Low Interrater Agreement on the Semantic Base of Textual Material

Vicki L. Lee, K. Barnett, S. Crisp,
L. Young, M. Bron, P. C. Vatmanidis, and L. Atkinson
Monash University

The present paper describes three experiments which were conducted to determine whether independent raters could agree upon the semantic base of textual materials. These experiments were occasioned by an earlier experiment in which the investigators reported success in increasing the ability of students to extract the semantic base from textual materials. The present paper reports our unsuccessful attempts to obtain an acceptable level of agreement among independent raters about what constituted the semantic base of a number of texts. The paper concludes by raising some doubts about the strategy of extending behavior-analytic research to verbal behavior by combining behavior-change procedures with cognitive constructs.

Studying is a class of self-management behavior which might develop accidentally but is better taught directly (Skinner, 1968, pp. 128-129). Studying includes behaviors that might be prompted by instructors, by rules that constitute advice on how to study, by questions inserted in the text, or by questions which we ask ourselves. The behaviors of answering questions, of writing paraphrases and summaries, of drawing concept maps, of annotating the text, and so on, are related thematically to the text. They can have the effects of adding new intraverbal forms to the student's repertoire, of strengthening existing intraverbals, and of diversifying the discriminative control of existing intraverbals. When thematic prompting is successful, we become able to talk more like the author of the text. In ordinary terms, we come to understand (Skinner, 1953, p. 409; 1957, pp. 277-280; 1968, p. 139) what the text is about. In other cases, we become able to act upon the world in a manner closer to the way in which the author of the text might act upon the world.

Behavior-analytic research on studying as a class of self-management behavior is not easy to find. For the most part, such research has been concerned with methods for training students to follow the rules that comprise the SQ3R strategy or one of its variants (e.g., Alexander, 1985; Fox, 1962; Schumaker, Deshler, Alley, Warner, & Denton, 1982). An exception to this trend is a study by Glover, Zimmer, Filbeck, & Plake (1980) in which a

unit of description, the semantic base, was adopted from the cognitive research literature. Glover et al.'s study provided the occasion for the present investigation and will be described in detail.

In Glover et al.'s (1980) study, college students were asked to underline the semantic base of texts on the natural history of canines, the physical development of felines, and descriptive astronomy. The semantic base was defined by the experimenters as the smallest number of words which, when taken together, would express the meaning of the text. The students completed the underlining task for successive five-line units of the text. Immediately after underlining each unit of text, the subjects compared the words they had underlined with the words that the experimenters had designated as comprising the semantic base of the unit. In addition to this comparison procedure, the students were also required to complete a 10-item multiple choice test on the content of each text.

Considered as a group, Glover et al.'s subjects began by underlining about 20 percent of words selected by the experimenters and about 200 percent of words not selected by the experimenters. After 30 days of training, the subjects underlined about 56 percent of words selected by the experimenters and about 50 percent of words not selected by the experimenters. Immediately after reading each text, subjects were given a 10-item multiple choice test on the content of the text.

Test performance improved from a mean of 56 percent correct in the baseline condition to a mean of 82 percent correct in the last condition of the experiment. In short, changes in the students' underlining choices were associated with an improvement in their test scores.

We set out to replicate the study by Glover *et al.* (1980); specifically, to determine whether we could teach high school students to underline the semantic base of textual materials. However, we ran into difficulties because we could not agree among ourselves about what constituted the semantic base of the textual materials we had chosen to use. At that point, we set aside our initial plan to replicate Glover *et al.*'s study, and we concentrated on the technical problem of determining whether we could develop a procedure that would increase our own level of agreement about what constituted the semantic base of textual materials. The results of this effort led us further to reconsider the conceptual basis of Glover *et al.*'s study, a matter which will be discussed later in this paper.

EXPERIMENT 1

METHOD

Participants

The participants were eight psychology graduates who were enrolled in a fourth-year course in educational psychology. The students undertook the present study as part of their course requirements.

