Graph Structure Supports Graph Description

Irvin R. Katz (ikatz@ets.org)
Center for New Constructs, Educational Testing Service
Princeton, NJ 08541 USA

Xiaoming Xi (xxm@ucla.edu)
Dept. of Applied Linguistics & TESL, Univ. of California
Los Angeles, CA 90095 USA

Hyun-Joo Kim (hk312@columbia.edu)
Applied Linguistics Program, Columbia University
New York, NY 10027-6696 USA

Peter C-H. Cheng (peter.cheng@nottingham.ac.uk)
School of Psychology, University of Nottingham
University Park, Nottingham, NG7 2RD U.K.

Abstract

This research applies theories of graph comprehension to investigate the factors affecting how easily a graph can be described. We find that the structure of a graph– the number of *visual chunks* (Shah, Mayer, & Hegarty, 1999) to be described – influences the communicative quality of elicited descriptions. The work extends our understanding of graph comprehension by investigating the relationship between comprehension and description processes. This research occurs in the context of understanding how to design graphical description tasks for the Test of Spoken English.

Introduction

Graphs are a ubiquitous communication tool. Instructors describe graphs to communicate concepts, perhaps requiring students to uncover a graph's main point. A doctor might describe a graph to a patient to make a point about treatment ("see how your cholesterol level has been decreasing since you began the new diet?"). Yet we know little about the cognitive processes engaged when people describe a graph. Research on graph description can contribute to our understanding of how people integrate visual and verbal information in the performance of everyday tasks. From a practical standpoint, such research can provide guidelines for designing graphs that facilitate description.

Instead, much of the research on graphs has focused on graph comprehension – how we encode and interpret elements of a graph to draw out key pieces of information (Carpenter & Shah, 1998; Lohse, 1993; Pinker, 1990), typically in response to proscribed tasks (e.g., "Who had a greater market share in 1983?"). The few studies that investigate spontaneous descriptions of graphs have focused on what is described (e.g., global trends vs. local, piecemeal descriptions [Carswell, 1993; Carswell et al., 1998]; trends vs. comparisons [Zacks & Tversky, 1999]) and the organization of the descriptions (Shah, Mayer, & Hegarty, 1999; see below) rather than on the *communicative quality* of the description. One reason for this oversight might be the lack of a rigorous measure of communicative quality.

In the work presented herein, we apply a theory of graph comprehension to predict the characteristics of graphs that facilitate descriptive communication. To measure the quality of descriptions produced by alternative graphs, we use a theoretically grounded and empirically validated measure of communicative quality: the scoring rubric from the Test of Spoken English (TSE®).

The next section provides some background on the TSE, its scoring rubric, and the real-world problem that motivated this research.

The Test of Spoken English

The real-world problem

The goal of the Test of Spoken English (TSE) is to measure a test-taker's communicative competence in Northern American English. It is taken by approximately 20,000 non-U.S. citizens each year, who are seeking to be teaching assistants or healthcare professionals in the U.S. The test consists of 12 questions that elicit a range of communication functions (e.g., describe, compare, state opinion). The questions are presented in a booklet and aurally by a taped interviewer; test-takers' spoken responses are recorded. Responses are scored by trained raters employing a well-defined scoring rubric (see below).

One question (illustrated in Figure 1) prompts for a description of a statistical graph. Test-takers are given one minute to respond. The task mirrors the type of communication using graphs done by teaching assistants and healthcare professionals.

The graph below shows what people of two age groups value about their work. Describe the information given in the graph.

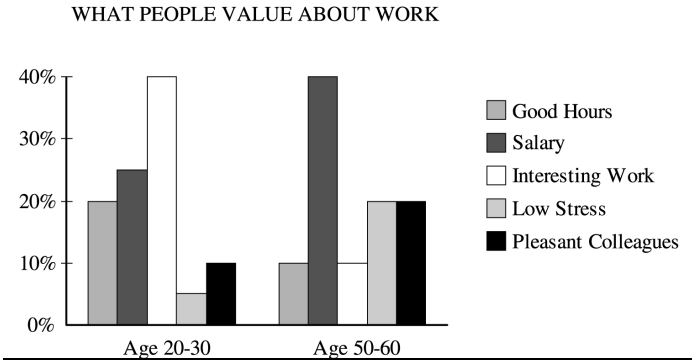

Figure 1: Illustrative TSE graph question [Fewer visual chunks].

