The different effects of thinking aloud and writing on graph comprehension

Nadia Ali (n.ali@hud.ac.uk)

David Peebles (d.peebles@hud.ac.uk)
Department of Behavioural and Social Sciences,
University of Huddersfield,
Queensgate, Huddersfield, HD13DH, UK

Abstract

We report an experiment which seeks to determine how novice users' conceptual understanding of graphs differs depending on the nature of the interaction with them. Undergraduate psychology students were asked to interpret three-variable "interaction" data in either bar or line graph form and were required to either think aloud while doing so or to produce written interpretations. Analysis of the verbal protocols and written interpretations showed that producing a written interpretation revealed significantly higher levels of comprehension than interpreting them while thinking aloud. Specifically, a significant proportion of line graph users in the verbal protocol condition was either unable to interpret the graphs, or misinterpreted information presented in them. The occurrence of these errors was substantially lower for the bar graph users in the verbal protocol condition. In contrast, analysis of the written condition revealed no significant difference in the level of comprehension between the two graph types. Possible explanations for these findings are discussed.

Keywords: Graph comprehension, diagrammatic reasoning, verbal protocols, writing to learn.

Introduction

Various tasks have been used to assess the different aspects of graph comprehension and use. Zacks and Tversky (1999) for example presented participants with bar or line graphs together with the instruction "Please describe in a sentence the relationship shown in this graph" (or a simpler variant) and required written responses, or in one experiment to draw graphs from written descriptions of data (e.g., "Height for 12-year-olds is greater than for 10-year-olds"). Shah and Carpenter (1995) asked participants to produce verbal descriptions of the graphs they saw, and to then either compare them to other graphs, reproduce them, or provide possible explanations for the data depicted. In a subsequent experiment Carpenter and Shah (1998) asked people to answer specific questions about relationships depicted (e.g., "What happens to vocabulary score as age increases?") while their eye movements were recorded. Other researchers have recorded participants' verbal protocols while they carry out various graph related tasks (e.g., Ratwani, Trafton, & Boehm-Davis, 2008)

Ericsson and Simon (1993) proposed the verbal protocol method as a means of tracing cognitive processes. One type of verbal protocol involves "thinking aloud", a process-tracing method in which subjects report their thoughts continuously whilst attempting to complete a task. This method can yield important information about the steps of problem solving that would be difficult, if not impossible, to observe using other measures, (e.g., the contents of working memory) which can suggest hypotheses concerning the strategies employed (A. Newell & Simon, 1972; Larkin, McDermott, Simon, & Simon, 1980; Koedinger & Anderson, 1990).

Since the original proposal, the think aloud method has been widely adopted, resulting in a large body of research into the processes underlying decision making, problem solving, text comprehension, diagrammatic reasoning, writing, and various other tasks (Crutcher, 1994).

An alternative approach to assessing conceptual understanding of material is to provide subjects with specific questions or goals and require them to produce written responses. Because the data are limited to the final written output, process tracing is not possible with this method alone. Written responses can yield rich information however which may be used to infer strategy choice in some cases (e.g., where different strategies result in different written responses).

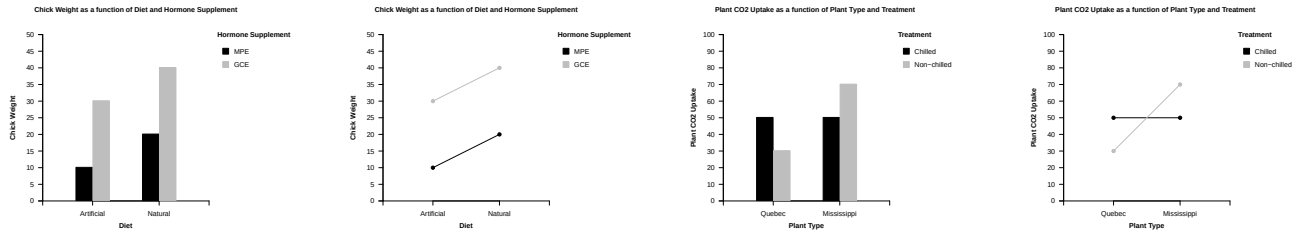
Some researchers have attempted to employ writing as a process tracing method by requiring people to write down everything that comes to mind (Pugalee, 2004) while others have used both methods simultaneously (e.g., Flower & Hayes, 1981)

