ETS TOEFL

# Research Reports

*Report 74*
*February 2004*

# Elicited Speech From Graph Items on the Test of Spoken English™

**Irvin R. Katz**

**Xiaoming Xi**

**Hyun-Joo Kim**

**Peter C.H. Cheng**

# Elicited Speech From Graph Items on the Test of Spoken English™

Irvin R. Katz
ETS, Princeton, NJ

Xiaoming Xi
University of California, Los Angeles

Hyun-Joo Kim
Teachers College, Columbia University, NY

Peter C-H. Cheng
University of Nottingham, UK

RR-04-06

*ETS is an Equal Opportunity/Affirmative Action Employer.*

**Abstract**

This research applied a cognitive model to identify item features that lead to irrelevant variance on the Test of Spoken English™ (TSE®). The TSE is an assessment of English oral proficiency and includes an item that elicits a description of a statistical graph. This item type sometimes appears to tap graph-reading skills—an irrelevant construct; TSE raters report that many examinees perform worse on this item type than they do on the other 11 items in the test. We adapted a cognitive theory of graph comprehension to predict the degree to which TSE graph items tap irrelevant skills such as graph reading. Through analyses of existing TSE data as well as an experiment, we show how the theory provides specific, empirically justified recommendations on the construction of graph items that minimize the influence of extraneous skills.

Key words: communicative competence, graph description task, visual processing, Test of Spoken English™ (TSE®)

The Test of English as a Foreign Language™ (TOEFL®) was developed in 1963 by the National Council on the Testing of English as a Foreign Language. The Council was formed through the cooperative effort of more than 30 public and private organizations concerned with testing the English proficiency of nonnative speakers of the language applying for admission to institutions in the United States. In 1965, Educational Testing Service® (ETS®) and the College Board® assumed joint responsibility for the program. In 1973, a cooperative arrangement for the operation of the program was entered into by ETS, the College Board, and the Graduate Record Examinations (GRE®) Board. The membership of the College Board is composed of schools, colleges, school systems, and educational associations; GRE Board members are associated with graduate education.

ETS administers the TOEFL program under the general direction of a policy board that was established by, and is affiliated with, the sponsoring organizations. Members of the TOEFL Board (previously the Policy Council) represent the College Board, the GRE Board, and such institutions and agencies as graduate schools of business, junior and community colleges, nonprofit educational exchange agencies, and agencies of the United States government.

A continuing program of research related to the TOEFL test is carried out under the direction of the TOEFL Committee of Examiners. Its 12 members include representatives of the TOEFL Board and distinguished English as a second language specialists from the academic community. The Committee meets twice yearly to review and approve proposals for test-related research and to set guidelines for the entire scope of the TOEFL research program. Members of the Committee of Examiners serve four-year terms at the invitation of the Board; the chair of the committee serves on the Board.

Because the studies are specific to the TOEFL test and the testing program, most of the actual research is conducted by ETS staff rather than by outside researchers. Many projects require the cooperation of other institutions, however, particularly those with programs in the teaching of English as a foreign or second language and applied linguistics. Representatives of such programs who are interested in participating in or conducting TOEFL-related research are invited to contact the TOEFL program office. All TOEFL research projects must undergo appropriate ETS review to ascertain that data confidentiality will be protected.

Current (2003-2004) members of the TOEFL Committee of Examiners are:

| | |
|---|---|
| Micheline Chalhoub-Deville | University of Iowa |
| Lyle Bachman | University of California, Los Angeles |
| Deena Boraie | The American University in Cairo |
| Catherine Elder | University of Auckland |
| Glenn Fulcher | University of Dundee |
| William Grabe | Northern Arizona University |
| Keiko Koda | Carnegie Mellon University |
| Richard Luecht | University of North Carolina at Greensboro |
| Tim McNamara | The University of Melbourne |
| James E. Purpura | Teachers College, Columbia University |
| Terry Santos | Humboldt State University |
| Richard Young | University of Wisconsin-Madison |

To obtain more information about the TOEFL programs and services, use one of the following:

**E-mail: toefl@ets.org**

**Web site: www.toefl.org**

**Acknowledgements**

This research was funded by the Test of Spoken English program of the TOEFL Policy Council. The UK Economic and Social Research Council supported Peter Cheng through the Centre for Research in Development, Instruction, and Training. We thank Shauna Cooper, Susan Lynn Martin, and Venus Mifsud for their assistance with this work; Jill Carey and Yong-Won Lee for their data analysis support; Kathy Sheehan and Alina von Davier for advice on experimental design and analysis; and Malcolm Bauer, Ann Gallagher, Pat Kyllonen, and Val Shute for useful comments on earlier drafts of this paper. We are grateful to the TSE program staff—especially Evelyne Aguirre Patterson, Emilie Pooler, and John Miles—and to the TSE raters for their contributions to this project. This project was initially directed by Hunter Breland, and we thank him for his many contributions to research planning.

