivated LBG's transformation procedure, since it allowed a direct comparison of externally obtained expert score estimates for the British parties in 1997 (on a 1- to 20-point scale) to the Wordscores estimates based on reference values set from the same 1- to 20-point scale from a 1992 expert survey.

The key question in this context is whether the Wordscores scaling issue constitutes a "significant limitation" of the technique itself or merely an impediment to intuition and interpretation in particular situations. From our viewpoint, it is understating the case to claim that raw scores cannot be compared in ways that are "meaningful" (MV). Indeed, the problem lies not in comparing virgin text scores to one another but rather in comparing them to reference values. As we shall argue, furthermore, there is a difference between comparing virgin text scores to reference text *values* assigned by the researcher and to reference text *scores* when these texts are treated as virgin documents. How and whether to transform raw text scores into quantities that better serve a researcher's intuition involves a substantive decision that should be motivated by the researcher's need for a particular type of comparison (as well by the researcher's subjective sense of what constitutes an intuition). In what follows, we explore the issue of scaling and comparisons of raw text scores in order to better illuminate the advantages and limitations of both the MV and the LBG transformations.

3 Solutions to the Text Score Bunching

3.1 *Solution 1: Use the Raw Scores*

The most direct way to use Wordscores output is to interpret the virgin text scores directly since these scores also contain substantive information on an interval scale. The set of raw scores, after all, contains the fundamental input into *any* transformation, which neither can nor should, in a strict sense, generate any *new* information. The level of information about relative party positions in the set of raw virgin text scores is at least high as that in the set of numbers produced by any transformation of these. If we wish to compare estimated virgin text positions to reference texts, furthermore, we can simply score reference texts too as if they were "virgin" texts. The resulting raw estimates are robust, in the MV sense of being the same regardless of the set of virgin texts chosen. *Our recommended policy is thus to use untransformed virgin text scores whenever this is feasible*. If the issue is simply that users get eyestrain by being forced to peer behind the decimal point at small (but statistically significant) differences between numbers, nothing is lost, e.g., by converting dollars into cents and simply multiplying everything by 100. We transform the raw scores because we want to compare virgin text scores with some *external* referent. How we transform, therefore, depends upon what that referent is.

One common solution to the problem of putting two different sets of numbers on a comparable metric is to standardize both sets. It is certainly possible to standardize a set of virgin text scores and indeed to include in this set the raw scores from reference texts, treated as virgin texts. In fact, as we shall see, the LBG transformation is a particular type of standardization procedure, adapted to the Wordscores environment. The problem with any standardization procedure, of course, is that the same input scores map into different output scores depending on the particular set of inputs being scored. Add new raw scores and the standardized scores of existing items are liable to change. This, indeed, is one of the main shortcomings of the LBG transformation, highlighted by MV.

If for reasons of research design, virgin text scores must be compared directly to reference values, then raw or standardized scores will not be enough to overcome the score-bunching problem. In this case, a more interventionist procedure will be required that transforms raw scores onto a metric more closely resembling that of the original reference scores. Note that the problem is not with the *center* of the raw text scores: This is approximately where it should be, drawn to the reference score mean but with each text differing from this mean according to the discriminating words found in the reference texts. The problem rather has to do with the *dispersion* of the raw text scores: They are "too bunched," and we would like them rescaled to look more like the reference text values that served as inputs. As we will show, both the LBG and the MV methods use linear rescaling to accomplish this expansion of the raw scores. LBG's transformation rescales

Table 1 British party scores reproduced

Texts	*Raw score*	*RDR*	*MV transformed scores*	*LBG transformed scores*	*1997 Expert survey*
Labour 1992 (5.35)	9.51	0.00	5.35		5.35
LD 1992 (8.21)	9.98	0.26	8.50		8.21
Conservatives 1992 (17.21)	11.28	1.00	11.31		17.21
LD 1997	10.22	0.40	10.11	5.00	5.77
Labour 1997	10.40	0.50	11.31	9.17	10.30
Conservatives 1997	10.74	0.69	13.59	17.18	15.05

the raw virgin text scores to match the variance of the reference values, whereas MV's transformation rescales raw scores based on absolute difference between two extreme reference texts, in order to fix the transformed virgin text scores of the reference texts at their original reference values.