This type of graph-description question occasionally poses problems for TSE scoring. According to the raters, certain graphs elicit speech that displays a lower ability in

English than would be expected based on responses to all other test questions. However, many graphs evidenced no such difficulties. These reported rating difficulties are typically not detectible through statistical analyses of scores (e.g., Myford & Wolfe, 2000).

The issue of what characteristics of a graph lead to descriptions that communicate better is critical to the TSE. If a graph is hard to describe, it might give an unfair advantage to test-takers with better graph-reading skills (i.e., a more sophisticated "graph schema"; Pinker, 1990), who can make sense of poorly constructed graphs. A test-taker's ability to read and interpret graphs should not influence their score on a graph question. Indeed, the accuracy of a person's response to a graph item is not considered in the score, only the degree to which the person evidences certain competencies associated with spoken English.

The challenge is to create graphs that contain enough information so as not to trivialize the description (which would eliminate any differences between test-takers) yet are straightforward to describe, allowing a test-taker to show off his/her communicative skill without other factors getting in the way. Ultimately we seek to develop guidelines for the development of graph questions that validly measure communicative competence.

TSE Scoring Rubric

Responses to TSE prompts are scored according to the published "TSE SCORE BAND DESCRIPTOR CHART" (TOEFL, 2001). This scoring rubric defines four key communicative competencies: discourse, functional, sociolinguistic, and linguistic competence. The chart also specifies the types of response characteristics for these competencies at each of the five possible score levels (20, 30, 40, 50, and 60). Although these several competencies are considered during scoring, each response receives a single, holistic score representing the raters' judgment of which score band level was best evidenced in the response. The score band chart and associated training materials were developed based on research into the components of communicative competence (Myford & Wolfe, 2000; Powers, Schedl, Wilson-Leung, & Butler, 1999).

Two communicative competencies are particularly relevant to the issue of graph comprehension: discourse competence and functional competence.

Discourse competence relates to the coherence and cohesiveness of a response. Is the response well organized and well developed, and does the speaker cue the listener to the organization (e.g., "First we see that…," "In contrast…")? For the graph in Figure 1, a partial response demonstrating low discourse competence is: (ellipses refer to short pauses in speech)

> *the good hours...ah for age...ah,...between age ...fifty and sixty is ten percent....And...the pleasant ...colleagues...for ...ah,...for age ...twenty to thirty ...is ten percent ...and ...ah, for ...fifty to sixty is twenty percent....*

Responses low in discourse competence tend to be list-like, consisting of phrases connected by "and" but showing neither a strong organizing structure nor development. A response showing stronger discourse competence is:

> *…for adults...uh,...between age two,...twenty to thirty,...they value interesting work as their most important thing....well...for the old man...that's not important....Other points I should compare is uh,...is the low stress ...for the old man they...they prefer low stress and...while for the younger men …*

This response guides the listener better by using phrases such as "for the old man…" and "Other points I should compare…".

Functional competence is using language to transfer information and ideas to accomplish a goal. Does the person communicate? For example, we all know people who "beat around the bush" while you are wondering when they will get to their point. For the graph in Figure 1, a partial response demonstrating low functional competence is:

> *Ok, people ...around the age ...twenty to thirty...I guess started like ...ah, ...just youngsters, ...they are...um...they good hours up like twenty percent ...and ...only...ah, ...at the age of twenty to thirty ...the people who are interested ...are only forty percent*

Based on this response, it is difficult to understand what information was provided in the graph, partially because the speaker misrepresents the meaning of "good hours" and "interesting work." The first response, in contrast, does a good job of describing the information and so was rated higher on functional competence than was the second response.

The other two competencies appear less likely to be affected by the particular characteristics of a graph. **Sociolinguistic competence** refers to "the speaker's ability to demonstrate an awareness of audience and situation;" **linguistic competence** refers to more basic speech issues such as vocabulary selection, pronunciation, and syntax.