## Table of Contents

## List of Tables

Page

## List of Figures

Page

## Introduction

How do we know whether an item introduces construct-irrelevant variance to a test score and, if it does, what can we do about it? Psychometrics offers several methods for investigating the validity or reliability of scores at the item level, including differential item functioning and inspection of item-total correlation. These and other methods have proven worthwhile for *detecting* possible construct-irrelevant variance, but they provide no guidance on how to *improve* the measurement of weak items. Test developers must decide whether to discard weak items or to modify them, with the latter typically based on intuition or guidelines stemming from accumulated expertise.

Cognitive psychology, combined with psychometrics, can fill this gap. Modeling the information processing demands of test items vis-à-vis their psychometric properties can provide guidance on how to modify items so that they are more likely to tap the constructs of interest. Rather than informal rules-of-thumb, such models are built around empirically supported theories concerning the structure and limits of human information processing.

This paper takes the unique approach of modeling the *extraneous skill* thought to introduce irrelevant variance. If we know what characteristics of tasks make the tasks more likely to require the extraneous skills, we can provide guidelines to test development that avoid these pitfalls. We adapted a cognitive model to identify item features that lead to irrelevant variance on a particular type of item from the Test of Spoken English™ (TSE®), an assessment of English oral proficiency.

The TSE includes an item that presents a statistical graph and prompts examinees to describe the information presented. This question type—one of 12 items in the TSE—is illustrated in Figure 1.[1] Test-takers are given one minute to complete their response. Only the communicative quality of the response is scored—the degree to which the description reflects "the ability of nonnative speakers of English to communicate orally in a North American English context" (ETS, 2001, p. 4). Even though the accuracy of the description is *not* scored, one might still suppose that an examinee's skills in reading graphs contributes to the score, potentially hindering (or helping) performance.

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

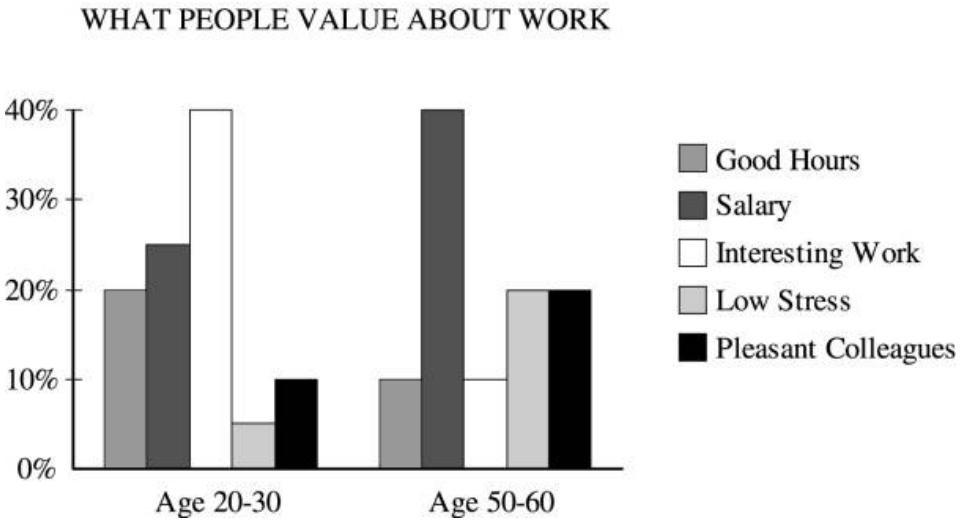


***Figure 1.*** **Illustrative TSE graph question.**

This graph-description question has posed problems for scoring. Anecdotally, TSE raters report that some graph items appear to elicit performance that is inconsistent with examinees' performance on the remainder of the test. In other words, certain graphs elicit speech that demonstrates a lower (or higher) ability in English than would be expected based on responses to other test questions. Raters also report that some graph items elicit unratable speech, such as simple listings of numbers depicted in the graph. In informal interviews, raters and test development staff provide many possible explanations for the observed problems with TSE graph items, including examinees being unfamiliar with, or uninterested in, the content of a graph (e.g., bicycle sales) and the visual complexity of a graph causing confusion for some examinees.

Whereas TSE raters report these difficulties, statistical analyses of TSE data have not revealed any systematic weakness in the TSE graph item type. Such analyses instead support the generally high internal consistency of TSE items, with most of the items contributing equally to measurement (Myford & Wolfe, 2000; Powers, Schedl, Leung, & Butler, 1999; Wang, Bradlow, & Wainer, 2000).

Critical to the TSE is the issue of which characteristics of a graph lead to descriptions that best indicate communicative skill. 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. As pointed out earlier, the accuracy of a person's response to a graph item is not considered in the score; rather, the score reflects the degree to which the person demonstrates 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 potentially narrow any differences between test-takers), yet are straightforward enough to describe, allowing a test-taker to show off his/her communicative skill without other factors getting in the way.

Given the difficulties raised with the TSE graph item, one might reasonably ask whether a graphical description task belongs in an assessment of general speaking proficiency. However, there are several compelling reasons to keep the TSE graph item. The item mirrors the types of descriptive and interpretative tasks undertaken by healthcare professionals and teaching assistants in their day-to-day work. Many of the test-takers are going into fields in which reading, describing, and interpreting graphs are an important part of their jobs. Thus, the task has a degree of face validity. Furthermore, in contrast to more verbal prompts, the graph item conveys information without providing language for the test-taker to quote. Because the information is largely visual, most of the language must come from the test-taker, providing a good measure of the test-taker's language usage.

