Examination and Critique of Codebook for Textual Analysis

L. Donna Fowler
Florida International University, USA

Abstract: The researcher presents the details, findings, and critique of a pre-pilot study conducted on a codebook created for a textbook comparison. She used Cohen's alpha and percent agreement to determine inter-rater reliabilities for coding categories. These values revealed changes needed in the coding scheme and in the coder training process for the future comparison study.

International comparison tests have placed students from Singapore at the top and students from the United States as average on these tests (Zhu & Fan, 2004). Studies of the mathematical systems from nations participating in these tests suggest that differences in textbooks may help explain this disparity in achievement (Ginsburg, Leinwand, Anstrom, & Pollock, 2005). To find the differences between mathematics textbooks from the two countries a textbook comparison can be used. The theoretical background for the textual comparison is from Vygotsky's concept of symbolic mediation, the idea that symbolic tools, including algebraic symbols, organize and control mental processes (Lantolf, 1994). A textual analysis will allow us to examine the use of symbolic tools in a text.

In preparation for a comparison of mathematics textbooks from Singapore and the United States, a pre-pilot study was conducted to determine the feasibility of a coding instrument (Appendix A) created by the author. The coding scheme was designed to examine 23 features of the text (See Appendix B for researchers who have influenced these features). According to Neuendorf (2002), one of the purposes of a pilot study is to address four main threats to reliability of a content, or textual, analysis. These four threats are "1. a poorly executed coding scheme, 2. inadequate coder training, 3. coder fatigue, and 4. the presence of a rogue coder" (Neuendorf, 2002, p. 145). The pre-pilot study addressed these threats by identifying problems in the coding scheme and within the coder training process. The goal was to find ways to improve the codebook before using it in the textbook comparison study. The purpose of this manuscript is to present the details, findings, and critique of this pre-pilot study.

Addressing Threats to Reliability

In a coding scheme, the categories need to be clear, unambiguous, and must consist of mutually exclusive sets (Neuendorf, 2002). Neuendorf recommended that the codebook and coding forms should be so well-defined as to virtually eliminate any differences in the coding by different individuals. One way the pre-pilot study will address the threats to reliability is by identifying problems in the coding scheme. The inter-rater reliabilities of two coders will be assessed using percent agreement and Cohen's alpha. From these values, the researcher will be able to determine any definitions or coding categories that need to be changed in the codebook.

The coder training process is also an issue that the pre-pilot study will examine. Neuendorf (2002) presented a 15-step process when creating a coding scheme. This process involves (a) creating the codebook, (b) three iterations of coder training, coder practice, coder discussion, and codebook revision, (c) the final coding, and (d) analysis of the experience (Neuendorf, 2002). Due to time restrictions, the coder training was truncated to one iteration of Neuendorf's proposed guidelines for coder training. This affected the results of the inter-rater reliability values. However, the study was beneficial in pinpointing things that need to be

Fowler, L. D. (2008). Examination and critique of codebook for textual analysis. In M. S. Plakhotnik & S. M. Nielsen (Eds.), *Proceedings of the Seventh Annual College of Education Research Conference: Urban and International Education Section* (pp. 38-45). Miami: Florida International University. http://coeweb.fiu.edu/research_conference/

changed in the codebook and the coder training process and in revealing sources of potential problems in the future textbook comparison study.

Methodology

Two coders independently coded a randomly chosen sample of text pertaining to linear functions in a mathematics textbook from the US with respect to 23 features. Inter-rater reliability values were assessed using Cohen's *kappa* and percent agreement. The two reliability values represent the differences in Cohen's kappa's conservative value (Lombard, Snyder-Duch, & Bracken, 2005; Neuendorf, 2002) and percent agreement's more liberal index for estimating inter-rater reliability. The desired value for the inter-rater reliability coefficient was chosen to be between .9 for percent agreement and .75+ for Cohen's kappa as this is the level acceptable in most situations (Lombard et al., 2005; Neuendorf, 2002). General features were also compared.