3.2 *Solution 2: Relative Distance Ratios*

MV propose a new way to interpret raw scores based on their positions relative to the raw scores for two anchoring texts, P_1 and P_2, which they suggest should be those with the most extreme reference values. When virgin texts' positions relative to P_1 are divided by the distance $|P_1 - P_2|$, the resulting *relative distance ratios* (RDRs) "provide all the information necessary for comparisons across texts," (MV) including comparisons of virgin texts to reference texts. Hence, in MV's Table 1, they are able to show using RDRs that the Liberal Democrats (LDs) in 1997 moved to the *right* compared to their 1992 manifesto. The LBG transformation (reported in LBG 2003) of the raw scores, by contrast, suggests that in 1997 the LDs moved to the left. We reproduce this finding in Table 1. The conclusion that MV draw from the raw scores, the RDRs, and their own transformation is that the LDs moved to the right since this party's RDR shifts from 0.26 to 0.40. Does this represent a "real" shift to the right, indicating that the conclusions drawn by the LBG transformation are wrong? Your answer depends on your point of reference when measuring a shift to the right.

MV would argue that our advice of using the reference texts directly implies that the LDs moved to the right in 1997, on the basis that the raw score of this party's 1997 manifesto (10.22) is to the right of the raw score of its 1992 manifesto (9.98) which also served as a reference text. The RDR and the MV transformation, furthermore, preserve this ordering, whereas the LBG transformation does not, leading to two very different conclusions about the movement of this party between elections. Which conclusion is correct?

Once again, the answer depends on your point of reference, which is why it is very important to understand what is being compared in the "relative" part of these distance ratios. The RDRs take as their benchmarks the raw scores of the two most extreme reference texts. Because of overlapping words from the reference texts, these reference text raw scores are highly bunched, as clearly seen in Table 1. The center of this bunching is the mean of the reference values (10.26), forming the general center of gravity toward which the scores of overlapping words in the reference texts are drawn. If all reference texts used the same words with the same relative frequencies, then all words would have scores of 10.26, and hence all virgin texts would also have raw scores of 10.26. If we compare the LD 1997 text score to this value, instead of to the LD 1997 manifesto's raw score when treated as a virgin text, then the LDs did indeed move to the left in 1997, Labour moved to the right, and the Conservatives stayed on the right. In a nutshell, this is the difference between the MV and the LBG transformations and highlights the substantive decision as to the target of comparison, to be guided by research design. In many circumstances, we shall demonstrate, our comparisons of substantive interest will be better served by transforming based on the reference value *metric* rather than the relative distance of the raw *scores* of the reference texts as per the RDR and also, as we shall see shortly, the MV transformation.

3.3 *Solution 3: The LBG Transformation*

LBG propose a transformation based on expanding the bunched raw text scores to have the same SD as the reference texts. This is accomplished by first normalizing the raw virgin text scores, zeroing the mean, and standardizing their variance to 1.0. The normalized scores are then multiplied by the SD of the reference scores to give them the same variance and then recentered to the original raw virgin text score mean by adding back this value (see MV, equation 1). This transformation can be interpreted as a linear rescaling of the raw scores. The $(P_t - \overline{P}_v)/\mathrm{SD}_v$ is the normalization, the SD_r the coefficient, and the $\overline{P}_v$ the additive component or "intercept" that restores the mean of the virgin text scores. Although the LBG transformation is completely independent of the RDRs that come from scoring the reference documents as virgin texts, it will depend on the overall distribution of raw scores, meaning it is sensitive to the set of texts chosen for transformation. Unlike the raw scores that are always invariant to the selection of virgin texts, as MV correctly point out, the values of the LBG transformed scores will change when virgin texts are added or dropped, especially when the set of texts was small (less than 10) to begin with.

3.4 *Solution 4: The MV Transformation*

MV propose an alternative rescaling procedure based on the difference between the reference values and the raw text scores of reference texts when scored as virgin texts. Like LBG's transformation, MV's rescaling is also linear. The transformation, analogous to the LBG normalization, comes from $(P_t - P_1)/(P_2 - P_1)$, which is the difference of the raw text score for text t relative to the difference between the raw scores of the two reference texts when treated as virgin texts. The rescaling coefficient is the difference between the reference scores $(A_2 - A_1)$ to which the lower reference score is added back to recenter the transformed scores.