The paper is structured as follows. In the next section, we present background information on the TSE and its scoring. The section following addresses the question of which characteristics of graph items make them better or worse indicators of communicative skill. We then present two investigations that support our characterization of the features of graphs that affect their quality as measures of general speaking proficiency. Finally, we present recommendations for the construction of TSE graph items.

## Background: The Test of Spoken English

The Test of Spoken English (TSE) is designed to measure a test-taker's communicative competence in Northern American English (Douglas & Smith, 1997). It is taken by

approximately 30,000 non-U.S. citizens each year, who are seeking to become teaching assistants or healthcare professionals in the United States. The test consists of 12 questions that elicit a range of language functions (e.g., describe, compare, state opinion). Most of the questions consist of verbal prompts, but a few questions utilize more visual prompts, such as pictures, maps, or graphs. The questions are presented visually in a booklet, and delivered verbally by a pretaped interviewer; the test-takers' spoken responses are recorded. Trained raters score responses by employing a well-defined scoring rubric that draws on a well-known model of communicative language ability (Bachman & Palmer, 1996) and includes four language competencies: functional, discourse, sociolinguistic, and linguistic.

Responses to TSE prompts are scored according to the published "TSE Score Band Descriptor Chart" (ETS, 2001). This scoring rubric defines four key communicative competencies: discourse, functional, sociolinguistic, and linguistic competence. The chart also specifies typical response characteristics for these competencies at each of the five possible score bands (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 major models of communicative language ability and analyses of linguistic features of sample responses that represent different proficiency levels (Bachman & Palmer, 1996; Douglas & Smith, 1997).

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 follows (ellipses refer to short pauses in speech):[2]

(1) *the good hours...ah for age...ah,...between age...50 and 60 is ten percent....And...the pleasant...colleagues...for...ah...for age...20 to 30...is ten percent...and...ah for...50 to 60 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:

(2) *...for adults...uh,...between age two,...20 to 30,...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 better guides the listener by using discourse markers such as "for the old man…" and "Other points I should compare…".

*Functional competence* is the ability to use appropriate language to transfer information and ideas to accomplish a goal. It is demonstrated by the extent to which a person communicates an intended goal. For the graph in Figure 1, a partial response demonstrating low functional competence follows:

(3) *Ok, people...around the age...20 to 30...I guess started like...ah...just youngsters...they are...um... they good hours up like twenty percent...and...only...ah...at the age of 20 to 30...the people who are interested...are only forty percent*

This response does not communicate the information provided in the graph, partially because the speaker misrepresents the meaning of "good hours" and "interesting work." Response (1), in contrast, does a good job of describing the information and therefore, was rated higher on functional competence than was response (3).

The other two competencies appear less likely to be affected by the particular characteristics of a graph. *Sociolinguistic competence* is the ability to demonstrate an awareness of audience and situation. *Linguistic competence* refers to speech features such as vocabulary use, syntax, pronunciation, and fluency.

The remainder of this paper focuses on identifying the characteristics of the graph items that lead to higher or lower quality items. If we understand which features of graphs lead to items that poorly assess communicative competence—that is, graphs that require more graph-reading skills than others—then we can make recommendations to test development staff on the

crafting of TSE graph items. First, drawing from research on the cognitive processes underlying graph comprehension, we present the *visual chunks* theory, which asserts that people tend to describe graph items in terms of visually identifiable graph features in the data, and which defines how to identify these features in a graph. Second, based on previous research on the cognitive processes underlying graph comprehension, we hypothesize that the number of visual chunks will predict item quality—the more that needs to be described and integrated to support the communication of major points in a graph, the lower the communicative quality of a response. Finally, we present two investigations of this hypothesis: a regression analysis of existing TSE data, and a controlled experiment that specifically manipulates the number of visual chunks in a graph.

**Applying Theories of Graph Comprehension: The Visual Chunks Hypothesis**

At the time this project was conducted, there were 39 TSE graph items for which administration data were available (those administered between July 1997 and August 2000). The items included 8 pie charts, 19 bar charts, 8 line graphs, 2 items including both a bar and a line graph, 1 unidentified graph (the test form on which it appeared was not available), and 1 table. However, within these general types are a variety of story contexts (e.g., bicycle sales, electricity usage, family budgets), visual formalisms (tic marks, shading, labeling) and data types (single function or multiple function), and x-axis scales (continuous, discrete). For example, there are items that present the comparison of two pie charts, sometimes with individual pie sections shaded or sometimes without shading. Some bar charts show the level of an individual variable over several months (e.g., number of books checked out) while other bar charts show how two different types of data change over time (e.g., the relative popularity of different college majors). From this wide array of graph types, formalisms, and data types, a theoretical model can point out the important and unimportant differences among the graphs.

