

SMSO

**Survey of Mathematics
& Science Opportunities**

**CURRICULUM ANALYSIS
TECHNICAL REPORT SERIES
No. 3**

DOCUMENT ANALYSIS: GENERATION OF ANALYSIS FILES.

June 14, 1995

Project Director:
William Schmidt

MICHIGAN STATE
University

DOCUMENT ANALYSIS: GENERATION OF ANALYSIS FILES

Summary

- Coders in each TIMSS country provided Document Analysis (DA) data that are entered and processed by the TIMSS Curriculum Analysis Project (see Reports 1 and 2 in this series).

- A series of routines are performed to generate DA analysis files that include
 - calculation of frequency of occurrence of codes
 - calculation of frequency of occurrence of paired codes
 - calculation of frequency of occurrence of code triplets
 - calculation of frequency of occurrence of content codes and performance expectation codes used in tandem
 - calculation of frequency of occurrence of codes within a national document sample
 - calculation of indices of coverage
 - calculation of summary statistics

Overview

The document analysis is done by coders in each country following the instructions and content frameworks provided by TIMSS. Each document is first divided into units that would take one to three days to cover. Each unit of the text is further sub-divided into coherent content blocks. The coder codes each block according to its content.

There are three different aspects of the content of the text that the coder attends to: content topic, performance expectation, and perspective. The coder potentially codes each block with multiple codes in each of the three aspects. For both content topic and performance expectation, the coder also identifies whether the codes are primary or secondary coverage in the content of the text.

The coder also characterizes the type of unit and block according to pre-defined categories. For the present analysis, all the units and blocks for curriculum guides are included. But only instructional related units (unit type: 2 Lesson) and blocks (block type: 1 Narrative; 2 Related Narrative; 4 Related Graphic; 6 Exercise/question set; 8 Activity; and 9 Worked Example) for textbooks are included.

Following, we present a mathematical description of the procedures for generating the summary statistics files used to analyze these data. For the simple case where there is only one code per aspect per block, this file would be trivially derived and the example given in report #2 in this series illustrates what such a summary statistics file might look like. As indicated in this report complexity arises and the need for a more detailed analysis file structure when blocks are coded with multiple codes on each of the three aspects. This leads to what we have termed complex signatures. This case of which the simpler is a subset is described generically in the following pages.

Understanding the Codes

- To represent the data submitted by the coder, let

$$Code_{d'ub i}^{(c_o a_s p_o s_m d_t)}$$

represent the **i**'th aspect code of type **a_s** entered by the coder in country **c_o**, for block **b** of unit **u** of document **d'**. Document **d'** is for student target **p_o**, subject matter **s_m**, and book type **d_t**.

- For the superscripts,

c_o is the country code,

a_s represents the different aspects of content:

- c - content topic,
- p - performance expectation,
- o - perspective,

p_o represents the target student population for the document:

- 1 - Population 1 upper grade,
- 2 - Population 2 upper grade,
- 3 - Population 3 specialists.

s_m represents the subject matter of the document:

- m - Mathematics,
- s - Sciences,

d_t represents the kind of document:

- g - curriculum guide,
- t - textbook.

- For the subscripts,

d' is the document identification number,

u is the **u**th unit of document **d'**,

b is the **b**th block in the **u**th unit, and

i' is the **i**'th aspect code entered by the coder in this block,

For content topic, i.e., **a_s** = c above,

i' = 1, 2, ... , 8 for up to 8 possible primary content codes, and
i' = 9, ..., 13 for up to 5 possible secondary content codes

For performance expectation, i.e., $\mathbf{a}_s = p$ above,

$i' = 1, 2, \dots, 8$ for up to 8 possible primary performance expectation codes, and

$i' = 9, \dots, 13$ for up to 5 possible secondary performance expectation codes

For perspective, i.e., $\mathbf{a}_s = o$ above,

$i' = 1, 2, \dots, 5$ for up to 5 possible primary perspective codes.

The Frameworks

- The 4-level codes used for the Science and Mathematics Frameworks can be represented as:

$a_0.a_1.a_2.a_3$

where **a_0** corresponds to the aspect of the content of the document

- 1 - content topic
- 2 - performance expectation
- 3 - perspective

a_1 , **a_2** and **a_3** correspond to progressively more detailed descriptive categories. That is, **$a_0.a_1.a_2.a_3$** are sub-categories of **$a_0.a_1.a_2$** . The code **$a_0.a_1.a_2$** is used to represent the set of sub-categories in the same **$a_0.a_1.a_2$** category. Similarly, the main category **$a_0.a_1$** is used to represent the set of categories in the same **$a_0.a_1$** main categories.