Because the MV method does not rely on any virgin text-dependent quantities (such as SD_v), it is invariant to the set of virgin texts chosen, always producing the same transformations whether a single or any set of virgin texts is being rescaled. When $t = 1$ or $t = 2$, furthermore, the transformation easily produces A_1 or A_2, respectively, recovering as transformed scores the original values assigned to the reference texts. A constraint of the MV rescaling method is that it is based on the raw scores of only two reference texts, but an advantage is that it can be applied even to a single virgin text.[1] The LBG rescaling method works only if at least two virgin texts are used since otherwise no SD can be computed for virgin texts.

For purposes of simplicity in what follows, we limit our discussion to the case involving two reference texts, which also restricts analysis to only one possible dimension. With a single dimension defined by only two reference points, the numerical values assigned to

[1] In the Stata wordscores.pkg that now includes the MV transformation as an option, the transformation will automatically use (only) the two most extreme reference texts R_1 and R_n, where $A_1 < \ldots < A_\mathrm{n}$.

reference texts become essentially arbitrary since any two pairs of scores are simply a linear rescaling of any other. Popular choices are therefore to use reference values $A_1 = 0$ and $A_2 = 1$. In this case, the MV transformation is equivalent to the RDR. This can be shown by reproducing MV's equation (4) and substituting

$$\begin{aligned}\hat{P}_t &= (P_t - P_{\min}) \frac{A_{\max} - A_{\min}}{P_{\max} - P_{\min}} + A_{\min} \\ &= (P_t - P_1) \frac{1}{P_2 - P_1} + 0 \\ &= \frac{P_t - P_1}{|P_1 - P_2|}, \qquad (1)\end{aligned}$$

where equation (1) is the RDR definition supplied by MV. The implication is that the MV transformation will inherit all the properties of the RDR. We explore the full consequences of this issue in the next section.

4 Two Types of Sensitivity

4.1 *Sensitivity to the Selection of Reference Texts*

As we have always been careful to emphasize, the most crucial research design issue when using Wordscores concerns the identification of reference texts. Is there a "natural" set of reference texts with well-known positions (e.g., party manifestos)? How many of these are there? Should we use all of these or only the most extreme—in the latter case throwing away "good" information? How should we proceed when we are less confident about our ability to identify reference texts and measure their positions? These questions are vital because Wordscores estimates, axiomatically, can be no better than the reference texts that anchor them. This feature that Wordscores estimates are sensitive to the selection of reference texts is simply a specific instance of the classic information problem of "garbage in, garbage out"—a methodological issue we regard as far deeper than that of rescaling.

The issue is directly relevant here because the estimated relative distance between two (extreme) reference values P_1 and P_2, as recommended by MV, is determined by the degree to which R_1 and R_2 contain words in common. This can be clearly seen in the example in Table 2, which specifies two stylized pairs of reference texts and seven "virgin texts."[2] Each reference text pair is anchored at reference values of 0 and 1. The reference text pair $\{R_1, R_2\}$ contains a set of words identical to that in reference text pair $\{R_3, R_4\}$ except that the latter texts each contain an additional five "C" words.

The first thing we note is the difference in the raw scores between *reference* text pairs, scoring these as virgin texts. Regardless of the choice of reference texts in this example, all raw scores are the same. This applies both to the virgin texts and to the reference texts scored as if they were virgin texts. When we increase the proportion of noninformative content by adding five additional words "C" to each reference text with the same relative frequency (which adds no additional information that will change any of the word scores), this does not change any of the raw text scores for any documents scored as virgin texts. Note, however, from the center panel of the table, which uses R_1 and R_2 as reference texts, that the raw scores for R_1 and R_2, treated as virgin texts, are more distanced from one another (0.37 and 0.63) than those for R_3 and R_4 (0.43 and 0.57). In the denominator of

[2] This table is adapted from an example in an earlier draft of the MV paper.

Table 2 Text example with raw and transformed scores

	Reference texts: R_1, R_2			*Reference texts: R_3, R_4*		
Texts	*Raw score*	*MV transformed scores/RDR*	*LBG transformed scores*	*Raw score*	*MV transformed scores/RDR*	*LBG transformed scores*
R_1 (0) A A B C D	0.37	0.00		0.37	−0.50	
R_2 (1) A B B C E	0.63	1.00		0.63	1.50	
R_3 (0) A A B C D C C C C C	0.43	0.25		0.43	0.00	
R_4 (1) A B B C E C C C C C	0.57	0.75		0.57	1.00	
V_1 A C C D D	0.27	−0.38	−0.02	0.27	−1.25	−0.02
V_2 A A A A B C C	0.43	0.23	0.34	0.43	−0.04	0.34
V_3 A B D E	0.50	0.50	0.49	0.50	0.50	0.49
V_4 A B B C C E	0.61	0.92	0.74	0.61	1.33	0.74
V_5 A B B E E	0.73	1.38	1.00	0.73	2.25	1.00
V_6 D D D	0.00	−1.38	−0.60	0.00	−3.25	−0.60
V_7 E E E	1.00	2.38	1.59	1.00	4.25	1.59

