

SMSO

Survey of Mathematics & Science Opportunities

CURRICULUM ANALYSIS
TECHNICAL REPORT SERIES
No. 4

DOCUMENT ANALYSIS: REPRESENTATIVE COUNTRY PROFILES.

June 16, 1995

Project Director:
William Schmidt

MICHIGAN STATE
University

Document Analysis: Representative Country Profiles

Summary

- The index of the relative proportion of a country's textbooks or curriculum guides devoted to different categories of the TIMSS frameworks is called a country signature.
- A single country signature is useful for cross-country comparisons.
- However, an approach to characterizing the variability of different document signatures (within a country) is also needed especially within countries with a decentralized educational system or countries with a tracked system.
- One approach is to represent several sub-national or track signatures along with the country signature.
- Other approaches include various indices of variability.
- It is necessary to simulate the effect of various methods of constructing representative country signatures.
- Such tests indicate that current definitions of country signatures are seemingly robust.

Indices of Topic Coverage: Country Document Signatures

- A country signature is the relative proportion of a country's textbooks that are devoted to the different categories in the TIMSS Curriculum Frameworks.
- Each signature is defined for a given population and subject matter.
- The signature is an aggregate of indices of coverage. Numerically, the signature is an m-tuple of indices of coverage. The number m is the number of categories in the framework. A country signature can be used to contrast coverage of framework categories among different countries.
- Countries have submitted data from a national sample of textbooks and curriculum guides. Frequently these samples include data from more than one document per focal population.
- In order to accomplish among-country comparisons, it is necessary to combine the indices from each document or set of documents to create a single country signature
- This can be done by weighting the indices by the numbers of blocks in the document sets, by the proportions of students using the sets, or by a combination of the two.

- This type of weighting produces signatures that are analogous to set unions. Any topic that is covered in at least one document-set will have a non-zero component in the signature.
- We can mimic the notion of set intersection by defining an m-tuple of 1's and 0's: 1 if a topic is covered in all of the document-sets, and 0 otherwise. We can use this m-tuple as a mask to convert the m-tuple resulting from weighing schemes mentioned above to signatures that contain only categories that are covered in all of the document-sets.
- We can also manipulate the definition of a mask so that 1 would correspond to some percentage of the document-sets instead of all of the document-sets. Depending on the criterion chosen, the resulting signatures could attenuate similarities or differences among countries.

A Formal Specification of Country Signatures

In this section we present the formal definitions of a country signature and investigate the effect of weightings and masking on the relationships among the different definitions of the signatures.

Let

p_{di} = index of coverage for topic i , in document-set d .

m = number of categories.

n_d = number of blocks for document-set d .

w_d = proportion of students using document-set d .

D = number of document-sets, and

$$\sum_{d=1}^D w_d = 1$$

M_i = 1 if all $p_{di} > 0$, for $d = 1$ to D and

= 0 otherwise

Together, the M_i 's ($i = 1$ to m) form the mask. Note that for set union, the corresponding mask will have all 1's.

For a **country**, we may define the signature as:

1. Union of Document Sets:

$$x_{1i} = \sum_{d=1}^D \left[\frac{n_d}{\sum n_d} \right] p_{di}$$

$$\text{denote } c_{1d} = \frac{n_d}{\sum n_d} ; \sum_d c_{1d} = 1$$

2. Average of Document-Sets:

$$x_{2i} = \sum_{d=1}^D \left[\frac{1}{D} \right] p_{di}$$

$$\text{denote } c_{2d} = \frac{1}{D} ; \sum_d c_{2d} = 1$$

3. Weighted Union of Document-Sets:

$$x_{3i} = \sum_{d=1}^D \left[\frac{w_d n_d}{\sum w_d n_d} \right] p_{di}$$

$$\text{denote } c_{3d} = \frac{w_d n_d}{\sum_d w_d n_d} ; \sum_d c_{3d} = 1$$

4. Weighted Average of Document-Sets:

$$x_{4i} = \sum_{d=1}^D w_d p_{di}$$

$$\text{denote } c_{4d} = w_d ; \sum_d c_{4d} = 1$$

To contrast the different signatures, and their relationship to w_d 's, we investigate the correlations among the different definitions of signature.