Procedure

The eight participants worked together to develop a definition of the term "semantic base." To this end, they extracted verbatim every statement about the concept of the semantic base that they could find in several papers by Glover and his colleagues (Dinzel & Glover, 1985; Glover, Zimmer, Filbeck, & Plake, 1980; Glover, Zimmer, Filbeck, & Plake, 1982;). These statements were discussed in detail with the aim of developing a clear statement of the meaning of the term "semantic base." (A compilation of these statements is available from the first author.)

Agreement was reached on the following composite definition: "The semantic base consists of the least possible (minimum) number of words which, when taken

together, have the same effect upon the reader (or "meaning") as the original passage." We hoped that this definition would provide the basis for an instruction that would guide consistent underlining by pairs of raters working independently.

After the definition had been formulated, four participants worked together, without the assistance of the other four participants, to develop two sets of textual materials. The *Readers' Digest* was chosen as a source because it contains expository prose that fluent readers can easily read. The participants scanned issues of the *Readers' Digest*, looking for articles written in the third person which had a minimum of direct speech and which were about 800 words in length. The articles selected were on the topics of Euthanasia (Passage 1 from Rea, 1986) and Kookaburras (Passage 2 from Hutchinson, 1985).

To prepare the articles for use in this study, the four participants deleted subheadings, narrative, and unnecessary hyphens (e.g., "per-cent" became "percent"). Other minor changes were made to make the textual materials closer in length to Glover *et al.*'s passages, which had each contained 800 words with a mean 5-line length of 41 words. After the modifications, the passage contained 833 words with a mean of 44 words in the 5-line units, and the Kookaburras passage contained 793 words with a mean of 40 words in the 5-line units.

Each passage was typed in double spacing on a separate sheet of paper. The Kookaburras passage was presented as continuous prose. The Euthanasia passage was presented with each 5-line unit demarcated by a two-line space. Multiple copies of the two original sheets were prepared for use in this study.

While the textual materials were being developed, the other four participants worked together to develop two instructions to guide their own underlining of the passages. One of these instructions was written to prompt a rater to extract the semantic base *from each separate five-line unit* of the text in succession. This instruction read as follows: "(a) Carefully read the passage through first. (b) Underline the least possible (minimum) number of words in each five-line unit which, when taken together, would express the meaning of that unit." The second instruction was written to prompt subjects to extract the

semantic base from a passage as a whole. This instruction read as follows: "(a) Carefully read the passage through first. (b) Underline the least possible (minimum) number of words which when taken together would express the meaning of the passage." (We will refer to these two instructions as the "unit instruction" and as the "whole passage instruction," respectively.)

We developed these two instructions because it was not clear whether the raters in Glover et al.'s study had underlined the semantic base for each 5-line unit in succession or for the passage as a whole. Interrater agreement had been assessed for each 5-line unit, and the subjects had underlined each successive 5-line unit. Therefore it seemed likely that the raters had underlined the words in the same way in which the subject had (i.e., in successive five-line units). Even so, we decided to cover both possibilities in our study, and hence the two instructions.

After the instructions had been written, the four participants who had developed the instructions used them to guide their own underlining of the words in the texts prepared by the other four participants. Although the raters had been involved in preparing the instructions, they had not seen the textual materials before.

The raters sat at separate tables in a classroom. They did not talk to each other or to the other four participants during this part of the procedure. One of the other four participants ("the experimenter") wrote the unit instruction on the board at the front of the room. Subjects were asked to read the instruction and to follow it when they underlined parts of the text. The experimenter told the raters that they were free to take as much time as they wished. The Euthanasia text was then distributed to the raters, and they began the underlining task. When all raters had completed the task, the experimenter collected the sheets of paper from them. The instruction was then erased from the board, and the whole-passage instruction was written on the board, and the raters were asked to read it. The Kookaburras passage was then distributed to the raters, and they completed the underlining task.

Immediately after completing the task, all the participants worked together to count the agreements and disagreements across successive words in each five-line unit of the

passage. This counting was done for each of the six possible pairings of the four raters. Interrater agreement was calculated using the method described by Glover et al. (1980). Total agreements (agreed underlinings and agreed non-underlinings) were divided by the total number of words in each five-line unit.