MV's RDR, $|P_1 - P_2|$, the values are 0.26 and 0.14, respectively. Adding the noninformative content {C, C, C, C, C} made the reference texts appear almost twice as similar, if we use this benchmark.

The MV transformed scores—identical here to the RDRs when $A_1 = 0$ and $A_2 = 1$—also become much more extreme, for exactly the same virgin texts, when we change the reference text pair. Text V_1 appears to more than triple in extremity following the inclusion of noninformative content {C, C, C, C, C}, and other large changes can be observed for V_2, V_4, and V_5.[3] The extreme version of the problem is shown by the difference between the MV transformed scores of the most extreme texts, V_6 and V_7. As indicated by the raw scores, V_6 is full of pure zero-scored words "D" and V_7 full of pure 1.0-scored words "E". The effect of adding uninformative junk words "C" to the reference texts, however, is to stretch out the RDRs (identical to the MV transformed scores) at a rate inversely proportional to degree of overlap. If $P_1 = 0$ (as in the case of V_6), then

$$\mathrm{RDR}_i = P_i \frac{1}{|P_1 - P_2|} . \tag{2}$$

The greater the number of overlapping words between R_1 and R_2, the smaller the distance $|P_1 - P_2|$, and hence the greater the multiplier applied to the raw score. In Table 2, the raw scores for reference texts R_1 and R_2 are $P_1 = 0.367$ and $P_2 = 0.633$, but this narrows for R_3 and R_4 to just 0.43 and 0.57. At the extreme where both reference texts are identical, then the multiplier in equation (2) would be $1/0 = \infty$ suggesting an infinite relative distance.[4] Thus, although not so sensitive to the set of virgin texts, the MV transformation is very sensitive to the choice of reference texts and the degree of overlap they contain.

The key problem for making substantive conclusions about movement—as with the previous example of the LDs in 1997—can be seen by comparing the MV scores for V_2

[3] Only the middle-valued text V_3 is unchanged because it contains only words from each text that are held in equal proportions.

[4] Note that where R_1 and R_2 are two reference texts, then $P_1 + P_2 = A_2 + A_1$.

and V_4. Depending on the reference text pair, V_4 can be seen as being a tenth more left than its rightmost reference text (0.92) or a third more right (1.33). Likewise, V_2 also moves from being classified as being to the right of the leftmost reference text to being to its left when the set of reference texts changes. This situation directly parallels that of the LD manifesto in 1997 and explains why the LBG and MV transformations point to different conclusions regarding the movement of this party's position in 1997 relative to 1992. Compared to the reference text score when the LD 1997 manifesto is treated as a virgin text, it is 0.40 of the distance to the right of the distance between the Labour and Conservative texts. Compared to the mean of either the reference values (10.26) or of the raw scores of the virgin texts (10.45), however, the LD 1997 raw score is to the left.[5] Since the LBG transformation preserves this virgin text mean, it suggests that the LDs moved left in 1997, not right.

Which is correct? From a literal standpoint, it could be argued that the MV approach is correct since the unprejudiced incorporation of *all* word frequencies is a fundamental feature of Wordscores, meaning that the overlapping word "C" in Table 2 does in fact make R_3 and R_4 more similar and that this similarity should affect our interpretation of any virgin text scores based on these texts. The problem, in other words, is not a by-product of transforming the raw scores but rather derives from the fundamental operation of the Wordscores method that makes no distinctions between words according to whether they are "discriminating." Nonetheless, this argument leaves open the question of interpreting text scores of virgin documents when we want to assess their position relative to reference values: Movement, but compared to what? Do we wish to compare the virgin text scores relative to the distance between reference texts or according to some absolute referent defined by the reference values? The former method will make virgin text scores appear more centrist or more extreme depending on reference text length and the proportion of nondiscriminating content but is robust to the selection of virgin texts. The latter, by contrast, depends on a substantive modeling assumption that the spread of the virgin texts matches that of the reference texts and is sensitive to the selection of virgin texts but is robust to varying the level of nondiscriminating content in the reference texts.