Implementation of the Study

After the creation of the coding scheme and codebook, the researcher held an informal training session with the second coder. First, the second coder was given the codebook and coding scheme and asked if there were any questions about the definitions. Next, the two coders discussed the relationship and differences between linear functions and linear equations. They then looked at a sample text, McGraw-Hill's Teacher Wraparound edition of *Algebra 1* (Holliday et al., 2004) and covered a few examples of linear functions as compared to linear equations. They discussed the classification of object-analytic images and object-illustrative images using examples from the text. A teacher text was chosen for this training session so that the available student text, *Algebra: Structure and Method: Book 1* (Brown, Dolciani, Sorgenfrey, & Cole, 2000) by Houghton Mifflin, could possibly be used in a second iteration of the coder-training process in the future. Neither text will be used in the comparison study. Preliminary revisions to the codebook included a less ambiguous definition of linear functions and the replacement of the word lessons with the word sections with a short explanation for the term.

After the codebook was revised, two sections pertaining to linear functions from the textbook, *Algebra: Structure and Method Book I* (Brown et al., 2000), were randomly chosen to be the sample text in the pre-pilot study. The textbook was a text that had previously been used in a U.S. classroom and represented a typical traditional U.S. mathematics textbook. The coders independently coded the two sections. An Excel program was used to facilitate the recording of the coded categories while a coding form was used to record the numbers within the general-feature categories. A rubric listing all the problems in each section by category heading (i.e. Oral Exercises or Mixed Review Exercises) was created. However, the category Self-Test 1 was inadvertently omitted from the first section, so these eight problems were not included in the study. Also, one coder coded all problems in the first section and the problems pertaining only to linear functions in the second section. The other coder coded problems only pertaining to linear functions in both sections. This discrepancy in the total number of problems coded would have given an inaccurate measure of the level of agreement for the coding instrument. For that reason, only the problems pertaining to linear functions that both coders had examined were used in the calculations for the inter-rater reliability values. The total number of problems examined was 54.

Results

The pre-pilot study was designed to determine the extent to which two coders agreed on rating 23 features within the text. The amount of inter-rater agreement on the codes for these items will determine the usefulness of the definitions and categories within the coding scheme and codebook to be used in the textbook comparison. This section will consist of two segments: a discussion of the inter-rater agreement results and a discussion of what the levels of agreement

reveal about the coding instrument. The results will be delineated by characteristic examined.

General Features of the Text

The general features of the text did not require a choice of code categories except for the images, which required a count for each type of image. Thus, Cohen's kappa and percent agreement were not found for these characteristics. The results consisted of a comparison of the two coders' assessments of the features as listed on their coding forms. These values are listed in Table 1. None of the coders' assessments were a perfect match. One category, number of pages for each lesson, was not included in Table 1. This was done because of the differences in interpretation of the characteristic and the fact that reporting the average number of pages per linear function section was a better approach than listing the number of pages for the 100+ sections. The values were different for the coders due to the numbers for pages and sections being different. The coders were most in agreement about the average number of pages per section pertaining to linear functions with values of 5.43 and 5.22.

Table 1

Comparison of Data for General Characteristics of the Text

Feature	Coder 1	Coder 2
Pages in Text	794	807
Number of Chapters	12	13
Number of Sections	109	120
Linear Function Pages	38	47
Pages for Development	26	21
Pages for Exercises	32.5	26
Average Pages/Linear Function	5.43	5.22
Object-Illustrative Images	14	7
Object-Analytic Images	24	31
Signposts/Attention-Getters	24	16