The Theory

Most theories of graph comprehension include the processes of (1) encoding a visual feature of the graph or data (sometimes referred to as a "visual chunk") and (2) interpreting that feature with respect to basic graph knowledge (e.g., a line going up means something is increasing) and specific graph content (e.g., "bicycle sales are increasing"). Carpenter & Shah (1998) provide evidence that comprehension occurs through repeated cycles of encoding and interpretation, building up more inclusive understanding of the graph. Thus, the more information (the greater the number of visual chunks) in a graph to integrate, the longer it takes to comprehend a graph.

We hypothesize that fewer visual chunks will similarly lead to higher quality descriptions. Fewer pieces of information to be described potentially leaves more time and cognitive resources for communicative tasks such as

providing cues for the listener as to the organization of the description, describing each piece of information succinctly, and so forth.

What are the visual chunks in multi-variable bar graphs? Shah, Mayer, & Hegarty (1999) argue that each group of bars associated with a particular value along the x-axis forms a visual chunk. Consistent with this theory, the researchers showed that descriptions of bars graph tend to be organized around these chunks.

Consider the graphs shown in Figures 1 and 2. These graphs represent the same data set, but switch the variables represented along the x- and z- (bar shades) dimensions. Which should be easier to describe? Figure 1 incorporates fewer visual chunks than does Figure 2 (two vs. five), so according to our hypothesis should elicit descriptions with higher communicative quality. Figure 1 has two groups of bars, each with one category that is much higher than the rest: describing this feature succinctly summarizes the data represented in the group. Thus, a straightforward description would be to make the global comparison within one age group (e.g., "For Age 20-30, interesting work is the most important") and then the other age group. While such a response does not necessarily capture every nuance of the data, it does capture the essential difference between the two groups. Note that it is important that the visual chunks in the low-chunks graphs contain an obviously maximal value. Otherwise, the group might be perceived as separate chunks (each bar), potentially diminishing the quality of descriptions that the graph elicits. In contrast, Figure 2 has five visual chunks: the relative height of the bars within each category.

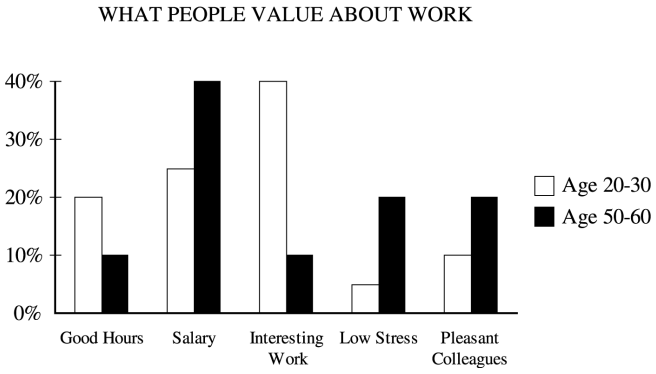

Figure 2: Many-chunks graph [More visual chunks].

This task analysis is not necessarily intuitively obvious. Although there are fewer visual chunks in Figure 1, the graph introduces five different shade-category mappings that might need to be either remembered or refreshed by looking at the legend (Lohse, 1993). Figure 1 should impose a heavier WM burden than Figure 2 because the latter has only two shades representing the two age groups. This alternative task analysis suggests that Figure 2 should elicit descriptions of superior communicative.

To test the visual chunk hypothesis, we conducted an experiment that manipulated two factors with the potential to affect the descriptive ease of a graph. First, as illustrated by Figures 1 and 2, we created two graph organizations for each of four data sets by switching the variables represented along the x-axis and by the differently shaded bars (the z-variable). One graph organization presents a smaller number of visual chunks (2-3 chunks depending on the data set) than the other organization (4-6 chunks). These two graph organizations will be referred to as the **few-chunks** (e.g., Figure 1) and **many-chunks** (e.g., Figure 2) graphs. The few-chunks graphs' organization minimizes amount of information to be described, and is therefore predicted to elicit better descriptions.

An alternative to the visual chunks hypothesis is that a comparison between two groups is simply a more natural way to describe a graph. In other words, any the superiority of the few-chunks graphs might be due to a particular descriptive strategy.