It is the second approach that is taken by the LBG transformation based on the reference values rather than the reference texts' scores when treated as virgin texts. Our objective in developing this transformation was to allow direct comparison of the virgin texts, not to the reference texts per se, but to the reference values assigned to those texts. If the proof is in the pudding, then the LBG transformation certainly seems to work better in the British example from Table 1. Completely independent expert surveys from 1992 to 1997 strongly imply that the LDs did in fact move left between the two elections. These independent estimates seem to validate, if not almost perfectly match, the LBG transformed scores for the LDs. A separate word-scoring method based on preconstructed dictionaries, furthermore, also indicates that the LDs shifted to the left from 1992 to 1997 (Laver and Garry 2000, 630–31).

4.2 *Sensitivity to the Selection of Virgin Texts*

The strong assumption made by the LBG transformation is that the *dispersion* (or variance) of these scores will be the same among virgin texts as among reference texts. This is

[5] The movement of the virgin text scores mean to the right of the reference value comes from the fact that the Conservative manifestos in both 1992 and 1997 were far longer than the other parties' texts (see LBG 2003, 320). This "movement" reflects word overlap and text length, however, rather than a substantive overall shift to the right by all parties.

the key to understanding why the LBG transformation is sensitive to the selection of virgin texts but also represents a key element of the implicit model for comparison. The assumption of matching variance is also why with the LBG transformation, like should only be compared with like.

How concerned should we be that the LBG transformation depends on the selection of virgin texts?[6] It is understandably unsettling for researchers if their transformed scores change as they add and drop reference texts, although the same will happen with any standardization procedure, of which the LBG transformation is a special case. Take the example of the LBG transformed scores for V_5, for instance, whose raw scores indicate that it lies to the right of all the reference texts (regardless of which pair is chosen). The LBG transformed score for this text, however, places it at 1.00 rather than showing it to be more extreme than the reference texts. This occurs because the standardization is affected by the inclusion of the extreme right V_7 in the set of virgin texts. When only V_1 through V_5 are chosen as virgin texts, the LBG transformed score for V_5 is 1.41, very close to the MV transformed score when using R_1 and R_2 as reference texts.

Once again, the fundamental issue here relates to research design. If we know, e.g., we are comparing like with like—such as three parties' manifestos in adjacent elections—then the assumption of constant dispersion can be plausibly upheld. If on the other hand we have a small and/or biased sample of the population of virgin texts under investigation, then the LBG transformed scores will poorly reflect the target metric of the reference values. For instance, we would not want to apply the LBG transformation with three main parties' manifestos as reference texts—such as Labour, the Conservatives, and the LDs—but just Labour and two small right-wing parties' manifestos as virgin texts. Rather, the standardization inherent in LBG calls for selecting virgin texts that are more balanced vis-a-vis the reference text sample, which yields the type of successful results as shown in LBG (2003) which followed this balanced virgin text selection approach.

Absent a selection of balanced virgin texts, then the LBG transformed scores will tend to move around as texts are added or deleted from the set, just as in any transformation involving standardization. The strong warning labeled on the LBG transformation, therefore, is to begin by making strenuous efforts to identify and analyze the populations of virgin texts under investigation or at least an unbiased sample of these.

5 Guidelines and Concluding Remarks

Summarizing our discussion of the issues involved in interpreting text scores, we offer some guidelines for researchers wanting to know which transformation is best for their application.

1. *Compare raw scores to one another whenever possible*. Raw scores are informative relative to each other, convey substantive information about the word overlap, and are also bounded on the [$A_{\min}$, $A_{\max}$] reference score interval. In many research problems, absolute comparison to a reference metric or to external scores is not necessary; many other commonly used scaling techniques do not offer this possibility. In much of our own applied work, for instance, we have found raw scores to be informative without transformation (e.g., Laver and Benoit 2002;

[6] We note that for any two texts, the LBG transformation will exactly recover the reference values as transformed scores if *only* those two reference texts are scored and transformed as virgin texts. LBG only fails to recover the reference texts' input scores when additional (virgin) texts are also included. In a sense, then, the problem of recovering reference values can be seen as one of sensitivity to virgin texts, rather than two separate problems.