There is a large body of literature investigating whether writing improves conceptual understanding of material in a number of disciplines (e.g., Britton, 1978; Young & Sullivan, 1984; G. E. Newell, 1984; Bereiter & Scardamalia, 1987; Flower & Hayes, 1980). This research comes under the umbrella of "writing to learn" and advocates assert that writing can help engender critical thinking and the formation of new relationships between ideas, leading to knowledge construction (Klein, 1999).

Several processes involved in writing have been identified as possible causes for these observed improvements in conceptual understanding. For example, the self-paced nature of writing allows for reflection (Emig, 1977; Ong, 1982) while the permanence of the text allows material to be reviewed (Emig, 1977; Young & Sullivan, 1984). The process of reviewing allows the writer to judge what is written against what is intended to be communicated and to evaluate (and improve) the logical coherence of sets of sentences within the text (Galbraith, 1992).

Furthermore, the context in which writing is produced can result in improved conceptual understanding of material. For example, the absence of an immediate audience requires writers to be explicit in their interpretation and presentation of material (Olson, 1977).

In contrast, according to Ericsson and Simon (1993), the think aloud method allows access to participants' short-term



(a) Bar graph 1 (b) Line graph 1 (c) Bar graph 3 (d) Line graph 3

Figure 1: Example bar and line graphs used in the experiment.

memory stream, and verbalisations uncover cognitive processes involved in task completion. According to their theory of protocol generation, the act of thinking aloud concurrently during a task should neither impair nor enhance performance as participants are simply verbalising their thought processes.

When attempting to compare and evaluate performance with different graphical formats, it is essential to have a set of behavioural criteria or categories with which to do so. From the considerable number of studies conducted into graph comprehension a consensus has emerged on the broad three-level taxonomy of skills required for the task. In a review of five studies, Friel, Curcio, and Bright (2001) characterised the three levels as *elementary*, *intermediate*, and *advanced* (or more descriptively as "read the data", "read between the data" and "read beyond the data" respectively).

In previous studies (Peebles & Ali, 2009; Ali & Peebles, submitted) we have demonstrated that undergraduate college students' ability to understand statistical data can vary significantly depending on the form of the graphical display. Specifically, we have shown that for a considerable number of students, conceptual understanding of three variable line graphs of the type shown in Figure 1 does not meet the lowest level of graph comprehension ability identified in the literature.

Peebles and Ali (2009) conducted an experiment in which people were asked to interpret informationally equivalent bar or line graphs representing two-way factorial design data as fully as possible while thinking aloud. Analysis of the verbal protocols revealed significant differences in how people interpreted the two graph formats. It was found that 39% of line graph users were either unable to interpret the graphs, or misinterpreted information presented in them. No bar graph users performed at this level. This finding led us to propose a fourth, lower category of comprehension ability which we termed "pre-elementary" and subsequently to propose and test a novel line graph design which we found successfully reduces the error level to that of the bar graphs (Ali & Peebles, submitted).

Developing a adequate model of diagrammatic reasoning requires taking into account three interacting factors: the nature of the graphical representation, the characteristics of the user and the nature of the task. Our previous work explored the role of graphical features in comprehension performance. The aim of this study is to determine how, given the same open-ended task (try to understand what the graph is portraying), the nature of the interaction can also significantly affect performance. Specifically, we seek to determine whether the reduction in performance found in novice line graph users may be partially accounted for by the additional cognitive demands imposed by producing a think aloud protocol and whether this may be mitigated by engaging in a different way.

Experiment

Method

Participants Sixty-five undergraduate psychology students (54 female, 11 male) from the University of Huddersfield were paid £5 (approximately $8) in grocery store vouchers to take part in the experiment. The age of participants ranged from 18.5 to 39.5 years with a mean of 21.5 years ($SD = 3.82$). All participants were in their first year of a three-year psychology degree.