There is a large body of literature on the comprehension and interpretation of statistical graphs, stemming from research in cognitive psychology, statistics, education, and management to name a few fields. Much of this research consists of either expert discussion on what makes a good graph (e.g., Tufte, 1983; Wainer, 2000) or empirical studies of the relative comprehensibility (typically measured via narrow laboratory tasks) of graphs containing different visual features (e.g., see reviews in Friel, Curcio, & Bright, 2001; Lewandowsky &

Behrens, 1999; Shah & Hoeffner, 2002). Additionally, several authors (Carpenter & Shah, 1998; Kosslyn, 1989; Lohse 1993; Pinker 1990) have compiled their and others' empirical results into comprehensive theories that specify, the detailed cognitive processes underlying graph comprehension. Although the various theories may differ in details, there is much agreement on the broad outlines of how people go about comprehending statistical graphs.

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 and Shah (1998) provide evidence that comprehension occurs through repeated cycles of encoding and interpretation, building up more inclusive understanding of the graph. Through reaction time studies and analyses of eye movements during graph comprehension, the researchers show that the more information (the greater the number of visual chunks) in a graph to integrate, the longer it takes to comprehend a graph. Furthermore, several empirical studies have shown that people tend to describe graphs in terms of these visual chunks (Carswell, 1993; Shah, Hegarty, & Mayer, 1999).

*We hypothesize that fewer visual chunks will similarly lead to higher quality descriptions.* Having fewer pieces of information to be described potentially leaves more time and cognitive resources for participants to monitor their language and organize their response.

To apply the theory to the TSE graph items requires strict definitions of the visual chunks represented in the graphs. Because the relevant literature focuses on bar and line graphs, we limit our discussion to these graph types (approximately three quarters of TSE graph items). This focus is justified as, anecdotally, test development staff report fewer difficulties associated with scoring pie charts as compared to bar and line graphs. Thus, the pragmatic need is to understand how to create "good" bar and line graphs for TSE graph items.

What are the visual chunks in line graphs? Carpenter and Shah (1998) provide empirical evidence that each line forms a separate chunk, unless the lines are parallel. That is, if a graph has more than one line (e.g., Figure 2[3]), each line is a separate visual chunk. Carswell (1993) builds on this claim by showing that *reversals* in a line—such as the switching of the slope from positive to negative—breaks a line into separate visual chunks. In contrast, other features of a line such as the number of points represented in a line or a simple change in rate (but not

direction) of slope does not add further information (i.e., something else to describe). Thus, Figure 2 contains four visual chunks—each line represents two visual chunks because there is one reversal in each line. Figure 3 contains two visual chunks because neither line has reversals.