Benoit et al. 2005; Laver, Benoit, and Sauger 2006). Since all metrics are simply linear rescalings of one another when only two reference texts are used, raw scores relative to a [0, 1] reference interval may be completely adequate for comparing virgin texts. When comparisons also extend to the raw scores of reference texts scored as virgin texts, however, it must be kept strongly in mind that reference text scores will be bunched according to the proportion of nondiscriminating words they contain. Whether this nondiscriminating content reflects substantive difference or simply noninformation will depend on the researcher's assessment of the particular question of interest.

2. *When using the LBG transformation, only compare like with like*. The researcher should make determined efforts to analyze the population of virgin texts of interest or an unbiased sample of these. Any standardization technique, of which LBG is an example, explicitly or implicitly assumes that the rescaled items are the population of interest or a large unbiased sample from this. For like to really be like, furthermore, this means a researcher must be willing to assume that virgin and reference text positions are drawn from a distribution with the same variance. Although this may seem like a strong and possibly unrealistic assumption, it is hardly uncommon to make strong assumptions about similarities in variances in the practice of applied statistical research.
3. *If more than two high-quality reference texts are available and transformation is motivated by a desire to compare like-for-like reference and virgin texts on the same absolute metric, use the LBG transformation*. If a researcher can accept the assumptions built in to the LBG transformation, the LBG transformation appears to work quite well in faithfully rescaling raw virgin text scores to the reference metric, as judged by external benchmarks in a variety of tested applications. By this standard, for instance, the LBG transformation performs demonstrably better in the British manifesto example selected by MV.
4. *When directly comparing a small number of virgin texts to only two reference texts, consider using the MV transformation*. The main advantage of the MV transformation is its fixing of the virgin texts to values relative to the virgin text scores of the reference texts, and this may aid interpretation when very few virgin texts are being used. In the extreme case where only two virgin texts are scored, it often makes more sense to anchor on the reference text range rather than on a virgin text variance based on two cases. It must be kept in mind, however, that the results will place the virgin texts relative to the dissimilarity of reference texts and not to the absolute difference of reference values chosen, although we can imagine situations in which such a focus might be explicitly warranted. Before the MV transformation is approved for general audiences, however, we would want to see concrete demonstrations of its superior results in a practical research context, validated by external estimates.

In conclusion, we view the MV transformation as a valuable addition to Wordscores, although not as one that should replace the LBG transformation in every circumstance. We remain convinced, furthermore, that using no transformation at all of the raw scores of the virgin texts may often be the best solution. The statistical analysis of text is a hugely important ongoing project, of which the Wordscores method is but a small part. The present healthy discussion over rescaling raw Wordscores estimates is certainly not the final word on this type of problem, which involves translating the output of statistical or numerical models onto substantively intuitive metrics—and thus extends to many other methods for the statistical analysis of text.

Funding

The Irish Research Council for Humanities and the Social Sciences (Government of Ireland Research Fellowship to K. B.).

References

Benoit, Kenneth, Michael Laver, Christine Arnold, Madeleine O. Hosli, and Paul Pennings. 2005. Measuring national delegate positions at the convention on the future of Europe using computerized wordscoring. *European Union Politics* 6:291–313.

Clinton, Joshua, Simon Jackman, and Douglas Rivers. 2004. The statistical analysis of roll call voting: A unified approach. *American Political Science Review* 98:355–70.

Gabel, Matthew, and John Huber. 2000. Putting parties in their place: Inferring party left-right ideological positions from party manifesto data. *American Journal of Political Science* 44:94–103.

Laver, Michael, and John Garry. 2000. Estimating policy positions from political texts. *American Journal of Political Science* 44:619–34.

Laver, Michael, and Kenneth Benoit. 2002. Locating TDs in policy spaces: Wordscoring Dáil Speeches. *Irish Political Studies* 17(1):59–73.

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, Kenneth Benoit, and Nicholas Sauger. 2006. Policy competition in the 2002 French legislative and presidential elections. *European Journal of Political Research* 45:667–97.

Martin, Lanny W., and Georg Vanberg. 2007. A robust transformation procedure for interpreting political text. *Political Analysis* (forthcoming).

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

Poole, Keith, and Howard Rosenthal. 1997. *Congress: A political-economic history of roll call voting.* New York: Oxford University Press.

Slapin, Jonathan, and Sven-Oliver Proksch. 2007. A scaling model for estimating time-series policy positions from texts. Paper presented at the annual meeting of the Midwest Political Science Association, Palmer House Hilton and Towers, Chicago, IL, April 12, 2007.