Are visual chunks defined by the organization of the graph or can participants' attention be drawn to see the fewer chunks in the "many chunks" graph? To investigate this possibility, we introduced alternative task prompts. **Open-ended** prompts were the same for all graphs and asked the participant to "Describe the information given in the graph." **Directive** prompts identified the critical contrast in the graph, suggesting more directly what should be described. For example, for Figure 1 the prompt was "Describe the changes in work values between the two age groups."

Method

Participants

Thirty-nine students (19 female, 18 male) participated in the experiment. Ten students[1] were recruited from each of four universities in the U.S., and students participated at their local institution. Eighty-five percent of participants were doing graduate or post-graduate work; others were juniors or seniors. Participants ranged in age from 21 to 45, with an average age of 29. Students' reported fields of study were medicine (20%), math or science (18%), humanities (12%), business (8%), and social science (7%).

Each institution was asked to recruit eight non-native English speakers and two native English speakers. Most of the participants (n = 19) were native speakers of a Chinese dialect; other languages were reported by no more than two or three participants (a mix of Asian, European, and Middle Eastern languages). There were seven native English participants because one institution recruited only one native English speaker instead of the request two. Most of the students had been living in the U.S. for fewer than two years (n = 22); the remaining students were evenly split between those that had lived in the U.S. 10 or more years (n = 9) and between 2 and 10 years (n = 8).

[1] Due to technical difficulties, one participants' data were lost, so one school contributed only nine students.

Materials

We constructed four data sets to be graphed as bar charts. Each data set had its own story line, which had been reviewed by professional test developers for comprehensibility to non-native speakers of English. The data had the following properties:

- The data showed changes in several nominal categories between two (or three) time periods. The time periods were either years or age groups as in Figures 1 and 2.
- Each time period had one category (unique to that group) that was clearly higher than the other categories.

We created two graphs from each data set, for a total of eight graphs. One graph in a pair placed the time periods along the x-axis and represented the categories on the z-variable (the different shades of bars)—this organization created the few-chunks graphs. The other graph was created by switching the variables represented along the x and z dimensions, creating the many-chunks graphs.

Design

The independent variables of graph organization and prompt directness were implemented in a completely within-subjects design: each participant received all four graph types. The organization type alternated, with half the subjects receiving few-chunks graphs first and half receiving many-chunks graphs first. Because of the possibility of one prompt type influencing the next, that variable was implemented using an ABBA design, with half the subjects receiving an open-ended prompt first and half receiving a directive prompt first.

Preliminary analyses suggested no *a priori* differences among the participants from each school in terms of their communicative competence in English or in their familiarity with reading graphs.

Procedure

Each university conducted one data collection session of 10 students. Sessions were typically conducted in a language lab or similar equipped facility. Each student had a tape recorder and headphones. Students heard the prompts over their headphones and spoke their responses, which were recorded on audiotape.

The questions were administered in two sets, with a short break between the sets; each set consisted of nine non-graph questions followed by two of the experimental questions. After both sets were administered, students were given a brief graph familiarity questionnaire. The questionnaire consisted of several questions concerning graph interpretation, a section on self-reported graph familiarity, and a short demographic questionnaire.

Measures

We obtained three types of dependent measures from each response: response latency, holistic scores, and four component scores. **Response latency** is the number of seconds between the end of the spoken prompt and when the participant began speaking. The timing was done by a research assistant unaware of the purpose of the experiment, using an on-line stopwatch while listening to each tape. Thus, the latencies recorded are probably only accurate to the nearest second, but there should be no systematic bias in these inaccuracies.

Each response was also scored by highly experienced TSE raters, each rater having participated in many rating sessions each year for five or more years. Raters produced a **holistic score** in a way identical to how actual TSE responses are scored. To provide finer-grain scores than the 5-level scale described earlier, each rater was asked to indicate whether a score fell into the high, middle, or low end of the score band. Thus, raters provided scores such as "high 40" or "low 60." Raters often discuss responses in this way, so producing this additional information was not difficult. In the analyses, a "high" score adds 3.3 to the band level (e.g., "high 40" becomes 43.3) whereas a "low" score subtracts 3.3 from the band level ("low 60" becomes 56.7). "Middle" scores are unadjusted.