Denote the matrix P:

$$P = \begin{bmatrix} p_1 & p_2 & L & p_D \end{bmatrix}$$

$$\text{where } p_1' = \begin{bmatrix} p_{d1} & p_{d2} & L & p_{dm} \end{bmatrix}$$

and

$$C = \begin{bmatrix} c_1 & c_2 & c_3 & c_4 \end{bmatrix}$$

$$\text{where } c_1' = \begin{bmatrix} c_{11} & c_{12} & L & c_{1D} \end{bmatrix}, \text{ etc.}$$

The 4 signatures are then:

$$X = \begin{bmatrix} x_1 & x_2 & x_3 & x_4 \end{bmatrix}$$

$$\text{where } x_1' = \begin{bmatrix} x_{11} & x_{12} & L & x_{1m} \end{bmatrix}, \text{ etc.}$$

$$X = PC$$

And the covariance matrix is:

$$\hat{\Sigma}_x = \frac{1}{m} C' P' \left[I_m - \mathbf{1}_m (\mathbf{1}_m' \mathbf{1}_m')^{-1} \mathbf{1}_m' \right] P C$$

$$= C' \hat{\Sigma}_p C$$

$\hat{\Sigma}_p$ is the $D \times D$ covariance matrix for the p 's.

To mask the signatures, define a $m \times m$ diagonal matrix M such that the i th entry corresponds to M_i . Then X becomes

$$X = MPC$$

and the covariance matrix becomes

$$\begin{aligned}\hat{\Sigma}_x &= \frac{1}{m} C' P' M' \left[I_m - 1_m (1_m' 1_m')^{-1} 1_m' \right] MPC \\ &= C' \hat{\Sigma}_{MP} C\end{aligned}$$

$\hat{\Sigma}_{MP}$ is the $D \times D$ covariance matrix for the p 's masked by M .

The Effects of Masking

Masking affects the computation of the variances and covariances among the p 's. It increases the number of 0,0 pairs in the computations of variances and covariances. The effect is to increase the magnitude of the correlations; at worst, it leaves the correlations unchanged. This suggests that the effects of the weights in C will similarly affect the correlations among the four definitions of signatures, masked as contrasted to not masked. In fact, the correlations for unmasked or set union based signatures are lower bounds for the corresponding masked or set intersection based signatures.

To determine the bounds on the correlations among the X 's, consider the following extreme assumptions.

1. p 's are perfectly correlated with identical variances. This is equivalent to the case where all the students are using similar documents.

The $\hat{\Sigma}_p$ can be written as:

$$\hat{\Sigma}_p = \hat{\sigma}^2(1_D 1_D')$$

and

$$\begin{aligned}\hat{\Sigma}_x &= \hat{\mathbf{S}}^2 C' (1_D 1_D') C \\ &= \hat{\mathbf{S}}^2 (1_4 1_4')\end{aligned}$$

In other words, the x 's are perfectly correlated or the w_d 's will have no effect.

2. p 's are perfectly correlated but with different variances:

$$\hat{\Sigma}_p = \text{diag}(\hat{\sigma}_d) 1 1' \text{diag}(\hat{\sigma}_d)$$

and

$$\hat{\Sigma}_x = C' \text{diag}(\hat{\sigma}_d) 1 1' \text{diag}(\hat{\sigma}_d) C.$$

The correlations among the x 's would also depend on the $\hat{\sigma}_d$. If they are similar and do not vary greatly in magnitude, the correlations should approach 1.

3. p 's are uncorrelated with different variances. This is equivalent to the case where all the document sets have nothing in common. Then $\hat{\Sigma}_p$ is a diagonal matrix of variances.

Suppose the variances are similar or identical and

$$\hat{\Sigma}_p \approx \hat{\sigma}^2 \mathbf{I}_D$$

then

$$\hat{\Sigma}_x = \mathbf{C}' \hat{\Sigma}_p \mathbf{C} = \hat{\sigma}^2 \mathbf{C}' \mathbf{C}$$

$$\approx \hat{\sigma}^2 \begin{bmatrix} \sum_1^D \frac{n_d^2}{[\sum n_d]^2} & \frac{1}{D} & \sum_1^D \frac{w_d n_d^2}{[\sum n_d][\sum w_d n_d]} & \sum_1^D \frac{w_d n_d}{[\sum n_d]} \\ & \frac{1}{D} & \frac{1}{D} & \frac{1}{D} \\ & & \sum_1^D \frac{w_d^2 n_d^2}{[\sum w_d n_d]^2} & \sum_1^D \frac{w_d^2 n_d}{[\sum w_d n_d]} \\ \text{Sym.} & & & \sum_1^D w_d^2 \end{bmatrix}$$

Then the correlation between x_1 and x_2 is

$$\hat{\rho}_{x_1 x_2} = \sqrt{\frac{D \bar{n}^2}{\sum n_d}} \leq 1$$

If the n_d 's are constant, then

$$\sum n_d^2 = n^2 D = \bar{n}^2 D$$

or

$$\hat{\rho}_{x_1 x_2} = 1$$

The correlation between x_1 and x_3 is

$$\hat{\rho}_{x_1 x_3} = \frac{\sum w_d n_d^2}{\sqrt{(\sum n_d^2)(\sum n_d^2 w_d^2)}}$$