Design The experiment was an independent groups design with two between-subject variables: type of diagram used (bar or line graph) and methodology employed (think aloud or written responses). Sixty-five participants were randomly allocated to each condition. There were 14 participants in the verbal protocol bar condition, 16 in the written bar condition, 15 in the verbal protocol line condition and 20 in the written line condition.

Materials The stimuli used were six bar and six line three-variable interaction graphs depicting a wide range of (fictional) content. The graphs were generated using the *PASW Statistics* software package (produced by SPSS Inc.). Examples are shown in Figure 1.

The bar and line graphs were constructed from the same six data sets (the variables of which are shown in Table 1). The numerical values for the variables were selected in order to provide the range of effects, interactions and other relationships between three variables commonly encountered

Table 1: Variables of the six graphs used in the experiment

Graph Number	Dependent Variable	Scale Range (Increment)	Independent Variable 1 (Levels)	Independent Variable 2 (Levels)
1	Chick weight	0–50 (5)	Diet (Artificial, Natural)	Hormone Supplement (MPE, GCE)
2	Maize yield	0–10 (1)	Plant Density (Low, High)	Nitrogen level (Low, High)
3	Plant CO_2 uptake	0–100 (10)	Plant type (Quebec, Mississippi)	Treatment (Chilled, Non-chilled)
4	Cutting tool wear	0–10 (1)	Rock Type (Limestone, Granite)	Diamond type (Bead, Wire)
5	Fixtural strength	0–1000 (100)	Cement type (Monochem, Bischem)	Curing method (Photocuring, Autocuring)
6	Dopamine activity	0–500 (50)	Rat breed (Lewis, Fischer)	Brain region (SNC, VTA)

in these designs (typically depicted in line graphs as parallel, crossed and converging lines, one horizontal line and one sloped line, two lines sloping at different angles, etc.). Stimuli were printed in colour (with the levels of legend variable in blue and green) on white A4-sized paper

Procedure Participants were instructed that they would see six graphs and that their task was to try to understand each graph as fully as possible whilst writing their response down or thinking aloud. They were instructed to write or talk aloud about the relationships each graph was showing, until they felt they had provided as much detail as they could.

The instructions drew attention to the fact that the graphs may depict more than one relationship, and that participants should imagine they are in an exam in which more detailed interpretations produced higher scores. In order to produce as close a similarity as possible to the think aloud condition, participants in the written condition were also encouraged to write down their thoughts as they went along.

In the written condition the six stimuli were compiled as a booklet with graph pages interleaved with blank paper for writing. Participants completed these under the supervision of the experimenter. In the verbal condition the graphs were handed over to participants one at a time for them to interpret while their verbal protocols were recorded using a portable digital audio recorder. Stimuli were presented in random order and all participants were informed that there was no time limit to the task.

Results

The verbal condition participants' protocols were transcribed and the content of the transcriptions and the statements from the written condition participants were analysed. Only statements in which a sufficient number of concepts could be identified were included for analysis. For example, the statement "Chick weight is higher for the GCE hormone supplement than for the MPE supplement" was included whereas "Chick weight is higher when…um…I'm not sure" was not.

Data analysis was conducted according to the procedure and criteria employed in our previous studies (Peebles & Ali, 2009; Ali & Peebles, submitted). For each trial, the participant's statements were analysed against the state of affairs represented by the graph. If a participant made a series of incorrect statements that were not subsequently corrected, then the trial was classified as an 'incorrect interpretation'. If the participant's statements were all true of the graph or if an incorrect interpretation was followed by a correct one however, then the trial was classified as an 'correct interpretation'. In this way, each participant's trials were coded as either being correctly or incorrectly interpreted.

The statements for each trial were initially scored as being either a correct or incorrect interpretation by the first author and a sample (approximately 25% from each graph type) was independently scored by the second author. The level of agreement between the two coders was approximately 96% ($\kappa = 0.91$). When disagreements were found the raters came to a consensus as to the correct code.