General Classifications of the Problems

Cohen's kappa and percent agreement were the inter-rater reliability coefficients calculated for the six problem characteristics pertaining to linear functions. As seen in Table 2, the two inter-coder reliability coefficients differed drastically when agreement by chance was taken into account using Cohen's kappa and when chance was not taken into account using percent agreement. The percent agreements for the contextual feature and response-type were found to be 87% (0.87) and 69% (0.69), respectively. However, Cohen's kappa agreement coefficients were 0.46 and 0.12, respectively, for the two features. This demonstrated a criticism overlooked initially by the researcher but documented by Lombard et al. (2005) and Neuendorf (2002) that Cohen's kappa gives an estimate of agreement that is too conservative. However, Neuendorf (2002) presented a value of .4 to .75 as being a fair to good agreement when using Cohen's kappa. The fact that no value for Cohen's kappa was found for the given-information category presented a limitation of Cohen's kappa. If the coders do not use one of the coding categories but have complete agreement, Cohen's kappa cannot be calculated due to the fact that zero is obtained in the denominator. Thus, the coders agreed on every problem's classification and coded every problem with the same category. Similarly, both agreement coefficients were 1.0 for the characteristic application type, but not all problems were classified with only one code category. The inter-rater reliability coefficients varied widely with respect to problem characteristics and had little agreement between the two indices.

Table 2
Inter-Rater Reliability Coefficients for Categories Coded in Text

Feature	Cohen's kappa	Percent Agreement
	Problems (n = 54)	
Computational	.12	.69
Contextual	.46	.87
Response-Type	.52	.69
Cognitive Requirement	.22	.67
Given Information	---[a]	1
Application Type	1	1
Problem Practices	-0.04	.26

[a]Cohen's kappa does not yield a value due to only one characterization (i.e., code), being chosen for all problems. All problems were coded with the same code, so the coders agreed on the characterization of every problem in this category.

Characterization of Problem Practices in the Text

The inter-rater reliability results (See Table 2) show the lack of agreement in the characterization of problem practices between the coders on both indices. Cohen's kappa value, -.04, signifies an agreement that is less than chance. The percent agreement is a scant .26. Thus, the coders did not agree on their labels of the practices in problems pertaining to linear functions.

Cultural Indicators

Both coders had difficulty using the coding scheme to classify cultural indicators. Due to the experimental nature of these characteristics, the author included a category similar to unable to determine within the choices. Inclusion of such a category is recommended for any content analysis (Neuendorf, 2002). Both coders selected this category for all four cultural indicator features. Thus, the coders were unable to determine the presence of cultural indicators in the text.

Changes Recommended for Coding Scheme

This section discusses recommended changes concerning (a) general features of the text, (b) general classifications of the problems, (c) characterization of problem practices, and (d) cultural indicators.

General Features of the Text

From the results of the pre-pilot study, the researcher successfully determined changes that need to be made in the coding scheme. First, the fact that every general feature of the text was classified differently by the coders indicated that the definitions of these characteristics should be changed. Indeed, there were issues that could not have been foreseen unless one had experience in coding textbooks or had extensive knowledge of textbooks. There was a portion of text at the end of the textbook called Looking Ahead. This portion of text looked like a chapter but was not labeled as a chapter. This pseudo-chapter even contained pseudo-sections (i.e., sections that were not labeled as sections). Contingencies like this need to be considered as the codebook is revised. Other things that need to be addressed in the definitions and directions in the codebook are as follows: (a) Does one count the introductory pages before the chapters begin? (b) Can one have partial pages in the count? (c) Do extra teaching lessons within a section count as a new section? (d) Are all problems; oral, written, mixed review, computer, and self-test; counted as problems? (e) Do graphs count as images to be classified? (f) What are more explicit examples of signposts? (g) Does the number of pages for every lesson, even ones without linear functions, need to be determined? and (h) To what extent are linear equations part of the linear functions topic? These suggestions represent some of the deficiencies that need to

be addressed in the codebook within the general features section. Another factor to consider is whether the comparison study should focus only on linear functions in view of the fact that the study's main purpose is to reveal information about the problems within the text.

General Classifications of the Problems

For general classifications of problems, three problem features received acceptable inter-rater reliability levels using percent agreement. The other three received close to .7, which is an acceptable level for some exploratory studies (Lombard et al., 2005). However, using Cohen's kappa, only two problem features had acceptable inter-rater reliability levels. This included the given-information feature that received an invalid answer for Cohen's kappa. This implies that careful thought needs to be given in setting acceptable inter-rater reliability levels. A look at Krippendorff's alpha may be warranted. However, with clarification of the category codes using examples and caveats for the coder, the inter-rater reliability coefficients are expected to change.