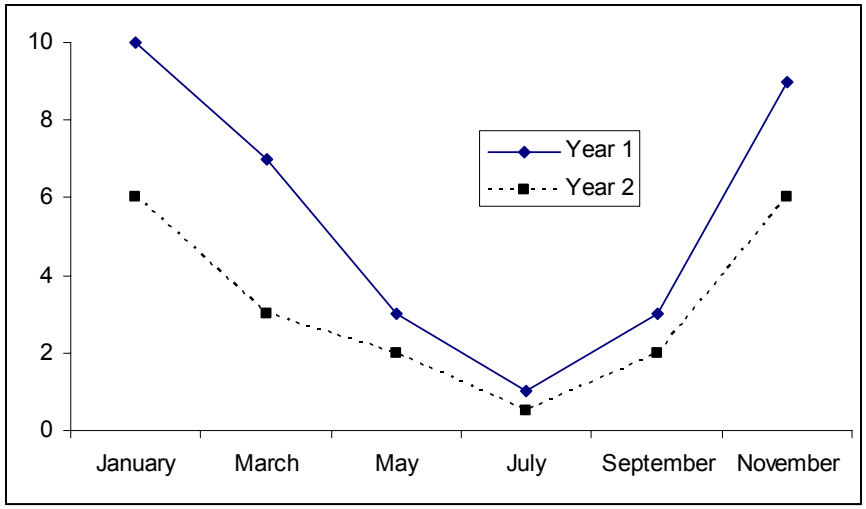


***Figure 2.*** **Graph with two data series and four total visual chunks.**

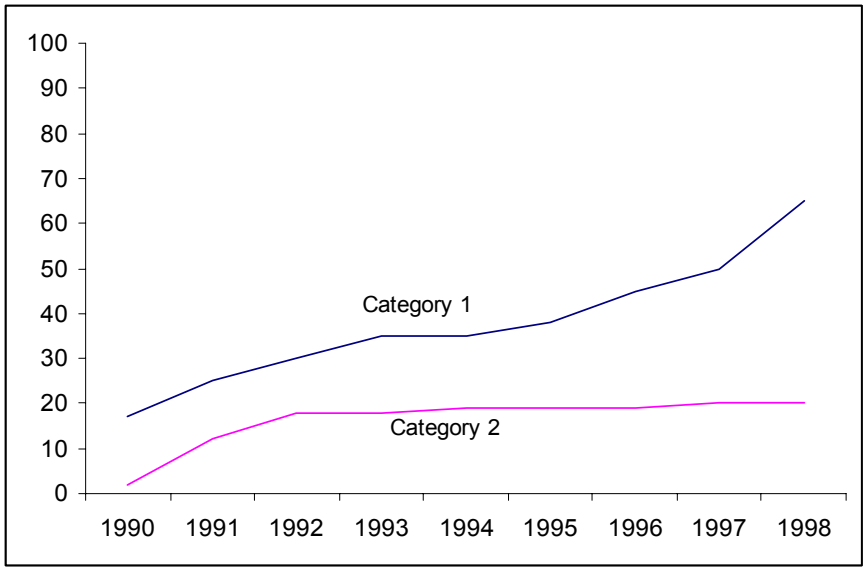


***Figure 3.*** **Graph with two data series and two total visual chunks.**

What are the visual chunks in bar charts? We consider first bar charts having discrete categories listed across the x-axis (i.e., a nominal scale; see Figure 9 for an example of a nominal scale of response categories). Shah, Hegarty, and Mayer (1999) demonstrated that each group of bars associated with a particular value along the x-axis form a visual chunk. Consistent with this theory, the researchers showed that descriptions of bars within a graph tend to be organized around these chunks—people tend to describe one group of bars, then the next, and so forth, rather than describing information associated with a particular shade of bar as it occurs in several groups. Thus, Figure 4 contains five visual chunks—each group of bars can be described as a simple unit of information (e.g., "For Category 1, Year 2 is greater than Year 1").

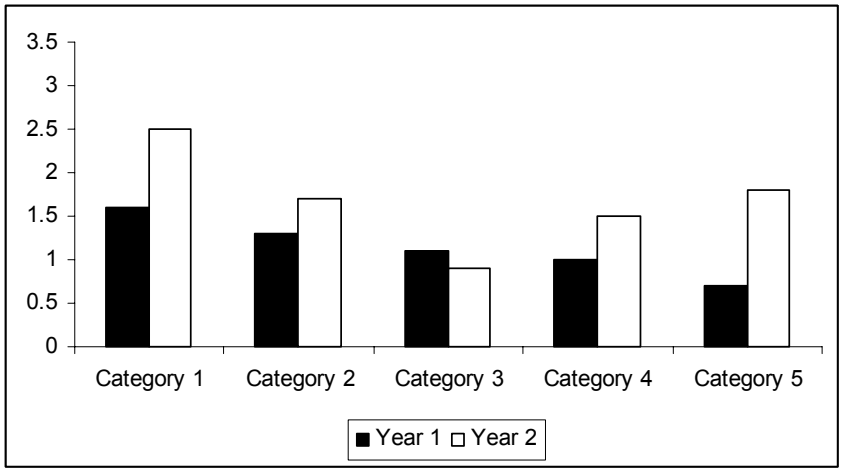


***Figure 4.*** **Graph with two data series and five visual chunks (one per x-axis group).**

Bar charts containing a continuous scale on the x-axis (e.g., years) may be treated the same as line graphs. That is, one can imagine a line connecting the tops of the bars of a particular shade to create a line graph. Thus**,** Figure 5 shows a bar graph representing four visual chunks—although there is an individual data series, three reversals are present in the data. For clarity**,** Figure 6 shows the same figure as a line graph with the visual chunks highlighted. Figure 7 shows a bar graph containing two visual chunks, one for each "line," neither of which has a reversal.