This measure was then used as the basis for subsequent categorisation into elementary and pre-elementary groups. For the purpose of our analysis, we classified participants as pre-elementary for their graph type if they interpreted 50% or more trials incorrectly (i.e., at least three of the graphs were classified as incorrect interpretations). This criterion was considered appropriate because it indicates that the user is unable to produce an accurate description of the data (even such basic information as point values) after at least two previous encounters with the same graph type—suggesting a lack of understanding of the basic representational features of the format (rather than just the content of the graph) and resulting in comprehension performance that does not meet elementary level criteria (Friel et al., 2001).

Figure 2 displays the proportion of bar and line users in the verbal protocol condition in each of the three performance categories. In this condition the difference between bar and line graph users emerged as predicted; 60% of participants were classified as pre-elementary in the line graph condition

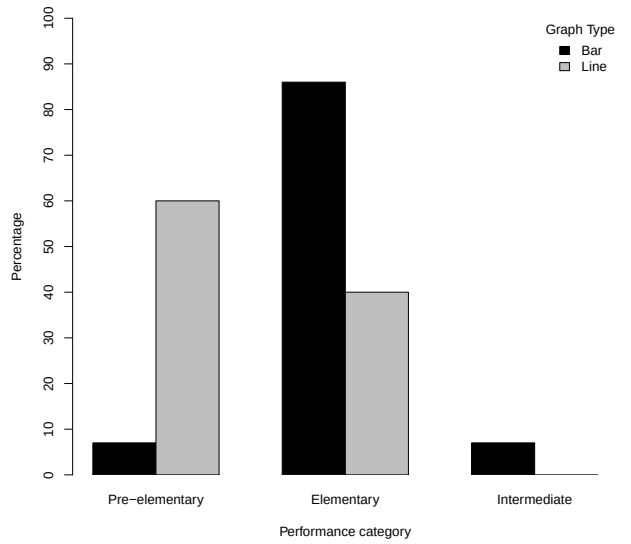


Figure 2: Percentage of bar and line graph users in the three performance categories, verbal protocol condition.

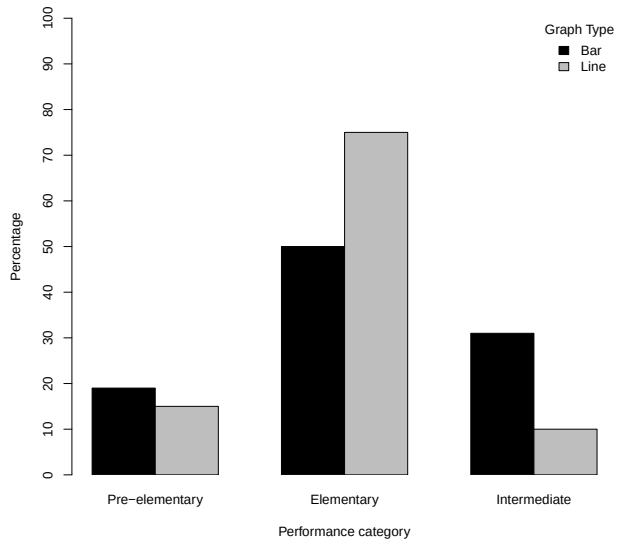


Figure 3: Percentage of bar and line graph users in the three performance categories, written condition.

compared to 7% in the bar graph condition. A chi-squared test revealed that this difference was statistically significant ($\chi^2 = .819$, $df = 1$, $p < .01$), replicating the result of the original Peebles and Ali (2009) and Ali and Peebles (submitted) experiments.

As Figure 3 shows however, this large difference in pre-elementary performance is not found in the written condition, with the number of participants classified as pre-elementary being roughly equal between the line and bar graph formats (15% and 19% respectively). A Fischer's Exact test revealed that this difference was not significant ($\chi^2 = .09$, $df = 1$, $p = 1.0$).