- The notation,

$fw_i^{(a_s s_m 3)}$

is used to represent the **$a_0.a_1.a_2.a_3$** sub-categories as ordered in the framework. The **a_0** will correspond to **a_s** as previously defined. And **s_m** denotes the subject matter.

Similarly,

$fw_i^{(a_s s_m 2)}$

is used to represent the **$a_0.a_1.a_2$** categories as ordered in the framework, and

$fw_i^{(a_s s_m 1)}$

is used to represent the **$a_0.a_1$** main categories as ordered in the framework. For simplicity, we use the superscript **lv** to represent the three levels of the hierarchy:

$fw_i^{(a_s s_m lv)}$

Calculating the Frequency of Occurrence of a Code Used in a Document

- Define the indicator function

$$\begin{aligned} ind\{statement\} &= 1 \quad \text{if statement is true,} \\ &= 0 \quad \text{if statement is false.} \end{aligned}$$

- The frequency of a framework code mentioned in a document is the number of blocks in which the code is used:

$$f_{d'i}^{(c_o a_s p_o s_m d_t l_v)} = \sum_u \sum_b ind\{\bigcup_{i'} (Code_{d'ub i'}^{(c_o a_s p_o s_m d_t l_v)} \in fw_i^{(a_s s_m l_v)})\} \quad (\text{Eq. 1})$$

Note that the coder is not to use the same code multiple times in the same block. Operationally, the codes in each block are pre-processed to eliminate duplicates and are re-ordered according to the framework. Hence, Eq. 1 is equivalent to

$$f_{d'i}^{(c_a p_o s_m d_t l_v)} = \sum_u \sum_b \sum_{i'} ind\{Code_{d'ub i'}^{(c_o a_s p_o s_m d_t l_v)} \in fw_i^{(a_s s_m l_v)}\} \quad (\text{Eq. 2})$$

In addition, the coder is to code the materials using the lowest level of the framework codes. Again, the codes in each block are pre-processed so that if a higher level code is used, it will be replaced by the corresponding lower level codes.

Calculating the Frequency of Occurrence of Two Codes Used Together in a Document

- The frequency of two framework codes of the same aspect mentioned together in a document is the number of blocks in which both codes are used together:

$$f_{d'i<j}^{(c a p_o s_m d_t l)} = \sum_u \sum_b ind\left\{\left(\bigcup_{i'}(Code_{d' u b i'}^{(c_o a_s p_o s_m d_t)} \in fw_i^{(a_s s_m l_v)})) \cap \left(\bigcup_{i'}(Code_{d' u b i'}^{(c_o a_s p_o s_m d_t)} \in fw_j^{(a_s s_m l_v)}))\right)\right\} \quad (\text{Eq. 3})$$

Calculating the Frequency of Occurrence of Three Codes Used Together in a Document

- The frequency of three framework codes of the same aspect mentioned together in a document is the number of blocks in which the three codes are used:

$$f_{d'i<j<k}^{(c a p_o s_m d_t l)} = \sum_u \sum_b ind\left\{\left(\bigcup_{i'}(Code_{d' u b i'}^{(c_o a_s p_o s_m d_t)} \in fw_i^{(a_s s_m l_v)})) \cap \left(\bigcup_{i'}(Code_{d' u b i'}^{(c_o a_s p_o s_m d_t)} \in fw_j^{(a_s s_m l_v)}))\right) \cap \left(\bigcup_{i'}(Code_{d' u b i'}^{(c_o a_s p_o s_m d_t)} \in fw_k^{(a_s s_m l_v)}))\right)\right\} \quad (\text{Eq. 4})$$

Calculating the Frequency of Occurrence of Content Topics and Performance Expectation used Together in a Document

- To limit the number of combinations, only a subset of performance expectation codes are considered. They are further grouped into categories separately for science and mathematics. We use the $a_s = p$ to denote these grouped categories. For science, the categories are:

$fw_h^{(p\ s\ 1)}$,	where
h = 1,	Understanding Simple Information
2,	Understanding Complex Information
3,	Theorizing, Analyzing & Solving Problems
4,	Using Tools, Routine Procedures & Science Processes
5,	Investigating the Natural World
6,	Communicating.

And for mathematics, the categories are:

$fw_h^{(p\ m\ 1)}$,	where
h = 1,	Knowing and Using Vocabulary
2,	Using equipment/Performing Routine Procedures
3,	Using Complex Procedures
4,	Investigating & Problem Solving
5,	Mathematical Reasoning
6,	Complex Communication.

Furthermore, we use $a_s = x$ to denote these content topic by performance expectation combinations.