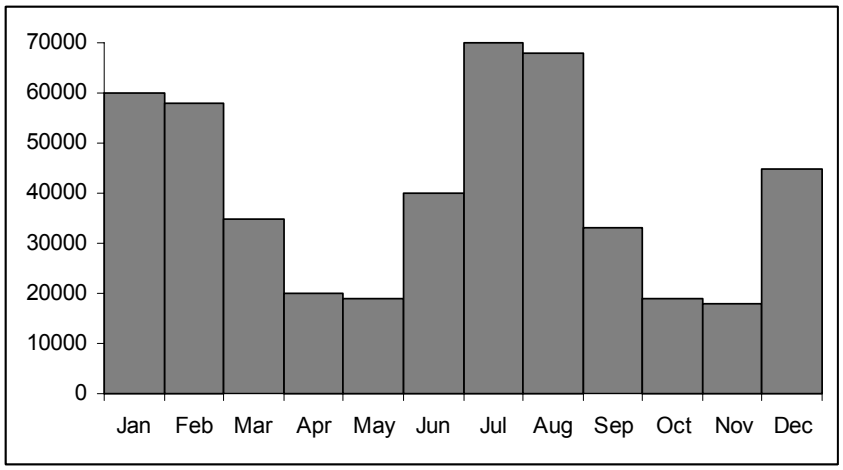


***Figure 5.*** **Graph with one data series and four visual chunks.**

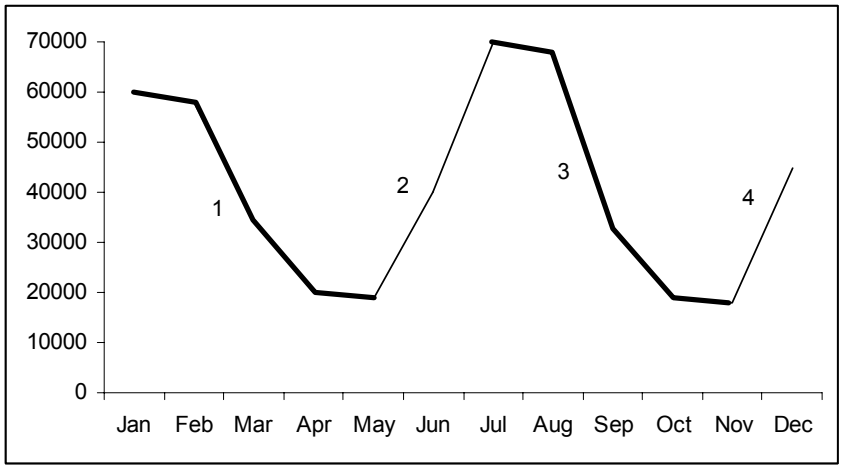


***Figure 6.*** **Previous figure with visual chunks highlighted.**

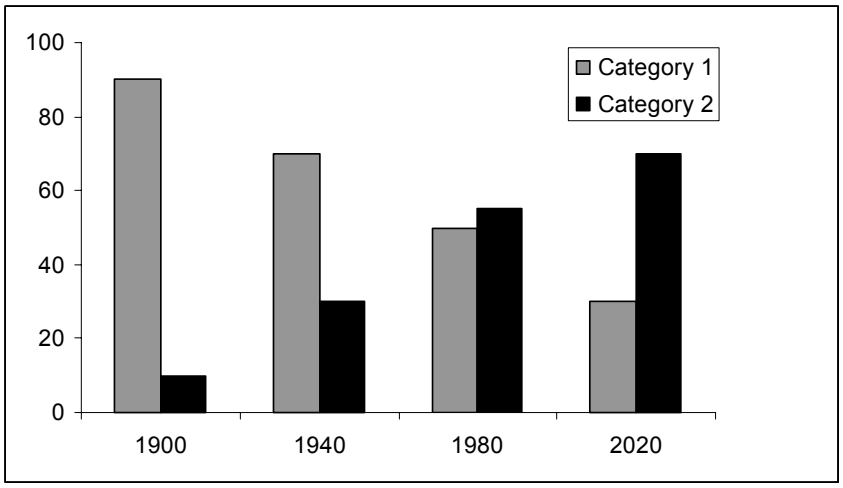


***Figure 7.*** **Graph with two data series and two visual chunks.**

Using the rules outlined above, one can determine the number of visual chunks in almost any bar or line graph. Note that these rules are objective, requiring no qualitative judgments to determine the number of visual chunks in a graph.

The next sections present two investigations of the visual chunks hypothesis: do more visual chunks lead to lower-quality items? First, we present a regression analysis that tests the predictive strength of two factors derived from the theory (number of visual chunks; individual or multiple data series) on the quality of the resulting item. Next, to supplement this correlational analysis, we present an experiment in which the number of visual chunks in a graph was systematically manipulated.

## Study 1: Modeling the Quality of TSE Graph Items

The goal of the regression analysis is to test the strength of the visual chunks theory to predict the quality of TSE graph items. In the following analysis, we use the visual chunks theory to classify the features of the 29 bar and line graphs among the set of administered items described earlier. By "item quality," we refer to the degree to which a graph item elicits a speech sample consistent with performance on the remainder of the test.

As noted earlier, TSE raters report that some graph items elicit performance that is lower than what would be expected given examinees' performance on the rest of the test. These anecdotal reports suggest a novel measure of item-whole comparisons: discrepancy scores.[4] We define a discrepant case as one where an examinee's score on an item is five or more points

*below* the average of the other items. This criterion was chosen because five points represents half a score band (TSE items are scored on a scale of five points ranging from 20 to 60 in increments of 10). If a graph item elicits a high percentage of discrepant cases, it suggests that the item might be tapping skills different than those assessed by the other items in the test.

## *Method*

### *Independent Variables*

According to the theories presented earlier, people comprehend graphs through repeated cycles of encoding, and then interpreting, the visual chunks in a graph. Carpenter & Shah (1999) showed that more visual chunks lead to a greater cognitive processing load, and, we speculate, will similarly lead to lower quality descriptions of TSE graph items. Graphs also differ in the ease with which people can interpret the visual chunks, that is to say, relating the quantitative relationships shown in visual patterns to variables. For example, if the visual chunks represent quantitative information about different variables, people might need to refer to a graph's legend to interpret each chunk. Lohse (1993) showed how people's eye movements return to the legend of a graph in a pattern consistent with the idea that they are refreshing their memory of how to interpret a symbol or portion of a graph. For example, the graph in Figure 5 should pose little cognitive load when a person interprets each chunk because each of them refers to the same entity: there is only one shade of bar (one data series). In contrast, Figure 7, which has two data series, should introduce greater cognitive load because a person needs to refresh his or her memory of the meaning of each bar shade by referring to the legend when attempting to describe the visual chunks.