To eliminate the possibility that these results may arise as an artifact of our classification system we also analysed the number of correct trials for each condition. This revealed that for the think aloud condition the mean ranks for the line graphs (11.0) was significantly lower than for the bar graph condition (19.29) $U = 45$, $z = 2.69$, $p < .01$. In the written condition, the mean ranks for the line and bar graphs were much closer (bar = 19.41, line = 18.69) and so there was no significant difference in number of correct trials between them ($U = 161.5$, $z = .212$, $p = .84$). There was also no difference in the number of correct trials between the bar graph users in the written (mean ranks = 16.16) and verbal (mean ranks = 14.75) conditions, $U = 141.5$, $z = .461$, $p = .667$.

A significant interaction was found between graph format and methodology (shown in Figure 4). Thinking aloud significantly reduced the comprehension of line graphs—but not bar graphs—compared to producing written interpretations $U = 69.0$, $z = 2.91$, $p < .01$.

Discussion

The results of our experiment reveal a remarkable interaction of methodology employed to assess graph comprehension and graph format. Consistent with the results of our previous experiments (Peebles & Ali, 2009; Ali & Peebles, submitted), a significant proportion of line graph users was classified as pre-elementary compared to bar graph viewers in the think aloud condition.

We have explained this effect using Gestalt principles of perceptual organisation (Ali & Peebles, submitted). In the line graph diagram, data points are connected by a line, resulting in two lines at the centre of the display. Pre-elementary line graph users were unable to integrate the information, primarily because they ignored the x variable entirely. This pattern of errors indicates that the salience of the lines is such that it draws users' attention to them and then—through a process of colour matching—to the legend variable, which they then try to interpret. Because they are focusing on the lines however, they are less able to identify the points at the ends of the lines and interpret them as discrete values associated with levels of the x variable.

In the bar graph however, each level of the legend variable is depicted by a bar projecting from the x axis. This allows participants to match bar colour to the appropriate variable level in the legend but, crucially, because the bars are located directly above the value labels on the x axis, also more easily identify the associated x variable levels. The results of this experiment reveal that this balances out attention to the two IVs and promotes a richer understanding of the relationship between them.

However, this effect does not emerge in the written re-

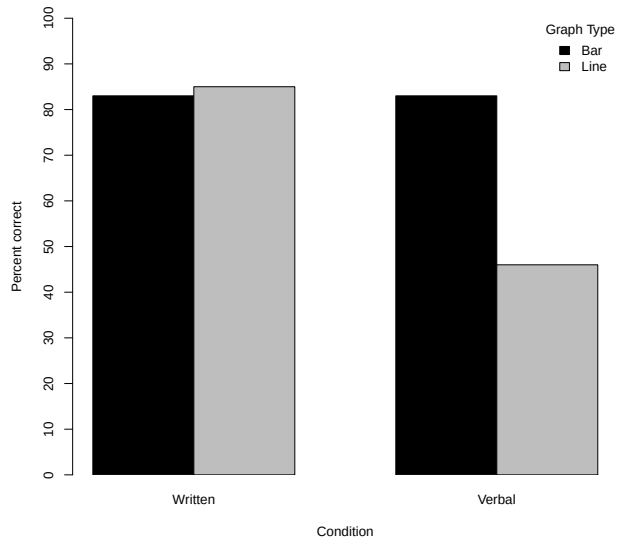


Figure 4: Percentage of correct trials for bar and line graphs in the verbal protocol and written conditions.

sponse condition. Despite the imbalance of Gestalt features associating the pattern to referents, the majority of graph readers demonstrate conceptual understanding of both graph formats at an elementary (and in a few cases intermediate) level. Our results therefore reveal that the reduction in line graph comprehension can be mitigated by changing the type of interaction the user is required to engage in. Being required to write down your understanding of the graph eliminates the overwhelming salience of the lines in the line graphs which produces the pre-elementary performance when also thinking aloud (Friel et al., 2001; Peebles & Ali, 2009).