Finally, each rater was asked to provide a score for each of the **component competencies** in the TSE Score Band Chart, as described earlier. Thus, each response received a discourse, functional, sociolinguistic, and linguistic score. These scores were rated on the typical 5-level (20-60) scale.

Results

We look at the effects of graph organization and prompt type from three perspectives. First, what are the effects on response latency? According to Carpenter and Shah (1998), a greater number of visual chunks should lead to longer latencies because of the greater number of encode-interpret cycles need for comprehension. Second, what are the effects on holistic scores? As we are looking at within-subject performance, any effects suggest an influence other than a person's own communicative competence on the score (i.e., variance irrelevant to the construct intended to be measured). Finally, as a follow-up to the effects on score, we look at the effects on the components of the score – the individual scores on discourse, functional, sociolinguistic, and linguistic competence.

We ran a 2x2 repeated-measures MANOVA, with graph organization (few- or many-chunks graphs) and prompt type (directive or open) as within-subjects factors and response latency as the dependent measure. There was a significant main effect of graph organization ($F(1,37^2)=4.0$, $p = .034$). Participants spent less time inspecting the few-chunks graphs before responding ($M = 5.5$; $SD = 3.7$) compared to the many-chunks graphs ($M = 6.8$; $SD = 4.6$). The main effect of prompt type was not significant nor was the interaction of graph organization and prompt.

Similar results were obtained for holistic scores. An identical 2x2 repeated-measures MANOVA revealed a significant effect of graph organization ($F(1,38)=8.1$,

[2] Due to technical difficulty, one participants' latency was not obtained.

p=.007). Participants received higher scores when responding to the few-chunks graphs (*M* = 47.7; *SD* = 9.1) compared to the many-chunks graphs (*M* = 46.1; *SD* = 9.5). The main effect of prompt type was not significant nor was the interaction of graph organization and prompt.

What types of effects does graph organization have on participants' responses? Are the responses to few-chunks graphs more expressive or more linguistically precise? While we might expect graph organization to affect how well organized a response is (i.e., discourse competence), it might be the case that a poorly organized graph increases WM load, so impinges on all language competencies.

Table 1 shows the effect of graph organization on each of the competency scores. As expected, discourse scores were significantly higher (via two-tailed, paired-samples t-test) for the few-chunks graphs: responses to these graphs were rated as more coherent and cohesive. There was an almost significant difference on the functional scores, whereby participants' responses to few-chunks graphs reflected language more appropriate to the task than did their responses to many-chunks graphs. There were no differences between the graph types in participants' ability to express their knowledge of audience (sociolinguistic) or in their pronunciation or grammar (linguistic).

Table 1. Mean (SD) scores by graph type.

Competence Component	Graph Type	
	Few-chunks	Many-chunks
Discourse	47.1	45.3*
	(8.6)	(9.9)
Functional	47.1	45.8
	(8.7)	(9.9)
Sociolinguistic	46.2	45.5
	(8.8)	(9.1)
Linguistic	48.0	47.2
	(8.8)	(8.5)

Note. Each graph type score is the mean of the two scores for each participant. N = 37 per cell because one participants' component scores were unavailable. * $p < .05$

Thus far, the results are consistent with the model that better performance is achieved with graphs that have fewer visual chunks. But are participants describing the visual chunks predicted by the theory? That is, for the fewer-chunks graph in Figure 1, participants' descriptions should include the global comparison between the highest category in a bar group and the other bars in that group (e.g., "Interesting Work is most important for the 20-30 year olds"). For the many-chunks graph in Figure 2, descriptions should instead include discrete comparisons within a category (e.g., "Interesting Work is more important to the 20-30 year olds than 50-60 year olds").

To address whether participants are describing the expected visual chunks for these two graphs, we analyzed the first piece of information mentioned in their response to the graphs. Given the speeded nature of the task, the first graph feature mentioned should be the most salient to the participant.