Issues pertaining to general problem characteristics need to be resolved in the codebook. For example, the coder needs to know how the directions given in a problem affect the coding of contextual features. Also, the difference between conceptual understanding and problem solving must be clarified through examples or wording of the definitions. An example is the case where a computation is simple, but the student is asked to do the computation multiple times in one problem. This case needs to be highlighted in the codebook for classification as a single computation. More training for the coders with a concerted effort to discuss each characteristic of the problem will also affect the inter-rater reliability coefficients.

Within the general problem features, there was perfect agreement pertaining to the given-information feature. There were no problems that contained insufficient or extraneous data. While the U.S. text showed no variation in the data given in the problem, an examination of the Singapore text may not yield the same result. Thus, some of the benefits of the comparison study may not be ascertained from the results of this pre-pilot study.

Characterization of Problem Practices

While the results for the problem practices feature had very low levels of inter-rater reliability, the initial examination of this feature was encouraging. The prospect of determining differences in texts based upon the practices within the problems seemed feasible as there were several categories within the text. The differences in coding seemed to be due to inadequate understanding of the categories. Thus, further clarification of the type of problem that each category represents needs to be in the codebook. This can be done by listing examples for each category. Also, the training for coders should include some background of how and why the categories were created by Mesa (2004). Once all coders understand the code categories, suitable inter-rater reliability coefficients should follow.

Cultural Indicators

The results revealed that the definitions for the cultural indicators should be revised. Some concerted thought needs to go into determining if these categories can be found and classified in a textbook. Perhaps these features are not passed on solely through the content of the textbook but through its connection to the process i.e., how the book is used (Haggarty & Pepin, 2002). More information on culture and how it is passed on needs to be examined to create unambiguous definitions for determining how cultural indicators are seen in the text.

Critique of Pre-Pilot Study

The pre-pilot study was very successful in revealing ways to revise the coding instrument to more accurately reflect the characteristics within the text. The revisions suggested above will allow for more concrete definitions and examples within the codebook. The study also uncovered

deficiencies in coder training and differences in using Cohen's kappa and percent agreement to determine inter-rater reliability. Krippendorff's alpha may be looked at in a future study to obtain acceptable inter-rater reliability coefficients. The researcher will also consider leaving some of the general features of the text, such as number of pages for each section, out of the final comparison study. She will also ponder carefully how to view cultural indicators within the text.

Even though this pre-pilot study has informed the future revision of the codebook, there were weaknesses in the study. First, due to time constraints, Neuendorf's (2002) 15-step coder training/codebook development process was truncated to only the first iteration of coder training. This abbreviated process affected the inter-rater reliability values obtained. This was the first time for both coders to use a coding instrument. A second iteration would have taken care of some of the discrepancies in the inter-rater reliabilities. The proposed inclusion of examples and changes to the definitions demonstrate how the coder training process can inform the creation of a codebook. Another weakness was the researcher's inexperience with coder training. This was obvious as the trainee misread some of the directions and definitions, which should have been covered in the training session.

A limitation of the pre-pilot study was the fact that the coding is biased due to human error and misinterpreting the definitions. For example, one of the coders used other as a category for one feature in almost all of the problems. This consistent incorrect coding was due to the misinterpretation of the definitions. Some of these misinterpretations may be prevented by changes made to the codebook based on this pre-pilot study.

References

Brown, R. G., Dolciani, M. P., Sorgenfrey, R. H., & Cole, W. L. (2000). *Algebra: Structure and method book I*. Evanston, IL: Houghton Mifflin.

Ginsburg, A., Leinwand, S., Anstrom, T., & Pollock, E. (2005) *What the United States can learn from Singapore's world-class mathematics system (and what Singapore can learn from the United States): An exploratory study*. American Institutes for Research. Retrieved March 17, 2008, from http://www.air.org/news/documents/Singapore%20Report%20(Bookmark%20Version).pdf

Haggarty, L., & Pepin, B. (2002). An investigation of mathematics textbooks and their use in English, French and German classrooms: Who gets an opportunity to learn what? *British Educational Research Journal, 28*(4), 567-590.