There are a number of potential competing explanations for why this difference in conceptual understanding in the line graph condition emerges between the two methodologies. Firstly, a number of researchers have noted the potential problems with employing the verbal protocol method to investigate underlying cognitive processes. There is an ongoing debate concerning whether thinking aloud is *reactive* (i.e., alters other cognitive processes). Reactivity can result in either an improvement or a deterioration in task performance. The question of whether producing a verbal protocol is reactive is a complex one however and the current conclusion appears to be that the demands of verbalisation can interact with task demands to affect output in at least some cases (Russo, Johnson, & Stephens, 1989; Schooler, Ohlsson, & Brooks, 1993).

In terms of the detriment in performance, researchers have argued that the additional demands for processing resources (which occurs when individuals are required to verbalise whilst performing a task) can explain this form of reactivity. In order to deal with these additional demands, participants can draw upon any unused resources which are not being employed by the task. When the demands of the task exceed processing resources however, reactivity effects can occur, resulting in a detriment in performance due to the resources being divided between completing the task and verbalising throughout (Russo et al., 1989; Wilson & Schooler, 1991).

Alternatively, verbal protocols could be providing an accurate reflection of underlying cognitive processes, and cognitive processes involved in writing could be facilitating task performance. The writing to learn literature argues that various factors unique to written assessments can result in improved conceptual understanding of material under scrutiny. For example, absence of an audience requires writers to be explicit in their interpretation of material and permanence of text allows them to review their ideas (Applebee, 1984; Klein, 1999).

Although the findings appear to be inconsistent (Ackerman, 1993), Tynjälä (1999) explains these conflicting findings as resulting from differing tasks demands. If the task simply involves learning factual knowledge, then a passive method such as reading text will not be affected by writing (Penrose, 1992). If higher-order thinking is required however, writing can result in learning gains. For example Tynjälä (1999) argues that, generally, writing is an effective learning tool when attempting to advance students' understanding and critical thinking skills, but not superior to any other method when students are required to simply "tell what they know".

In a similar vein, the second factor that can explain the conflicting results is how much information manipulation is required by the task. The larger the demands of manipulation of information are, the stronger the learning effects should be (e.g., Applebee, 1984; Greene & Ackerman, 1995; Langer, 1986; Tynjälä, 1999).

The current experiment does not allow us to differentiate between these competing explanations, and further empirical work is required to isolate and test the alternative hypotheses. What these initial results do suggest however is that researchers should take great care when deciding which methodology to employ to assess conceptual understanding.

Researchers make use of both methods to assess graph comprehension (e.g., Shah & Carpenter, 1995; Carpenter & Shah, 1998; Shah & Freedman, 2009), often interchangeably. This study demonstrates that even for what may superficially seem to be the same task, the precise details of the interaction can significantly affect performance.

References

Ackerman, J. M. (1993). The promise of writing to learn. *Written Communication*, *10*, 334–370.

Ali, N., & Peebles, D. (submitted). *The effect of Gestalt laws of perceptual organisation on the comprehension of three-variable bar and line graphs.*

Applebee, A. N. (1984). Writing and reasoning. *Review of Educational Research*, *54*(4), 577–596.

Bereiter, C., & Scardamalia, M. (1987). *The psychology of*

written composition. Hillsdale, NJ: Lawrence Erlbaum Associates.

Britton, J. (1978). The composing processes and the functions of writing. In C. R. Cooper & L. Odell (Eds.), *Research on composing: Points of departure* (pp. 13–28). Urbana, IL: NCTE.

Carpenter, P. A., & Shah, P. (1998). A model of the perceptual and conceptual processes in graph comprehension. *Journal of Experimental Psychology: Applied*, *4*, 75–100.

Crutcher, R. J. (1994). Telling what we know: The use of verbal report methodologies in psychological research. *Psychological Science*, *5*, 241–244.

Emig, J. (1977). Writing as a mode of learning. *College Composition and Communication*, *28*, 122–128.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data* (Revised ed.). Cambridge, Mass: MIT Press.

Flower, L., & Hayes, J. R. (1980). The cognition of discovery: Defining a rhetorical problem. *College Composition and Communication*, *31*, 21–32.