Participants' descriptions were consistent with their describing the two graphs in terms of the predicted visual chunks (Table 2). Participants mentioned first the global features of the data significantly more often when the graph was organized to accentuate these features (fewer-chunks graph) and mentioned first the discrete comparisons of the many-chunks graph ($\chi^2(1)=11.8$, $p<.001$).

Table 2. Graph type by first description.

Graph Type	Global Comparison	Discrete Comparison
Few-chunks (Figure 1)	19	1
Many-chunks (Figure 2)	8	10

Discussion

The research presented in this paper replicates and extends basic research on graph comprehension. The results provide support for the hypothesis that graphs with fewer visual chunks are easier to describe. Participants took less time to scan the few-chunks graphs before speaking, which replicates Shah & Carpenters' (1998) results. Graphs with fewer chunks also elicited descriptions of greater communicative quality. Furthermore, the organization of a graph had a very specific influence on the descriptions provided by participants: graphs with fewer visual chunks led to more cohesive and coherent descriptions. If the many-chunks graphs were worse because of lower overall comprehensibility, we would expect more aspects of descriptive competence to be affected. Future research might further extend Shah & Carpenter's processing model to explain the mechanisms by which the higher quality descriptions are facilitated.

The visual chunks hypothesis—fewer visual chunks leading to descriptions of higher communicative quality—has practical implications, suggesting desirable characteristics of graph questions for the Test of Spoken English. For example, two or visual chunks in a graph might be the limit of what is reasonably possible to describe within one minute. For multi-variable bar graphs as used in the current experiment, this recommendation would be limiting the number of bar-groups placed along the x-axis

The visual chunks hypothesis is applicable to a wider range of graph types, as long as we can adequately define the visual chunks. For example, other research (Carpenter & Shah, 1998; Carswell, 1993; Shah, Mayer, & Hegarty, 1999) suggests definitions of visual chunks for multi-function line graphs: each non-parallel line is a visual chunk, although each "reversal" in a line (e.g., changing

from an upwards to a downwards slope) is perceived as a separate chunk.

In line with the overall theme of the conference, applied research should adapt theories and results from the basic research literature to solve real-world problems, and then contribute back to the theoretical literature from which it drew. The applied research presented in this report achieves these goals.

Acknowledgments

This research was funded by the Test of Spoken English program of the TOEFL Policy Council. Peter Cheng was supported by the UK Economic and Social Research Council through the Centre for Research in Development, Instruction, and Training. We thank Susan Lynn Martin and Venus Mifsud for their assistance with this work, and Malcolm Bauer and Ann Gallagher for useful comments on earlier drafts of this paper. We are especially grateful to the TSE program staff and TSE raters for their contributions to this project.

References

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.

Carswell, C. M. (1993). Stimulus complexity and information integration in the spontaneous interpretations of line graphs. Applied Cognitive Psychology, 7, 341-357.

Carswell, C. M., Bates, J. R., Pregliasco, N. R., Lonon, A., & Urban, J. (1998). Finding graphs useful: Linking preference to performance for one cognitive tool. *International Journal of Cognitive Technology, 3*, 4-18.

Lohse, G. L. (1993). A cognitive model for understanding graphical perception. *Human-Computer Interaction, 8*, 353–388.

Myford, C., & Wolfe, E. (2000). Monitoring sources of variability within the Test of Spoken English assessment system (TOEFL Research Rep. No. 65). Princeton, NJ: Educational Testing Service.

Pinker, S. (1990). A theory of graph comprehension. In R. Freedle (Ed.), *Artificial intelligence and the future of testing*. Mahwah, NJ: Erlbaum.

Powers, D., Schedl, M., Wilson-Leung, S., & Butler, K. (1999). Validating the revisted Test of Spoken English against a criterion of communicative success. *Language Testing, 16*, 399-425.

Shah, P., Hegarty, M., & Mayer, R. E. (1999). Graphs as aids to knowledge construction: Signaling techniques for guiding the process of graph comprehension. *Journal of Educational Psychology, 91*, 690-702.

Test of English as a Foreign Language (TOEFL) (2001). *TSE and SPEAK score user guide*. Princeton, NJ: Educational Testing Service [also available through http://www.toefl.org/pubs/pubsindx.html]

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