Two independent variables were used in this analysis:

- **Number of visual chunks**. This variable encodes the number of visual chunks presented in a graph. The number of chunks was determined by the rubric presented earlier. Among the 29 items considered, the number of visual chunks ranged from one to six.
- **Data series type (individual/multiple)**. This variable encodes whether the graph shows an individual or a multiple series of data. An individual series is a graph with a single line or a single set of bars. Multiple-series graphs include more than

one line and more than one set (shading) of bars and, as a result, are expected to impose a greater cognitive load because of the need to refresh one's memory regarding the meaning of each line or bar shade (Lohse, 1993). Among the 29 items considered, 13 consisted of an individual data series, 14 depicted two data series, and one graph each contained three and four data series. Because the sample did not contain a wide enough range of number of series, we simplified this variable to encode only individual versus multiple data series. In the analysis, an individual series was coded as "0" and a multiple series was coded as "1."

### *Items*

As noted earlier, the items considered for this analysis include the TSE graph items containing either a bar or a line graph. However, the number of items used in the regression analysis was reduced because of an artificial interdependence between the two predictors of number of visual chunks and data series type. As described earlier, each data series adds an additional visual chunk. Therefore, graphs with one visual chunk cannot have more than one data series. Such a restriction does not exist for graphs with more than one visual chunk (e.g., a graph showing two visual chunks could consist either of an individual or multiple data series). To simplify the regression analysis, we included data only from the 23 items having graphs with two or more visual chunks. We acknowledge that this restriction limits the generality of the results and in the discussion separately consider the case of graphs with only one visual chunk.

### *Dependent Variable*

The percentage of discrepant cases was used as the dependent measure in the regression analysis. This measure potentially avoids the scaling issues inherent in comparing unequated scores across test administrations and provides greater variability than item-total correlation. Among the items analyzed, the percentage of discrepant cases ranged from 6.3 to 26.8 with a mean of 14.8 and standard deviation of 5.9.

## *Results*

Figure 8 plots the number of visual chunks and data series type by the measure of item quality. Higher percentages indicate more discrepant cases elicited by an item, and therefore, indicate an item of lower quality.

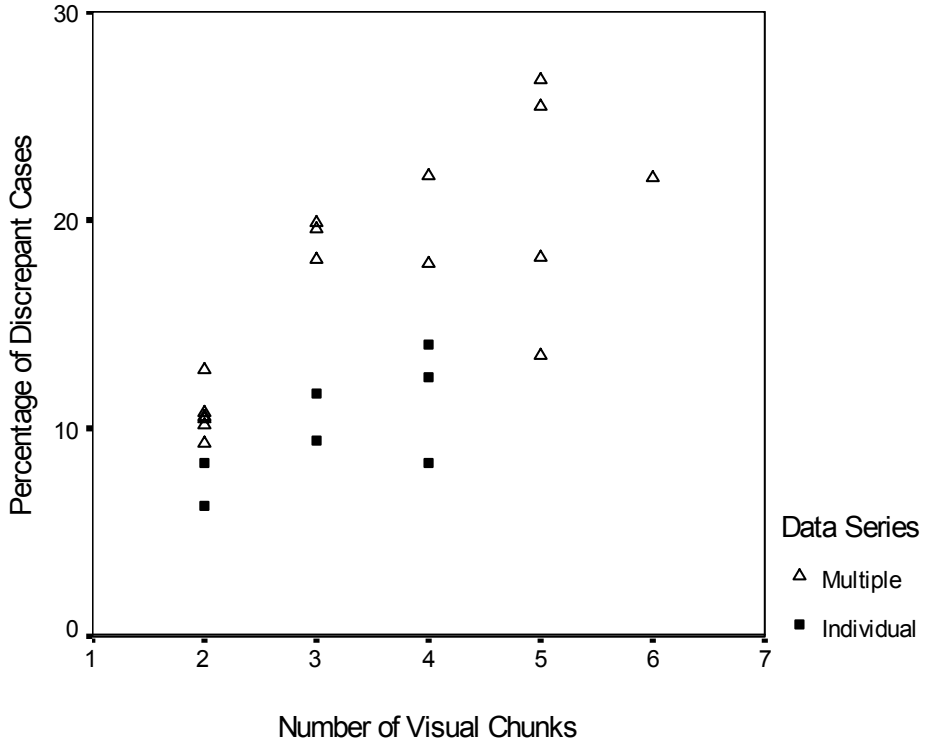


***Figure 8.*** **Scatterplot for regression analysis.**

The two predictors are both strongly correlated with the measure of item quality (see Figure 8). More complex graphs (more visual chunks and multiple data series) tend to elicit discrepant performance. The two factors are only weakly correlated with each other, suggesting unique potential predictive power of each (see Table 1).

**Table 1**

***Intercorrelations of Dependent and Independent Measures***

| | Percentage of discrepant cases | Number of visual chunks | Data series type |
|---|---|---|---|
| Percentage of discrepant cases | — | 0.690* | 0.535* |
| Number of visual chunks | | — | 0.110 |
| Data series type | | | — |

*$p < 0.01$.

To investigate the relative contributions to prediction of the two factors, we conducted a hierarchical regression analysis in which we incrementally added each factor, plus their interaction, to the model. The results of the hierarchical regression analysis are shown in Table 2 Each of the two main-effect factors contributes significantly to prediction, although the interaction effect does not add significant predictive power. Overall, the regression model accounts for approximately 70% of the variance in the measure of item quality.