Holliday, Marks, Cuevas, Casey, Moore-Harris, Day, et al. (2004). *Algebra 1 teacher wraparound Florida edition*. New York, NY: McGraw-Hill.

Lantolf, J. P. (1994). Sociocultural theory and second language learning: Introduction to the special issue. *The Modern Language Journal*. *78*(4), 418-420.

Lombard, M, Snyder-Duch, J., & Bracken, C. C. (2005). Practical resources for assessing and reporting intercoder reliability in content analysis research projects. Retrieved March 17, 2008, from http://www.temple.edu/sct/mmc/reliability/#How%20should%20content%20analysis%20researchers%20properly%20assess%20and%20report%20inter coder%20reliability.

Mesa, V. (2004). Characterizing practices associated with functions in middle school textbooks: An empirical approach. *Educational Studies in Mathematics, 56,* 255-286.

Neuendorf, K. A. (2002). *The content analysis guidebook*. Thousand Oaks, CA: Sage.

Zhu, Y., & Fan, L. (2004, July 4-11). *An analysis of the representation of problem types in Chinese and US mathematics textbooks*. Paper accepted for ICME-10, Discussion group 14, Copenhagen, Denmark.

Appendix A

Coding Instrument

Part I Background features

1. Number of pages in text
2. Number of chapters
3. Number of sections
4. Number of pages for each section
5. Number of pages pertaining to linear functions
6. Number of pages for development
7. Number of pages for exercises
8. Number of other pages
9. Number of problems pertaining to linear functions
10. Average number of pages per section
11. Average number of pages per section pertaining to linear functions
12. Type of images
    (OA) object-analytic images (OI) object-illustrative images
13. Number of signposts or attention-getters

Part II General classification of problem

1. Computational feature
    (S) single computation procedure (M) multiple computation procedures
2. Contextual feature
    (nu) numerical (vi) visual (ve) verbal (co) combined form
3. Response-type feature
    (A) numeric answer only (E) numeric expression only
    (ES) explanation or solution required (OP) other response
4. Cognitive requirement feature
    (PP) procedural practice (CU) conceptual understanding
    (PS) problem solving (SR) special requirement
5. Given-information feature
    (SF) sufficient (EX) extraneous (ISF) insufficient
6. Application type
    (AP) applied (NA) nonapplied

Part III Classification of problem practices

1. Characterization of problem practices
   (sr) symbolic rule (op) ordered pair (sd) social data
   (ph) physical phenomena (ci) controlling image (ot) other

Part IV Cultural indicator feature

1. Group dynamic
   (soc) social orientation (ind) individual orientation (utd) unable to determine

2. Level of importance of memorization
   (imp) important (nim) not important (utd) unable to determine

3. Responsibility for achievement and failure
   (srp) student responsible (orp) others responsible (utd) unable to determine

4. Attitudes toward study
   (hnf) hard work not fun (hfn) hard work fun (ota) other attitude

Appendix B

Influences on the Coding Instrument

Researcher	Work	Date
Anderson, Reder, & Simon	Applications and misapplications of cognitive psychology to mathematics education	2000
Ginsburg, Leinwand, Anstrom, & Pollock	What the United States can learn from Singapore's world-class mathematics system	2005
Harries & Sutherland	The representation of mathematical concepts in primary mathematics textbooks: A focus on multiplication	2000
Leung	The mathematics classroom in Beijing, Hong Kong, and London	1995
Li	A comparison of problems that follow selected content presentations in American and Chinese mathematics textbooks	2000
Mayer, Sims, & Tajika	A comparison of how textbooks teach mathematical problem solving in Japan and the United States	1995
Mesa	Characterizing practices associated with function in middle school textbooks: An empirical approach	2004
Tang	Textbook illustrations: A cross-cultural study and its implications for teachers of language minority students	1994
Tieso	The effects of grouping practices and curricular adjustments on achievement	2005
Zhu & Fan	An Analysis of the representation of problem types in Chinese and US mathematics textbooks	2004