Flower, L., & Hayes, J. R. (1981). A cognitive process theory of writing. *College Composition and Communication*, *32*, 365–387.

Friel, S. N., Curcio, F. R., & Bright, G. W. (2001). Making sense of graphs: Critical factors influencing comprehension and instructional implications. *Journal for Research in Mathematics Education*, *32*, 124–158.

Galbraith, D. (1992). Conditions for discovery through writing. *Instructional Science*, *21*, 45–72.

Greene s., & Ackerman, J. M. (1995). Expanding the constructivist metaphor: A rhetorical perspective on literary practice. *Review of Educational Research*, *65*(4), 383–420.

Klein, P. D. (1999). Reopening inquiry into cognitive processes in writing-to-learn. *Educational Psychology Review*, *11*, 203-270.

Koedinger, K. R., & Anderson, J. R. (1990). Abstract planning and perceptual chunks: Elements of expertise in geometry. *Cognitive Science*, *14*, 511–550.

Langer, J. A. (1986). *Children reading and writing: Structures and strategies*. Norwood, NJ: Ablex.

Larkin, J. H., McDermott, J., Simon, D. P., & Simon, H. A. (1980). Expert and novice performance in solving physics problems. *Science*, *208*, 1335–1342.

Newell, A., & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice Hall.

Newell, G. E. (1984). Learning from writing in two content areas: A case study/protocol analysis. *Research in the Teaching of English*, *18*, 265–287.

Olson, D. R. (1977). From utterance to text: The bias of language in speech and writing. *Harvard Educational Review*, *47*, 257–281.

Ong, W. J. (1982). *Orality and literacy*. New York: Methuen Inc.

Peebles, D., & Ali, N. (2009). Differences in comprehensibility between three-variable bar and line graphs. In *Proceedings of the thirty-first annual conference of the cognitive science society* (pp. 2938–2943). Mahwah, NJ: Lawrence Erlbaum Associates.

Penrose, A. M. (1992). To write or not to write: Effects of task and task interpretation on learning through writing. *Written Communication*, *9*, 465–500.

Pugalee, D. (2004). A comparison of verbal and written descriptions of students' problem solving processes. *Educational Studies in Mathematics*, *55*, 27–47.

Ratwani, R. M., Trafton, J. G., & Boehm-Davis, D. A. (2008). Thinking graphically: Connecting vision and cognition during graph comprehension. *Journal of Experimental Psychology: Applied*, *14*(1), 36–49.

Russo, J. E., Johnson, E. J., & Stephens, D. L. (1989). The validity of verbal protocols. *Memory and Cognition*, *17*, 759–769.

Schooler, J. W., Ohlsson, S., & Brooks, K. (1993). Thoughts beyond words: When language overshadows insight. *Journal of Experimental Psychology: General*, *122*, 166–183.

Shah, P., & Carpenter, P. A. (1995). Conceptual limitations in comprehending line graphs. *Journal of Experimental Psychology: General*, *124*, 43–62.

Shah, P., & Freedman, E. (2009). Bar and line graph comprehension: An interaction of top-down and bottom-up processes. *Topics in Cognitive Science*, *doi: 10.1111/j.1756-8765.2009.01066.x*.

Tynjälä, P. (1999). Towards expert knowledge? A comparison between a constructivist and a traditional learning environment in university. *International Journal of Educational Research*, *31*(5), 357–444.

Wilson, T. D., & Schooler, J. W. (1991). Thinking too much: Introspection can reduce the quality of preferences and decisions. *Journal of Personality and Social Psychology*, *60*, 181–192.

Young, R., & Sullivan, P. (1984). Why write? A reconsideration. In R. J. Conners, L. S. Ede, & A. A. Lunsford (Eds.), *Essays on classical rhetoric and modern discourse* (pp. 215–225). Carbondale, IL: Southern Illinois University Press.

Zacks, J., & Tversky, B. (1999). Bars and lines: A study of graphic communication. *Memory and Cognition*, *27*(6), 1073–1079.