RESULTS AND DISCUSSION

Tables 1 and 2 show that the mean interrater agreement varied minimally across rater pairs; specifically, from .63 to .69 for the Euthanasia passage and from .62 to .67 for the Kookaburras passage. The upper and lower limits of the range of agreement also showed little variation across subject pairs (Tables 1 and 2). For example, the upper limit of agreement for the Euthanasia passage varied from .88 to .95, and the lower limit for that passage varied from .41 to .50. The results were surprisingly consistent across the two passages (each with a different instruction) and across the six rater pairs.

Table 1
Interrater agreement
("Euthanasia" text).

Rater Pair	Percent Agreement
1	63 (41-93)
2	64 (45-93)
3	67 (50-88)
4	66 (48-95)
5	68 (48-88)
6	69 (50-88)

Table 2
Interrater agreement
("Kookaburras" text).

Rater Pair	Percent Agreement
1	67 (32-92)
2	63 (30-85)
3	63 (38-85)
4	66 (41-82)
5	65 (41-89)
6	62 (36-85)

The results differed from those of Glover et al. (1980) who reported interrater agreement ranging from .70 to 1.00 for five-line units with a mean of .88. Glover et al. defined the semantic base of their textual materials as the words upon which two independent raters had agreed. However, given our much lower interrater agreement, this strategy seemed unacceptable. Therefore, we pursued the task of looking for a method to increase our level of interrater agreement. During a lengthy discussion concerning the possible sources of our relatively low level of interrater agreement, we found that some, but not all, of our raters had proofread their underlining and, when necessary, had edited their underlining so that the underlined words formed a grammatically coherent sequence. This editing was not explicitly required by the instructions, but neither had the instructions proscribed it. We concluded, therefore, that perhaps we needed to rewrite our instructions to eliminate this source of variability across raters.

EXPERIMENT 2

In this experiment, we modified the instructions in an attempt to obtain greater agreement among the raters concerning the words that constituted the semantic base.

METHOD

Participants were the same as in Experiment 1. A new passage ("Salt," by Brody, 1984) was selected and was prepared using the same procedures as before. The passage was presented with each 5-line unit demarcated by a two-line space.

All participants contributed to discussions which resulted in the development of the following instruction: (a) "Carefully read the passage through first. (b) In each five-line unit of the passage, underline the least possible (minimum) number of words which, when taken together, would express the meaning of that unit. (c) After underlining each unit, reread the underlined words to ensure they form a coherent series of words when isolated from the passage. (d) Grammar is important only when it is necessary for the coherence of the underlined words."

The procedure followed for the underlining task was the same as in Experiment 1,

interrater agreement was calculated as before.

RESULTS AND DISCUSSION

Table 3 shows that the mean interrater agreement varied from .63 to .72, that the upper limit of agreement varied from .84 to .98, and that the lower limit varied from .30 to .46. The results indicated that the modified instructions had not increased agreement about which words should be underlined. Subsequent discussion among the participants led to the suggestion that the peculiar density of nouns in the passage might have affected the results. Specifically, the passage contained such strings of nouns as these: "No-one needs to be told that pretzels, potato chips, anchovies, soy sauce and sauerkraut are loaded with salt. But do you think of cheese, cereals, bread, pudding, pancakes, meat, soups, and tinned vegetables?" The discussion indicated much confusion among the raters about how to treat sentences of this kind. Because of this, we concluded that the low interrater agreement could have been due to the peculiar character of the passage.

Table 3
Interrater agreement ("Salt" text).

Rater Pair	Percent Agreement
1	68 (46-85)
2	67 (30-98)
3	65 (31-84)
4	72 (31-87)
5	70 (46-92)
6	63 (32-84)

EXPERIMENT 3

We conducted this experiment in a final attempt to obtain a higher level of interrater agreement. We used a passage previously used in Experiment 1, new raters, and both the initial (Experiment 1) and the extended (Experiment 2) instructions to underline a passage unit-by-unit rather than as a whole.

METHOD

Eight fourth-year psychology students who had not previously been involved in this

study participated as raters. Two sessions were held, and the Euthanasia passage was used in each session. In the first session, four raters were given the instruction used in Experiment 1 which asked them to underline words in each successive five-line unit of the passage. In the second session, the other four raters were given the elaborated unit instruction that had been used in Experiment 2. The procedures were the same as in the earlier experiments.