Again, if n 's are similar or constant,

$$\hat{\rho}_{x_1 x_3} = \frac{n^2 \sum w_d}{\sqrt{D n^2} \sqrt{n^2 \sum w_d^2}} = \frac{1}{\sqrt{D \sum w_d^2}}$$

$$[\text{If } \mathbf{W}_d = \frac{1}{d} \text{ then } \hat{\rho}_{x_1 x_3} = 1]$$

Similarly, the correlation between x_1 and x_4 is

$$\hat{\rho}_{x_1 x_4} = \frac{\sum w_d n_d}{\sqrt{\sum n_d^2 \sum w_d^2}}$$

and if n 's are constant

$$\hat{\rho}_{x_1 x_4} = \frac{1}{\sqrt{D \sum w_d^2}}$$

Note that

$$\sum w_d^2 = D(\sigma_w^2 + \bar{w}^2)$$

Since

$$\sum w_d = 1, \bar{w} = \frac{1}{D}, \text{ or } \sum w_d^2 = D\left(\sigma_w^2 + \frac{1}{D^2}\right)$$

or

$$D \sum w_d^2 = D^2 \sigma_w^2 + 1$$

This suggests the lower bound to

$$\hat{\rho}_{x_1 x_3} = \hat{\rho}_{x_1 x_4} = \frac{1}{\sqrt{D^2 \sigma_w^2 + 1}}$$

where D = number of tracks or regions and σ_w^2 is the variance of the w_d 's.

For example, when D is equal to 2, the maximum for σ_w^2 is .25 and the correlation is $\frac{1}{\sqrt{2}}$ or .71. When D is 5, and w_d 's are between .1 and .5, for example, .1, .1, .1 .2 and .5, then σ_w^2 is .0024 and the correlation is .79.

Implications for the Robustness of Alternative Country Signatures

We have explored how to combine multiple document-sets submitted by a country into one signature of topic coverage for that country. We devised four different approaches: weighted by the number of blocks in the document-sets, weighted by the proportions of students using each of the document-sets or weighted by a combination of both. These are signatures that are based on the notion of set union serving to represent topic coverage for the country.

To create signatures that were based on set intersection, we can simply devise a mask to mask out categories that were not covered in all of the document-sets. Applying the mask would actually increase the correlations among the different signatures. Or the correlations for the set union signatures were lower bounds to the set intersection signatures.

We further investigated the effects of weightings by considering situations that would produce low correlations among the different signatures. We determined the lower bound to the correlations, by considering some of the more extreme circumstances. For example, a situation in which different tracks or regions used textbooks that had no categories in common.

Not surprisingly, the correlation is affected by the distribution of the proportions of students using the different sets. With a diverse distribution of weights, the different signatures still maintained moderate to high correlations. This suggested that for the kind

of data we have for document analysis, the country signatures of topic coverage were robust to effects of the different weighting schemes.

Indices for characterizing the variability of document signatures within a country

- The collection of indices represents the signature of a document-set. Document-sets differ in the number of topics covered and the amount of coverage for these topics. It is desirable to have a single index or coefficient for an array of document-sets to facilitate comparison between sets, i.e., between countries. We propose different ways to capture the diversity or similarity of document-sets.
- Let the signature,
$$\mathbf{p}_d' = [p_{1d} \quad p_{2d} \quad \dots \quad p_{md}]$$
be the vector of coverage indices for document-set \mathbf{d} . p_{id} is the index of coverage for topic \mathbf{i} in the document-set. \mathbf{m} is the number of topics from the framework. Let \mathbf{D} be the total number of document-sets in a set. For convenience and without loss of generality, we drop the superscripts that identify the type and origin of the document-set.
- **Generalized Variance**
Generalized variance is one way to capture the variation of a set of random variables. Collect the \mathbf{D} document-set vectors into the matrix \mathbf{P} . Then the determinant of the variance-covariance matrix for \mathbf{P} , by treating each \mathbf{p}_d as a variable, is the generalized variance of the \mathbf{D} vectors.

- **Distances**

Each \underline{pd} is a m -tuple and can be represented as a vector in the m -dimensional Topic Space. Then another way to capture the dispersion of these D vectors is to compute the average distances between the vectors in the set.

Instead of using the average, the minimum or the maximum distance could also be used.

- **Angles between vectors**

As opposed to distances, the cosine of the angle between the vectors is a measure of similarity. The average cosine of the angles between vectors is another measure.

- **Threshold**

A document-set sometimes includes many topics but a great number of them have small coverage indices. Therefore, to accentuate the pattern of coverages, we may set a coverage threshold, such as 0.01, and set indices that are smaller than the threshold to zero. The procedure should result in increasing the magnitude of distances and similarly decreasing the magnitude of the cosines between vectors.