- The frequency of a content topic code mentioned with a performance expectation grouped category together in a document is the number of blocks that:

$$f_{d'ih}^{(c_o \times p_o \ s_m \ d_t \ l_v)} = \sum_u \sum_b \text{ind}\{(\bigcup_{i'} (Code_{d'ub i'}^{(c_o \ c \ p_o \ s_m \ d_t)} \in fw_i^{(c \ s_m \ l_v)})) \cap (\bigcup_{i'} (Code_{d'ub i'}^{(c_o \ e \ p_o \ s_m \ d_t)} \in fw_h^{(p \ s_m \ 1)}))\} \quad (\text{Eq. 5})$$

The frequency of two content topic codes mentioned with a performance expectation grouped category together in a document is the number of blocks that:

$$f_{d'ijh}^{(c_o \times p_o \ s_m \ d_t \ l_v)} = \sum_u \sum_b \text{ind}\{(\bigcup_{i'} (Code_{d'ub i'}^{(c_o \ c \ p_o \ s_m \ d_t)} \in fw_i^{(c \ s_m \ l_v)})) \cap (\bigcup_{j'} (Code_{d'ub j'}^{(c_o \ c \ p_o \ s_m \ d_t)} \in fw_j^{(c \ s_m \ l_v)})) \cap (\bigcup_{i'} (Code_{d'ub i'}^{(c_o \ e \ p_o \ s_m \ d_t)} \in fw_i^{(p \ s_m \ 1)}))\} \quad (\text{Eq. 6})$$

The frequency of three content topic codes mentioned with a performance expectation grouped category together in a document is the number of blocks:

$$f_{d' i < j < k h}^{(c_o x p_o s_m d_t l_v)} = \sum_u \sum_b ind\{(\bigcup_{i'}(Code_{d' u b i'}^{(c_o c p_o s_m d_t)} \in fw_i^{(c s_m l_v)})) \cap (\bigcup_{i'}(Code_{d' u b i'}^{(c_o c p_o s_m d_t)} \in fw_j^{(c s_m l_v)})) \cap (\bigcup_{i'}(Code_{d' u b i'}^{(c_o c p_o s_m d_t)} \in fw_k^{(c s_m l_v)})) \cap (\bigcup_{i'}(Code_{d' u b i'}^{(c_o e p_o s_m d_t)} \in fw_i^{(p s_m l)}))\} \quad (Eq. 7)$$

- **Unit Type, Block Type and Number of Blocks in a Document**

Denote

$$n_{d'}^{(c_o \bullet p_o s_m d_t)}$$

as the number of blocks in a document. As mentioned above, this corresponds to all the blocks in all the units for curriculum guides. In textbooks, it corresponds only to instructional blocks within lesson and multiple-lesson type units.

Calculating the Frequency of Occurrence of Codes for Document Set

- Documents from a country are sometimes grouped together as a document-set. This information is provided by the country's NRC.
- We use **d** to denote these combined document-sets. For example,

$$f_{d i}^{(c_o a_s p_o s_m d_t l_v)} = \sum_{d'} f_{d' i}^{(c_o a_s p_o s_m d_t l_v)} = \sum_{d'} \sum_u \sum_b ind\{\cup_{i'} (Code_{d' u b i'}^{(c_o a_s p_o s_m d_t l_v)} \in fw_i^{(a_s s_m l_v)})\}$$

- **Frequency of Occurrence of Codes for a Country**

For summary at the country level, we also combine document-sets into a country summary. For example,

$$f_i^{(c_o a_s p_o s_m d_t l_v)} = \sum_d f_{d i}^{(c_o a_s p_o s_m d_t l_v)}$$

- **Index of coverage**

The index of coverage is defined for each code or combination of codes in the framework. It is the ratio of the frequency of occurrence to the total number of blocks in the corresponding document or documents in a document-set. This index is usually reported as a percent. For example, the country level index of coverage for code i is

$$p_i^{(c_o a_s p_o s_m d_t l_v)} = \frac{f_i^{(c_o a_s p_o s_m d_t l_v)}}{n^{(c_o \bullet p_o s_m d_t l_v)}}$$

- **Content of the Summary Statistics File**

Since the analysis at this stage of the study is based solely on the summary statistics defined above, it is efficient to process the raw data files - codes provided by the coders for each document - and maintain a separate summary statistics file. This is sometimes referred to as the analysis file.

Each record in the file includes:

c_o, a_s, p_o, s_m, d_t, l_v, d', f, f1, f2, f3, f4, n, string

For textbook, only data from unit type 2 and 3 are used in creating the summary statistics. f1, f2, f3 and f4 are frequency of occurrences for different block types:

1. Narrative, Related Narrative and Related Graphic (block type 1, 2 and 4)
2. Exercise/question set (block type 6)
3. Activity (block type 8)
4. Worked Example (block type 9)

The string contains the actual framework code or code combination that corresponds to the frequencies. See Figure 1. (using Japanese mathematics textbook ID: M181C, as in Report No. 2)

Figure 1