Table 4
Interrater agreement
("Euthanasia" text).

Rater Pair	Percent Agreement	
	First Session	Second Session
1	63 (38-83)	55 (21-98)
2	51 (38-76)	62 (20-89)
3	52 (28-79)	52 (03-86)
4	59 (23-76)	61 (43-1.00)
5	72 (52-84)	61 (29-88)
6	65 (36-88)	54 (37-86)

RESULTS AND DISCUSSION

Table 4 shows that the mean interrater agreement varied from .51 to .72 for the raters who were given the unelaborated instruction and from .52 to .62 for the subjects who were given the elaborated instruction. For the raters given the unelaborated instruction (first session), the upper limit of agreement varied from .76 to .88 and the lower limit varied from .23 to .52. For the raters given the elaborated instruction (second session), the upper limit varied from .86 to 1.00, and the lower limit varied from .03 to .43. In short, once again our mean interrater agreement was considerably lower than the level of interrater agreement reported by Glover et al.

GENERAL DISCUSSION

An unimpressive level of interrater agreement was found with three different texts, with raters who had and who had not been involved in discussions about the concept of the semantic base, and with both the whole passage and unit instructions. We could not get more than an average of about 63 percent with a range of 3 to 100 percent, in contrast

to Glover et al.'s average agreement of 88 percent with a range of 70 to 100 percent.

This discrepancy might have occurred because of procedural differences between the two studies. Among other things, our textual materials were different from those which Glover et al. had used. However, we did make every effort to equate the length and content of our materials with those of Glover et al.'s. In any case, it seems reasonable to think that if the concept of the semantic base is robust, then we should be able to identify the semantic base of any textual material with a reasonable level of interrater agreement. Perhaps Glover et al. did not fully explicate the instructions that the raters in their study had used to extract the semantic base. Of course, it could be argued that we should have proceeded by defining the semantic base of our materials as those words upon which our raters had agreed. However, our level of agreement was much lower than Glover et al. had reported and, in any case, we were concerned to tackle the task implicitly set by Glover et al. of identifying the semantic base of textual materials.

Our failure to obtain agreement might indicate that the concept of the semantic base is flawed. This possibility led us to examine the concept. Glover et al. equated the semantic base of a text with the major ideas in, or the meaning of, the text. However, from a behavior-analytic perspective, a text does not have a meaning. As Skinner (e.g., 1968, p. 202; 1969, p. 128) has insisted, meaning does not reside in the form of behavior. From this point of view, a text, as a product of behavior, cannot have a single meaning which we can extract. Clarifying this, Skinner (1957, p. 195; 1974, p. 92) has written that the meaning of an utterance (i.e., a form of vocal-verbal behavior) is different for speaker and for listener. The meaning for the speaker is found in the conditions under which the speaker speaks, and the meaning for the listener is found in how the listener acts upon what is said. The same can be said for the writer and the reader and, thus, for the meaning of a text. Even if we insist that texts do have intrinsic meanings, it seems unlikely that we would be able to identify a single meaning that will have unity across readers irrespective of their past histories with respect to contingencies that have included textual and other verbal stimuli. The present

difficulty is therefore not surprising. We did try to maximize the chance of agreement among the raters, but each rater came to the task with a different history, both with respect to reading and underlining textual material and, presumably, with respect to the particular topics chosen for underlining.

Glover *et al.* concluded that their study suggested that they had modified their subjects' ability to underline the semantic base of prose passages. The present considerations suggest that a more cautious conclusion should be drawn from Glover *et al.*'s study. It seems more defensible to say that when subjects were given feedback consisting of the words selected by the experimenters, then the subjects' underlining became more like the experimenters' underlining. This conclusion is more closely tied to the events, and it avoids the confusion brought about by the notion that textual materials have a semantic base. Indeed, it might reasonably be said that this latter notion only obscures the events and contingencies that provided the data base in the study by Glover *et al.* Any subsequent attempt to replicate Glover *et al.*'s procedure might better be described explicitly as a procedure in which an expert learner selects the words to be underlined and composes the test items used to probe the students' verbal repertoires.

The question of interest in such research would be whether we can (a) instruct subjects to underline as an expert learner does and (b) thus improve the subjects' recall on test items composed by the expert learner. That is, the issue is in part whether the mismatch between experimenter- and subject-underlining evident during feedback can exert instructional control (e.g., Catania, 1985, 1986; Cerutti, 1989) over subsequent underlining behavior, or whether the feedback procedure will select subject underlining that is closer in its formal result to expert underlining, or both. As well, the issue is in part whether this procedure combined with exposure to successive tests designed by the expert learner will change how the student acts with respect to subsequent tests. This formulation of the problem would not distract us from the events and contingencies (Buskist & DeGrandpre, 1989) that comprise the phenomena under investigation. We would also want to consider how underlining might

facilitate subsequent performance on text-related probes. If so, then we will have to develop a functional interpretation of reading and studying, because it is likely that underlining is important in part because it occasions thematic elaboration and rehearsal. In this connection, we might do well further to develop the functional interpretation of studying (Skinner, 1953, p. 253; 1957, pp. 415-416; 1968, pp. 127-131) and to proceed empirically on the basis of it. This theoretically-driven strategy might be more fruitful in developing knowledge about the emergence of intraverbal repertoires than is the strategy of attempting to develop a behavior analysis of studying by combining behavior-change procedures with constructs developed elsewhere and for purposes other than the functional analysis of behavior.

REFERENCES

- Alexander, D. F. (1985). The effect of study skill training on learning disabled students' retelling of expository material. *Journal of Applied Behavior Analysis*, 18, 263-267.
- Brody, D. (1984) Licking the salt habit. *Readers' Digest*, 124, 742, 35-38.
- Buskist, W., & DeGrandpre, R. J. (1989). The myth of rule-governed behavior. *Experimental Analysis of Human Behavior Bulletin*, 7, 4-6.
- Catania, A. C. (1985). Rule-governed behavior and the origins of language. In C. F. Lowe, M. Richelle, D. E. Blackman and C. M. Bradshaw (Eds.), *Behavior analysis and contemporary psychology* (pp. 135-156). London: Erlbaum.
- Catania, A. C. (1986). On the difference between verbal and nonverbal behavior. *The Analysis of Verbal Behavior*, 4, 2-9.
- Cerutti, D. T. (1989). Discrimination theory of rule-governed behavior. *Journal of the Experimental Analysis of Behavior*, 51, 259-276.
- Dinnel, D., & Glover, J. A. (1985). Advance organizers: Encoding manipulations. *Journal of Educational Psychology*, 77, 514-521.
- Fox, L. (1962). Effecting the use of efficient study habits. *Journal of Mathematics*, 1, 75-86.
- Glover, J. A., Zimmer, J. W., Filbeck, R. W., & Plake, B. S. (1980). Effects of training students to identify the semantic base of prose materials. *Journal of Applied Behavior Analysis*, 13, 655-667.
- Glover, J. A., Zimmer, J. W., Filbeck, R. W., & Plake, B. S. (1982). Training subjects to utilize adjunct reading aids. *Journal of General Psychology*, 107, 3-10.
- Hutchinson, J. (1985). What makes the kookaburra laugh? *Readers' Digest*, 127, 761, 67-70.
- Rea, P. (1986). Who lives, who dies—Should doctors decide? *Readers' Digest*, 128, 774, 44-48.
- Schumaker, J. B., Deshler, D. D., Alley, G. R., Warner, M. M., & Denton, P. H. (1982). Multipass: A learning strategy for improving reading comprehension. *Learning Disability Quarterly*, 5, 295-304.

- Skinner, B. F. (1953). *Science and human behavior*. New York: Free Press.
- Skinner, B. F. (1957). *Verbal behavior*. New York: Appleton-Century-Crofts.
- Skinner, B. F. (1968). *The technology of teaching*. New York:

Appleton-Century-Crofts.

- Skinner, B. F. (1969). *Contingencies of reinforcement: A theoretical analysis*. New York: Appleton-Century-Crofts.
- Skinner, B. F. (1974). *About behaviourism*. London: Jonathan Cape.