**Table 2**

***Hierarchical Multiple Regression Analysis Results***

| Step | Predictor added | Cumulative $R^2$ | Change $R^2$ | *F* | *df* |
|---|---|---|---|---|---|
| 1 | Number of visual chunks | 0.476 | 0.476 | 19.1* | 1, 21 |
| 2 | Data series type | 0.690 | 0.214 | 13.8* | 1, 20 |
| 3 | Visual chunks-by-data series | 0.695 | 0.005 | .35 | 1, 19 |

*$p < 0.001$.

## Discussion

These results provide strong support for the visual chunks theory. The two factors derived from the theory predicted 70% of the variance in item quality among the bar and line graphs containing two or more visual chunks.

Items with only one visual chunk (and therefore having an individual data series) evidence a wider range of quality than would be expected, eliciting from 7%–17% discrepant cases. We might speculate that when there is not enough information to describe (i.e., only one visual chunk), examinees cast about for other aspects of the graph to talk about. As a result, other characteristics of the graph such as content, predictability of the data, and so forth, might have a stronger influence on performance.

Despite the strong results as predicted by the visual chunks hypothesis, this analysis has some limits. The analysis included a relatively small number of types of graph items: the bar and line graphs administered in the TSE. A wider range of graph items might show more variability that is not as well modeled by the two factors. The analysis was limited to graphs with more than one visual chunk and so the results cannot be generalized to graphs having one visual chunk.

Finally, the analysis is correlational and does not provide support for the idea that more visual chunks causes changes in performance. The next experiment explores the potential causal relationship between visual chunks and language quality of the elicited descriptions.

## Study 2: Experimental Investigation of Visual Chunks Hypothesis

In this experiment, we systematically manipulated the organization of a graph to create two versions of graphs that differed in the number of visual chunks. For example, the graphs shown in Figure 1 and Figure 9 represent the same data set, but the variables represented along the x- and z- (bar shades) dimensions are switched. Which should be easier to describe? Figure 1 incorporates fewer visual chunks than does Figure 9 (two vs. five), so according to our hypothesis, that graph 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 ages 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 fewer visual chunks in Figure 1 each include a visually obvious maximal value. Otherwise, each group might be perceived as separate chunks (each bar), potentially diminishing the quality of descriptions that the graph elicits. Figure 9, in contrast, has five visual chunks: the relative height of the bars within each category. Thus, more time is needed to comprehend the graph, and the communicative quality of any descriptions of this graph should be lower than those of Figure 1.

There is another way, however, in which these graphs may be interpreted. 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). The results of the previous regression analysis supported this notion of added complexity from additional data series (admittedly, however, the regression compared individual vs. multiple data series and the two graphs in this discussion both have multiple data series). From this alternative task analysis, Figure 1 might impose a heavier working-memory burden than Figure 9 because the latter has only two shades representing the two age groups. This alternative task analysis predicts that Figure 9 would elicit descriptions of superior communicative quality.

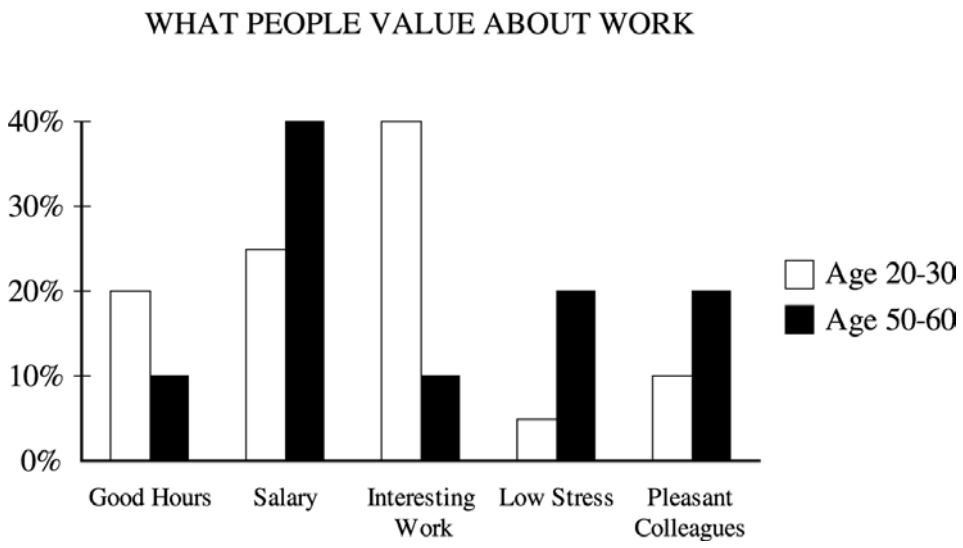


***Figure 9.*** **Alternative form of Figure 1 (more visual chunks).**

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. For the first factor, 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 (two to three chunks depending on the data set) than the other organization (four to six chunks). These two graph organizations will be referred to as the *few-chunks* (e.g., Figure 1) and *many-chunks* (e.g., Figure 9) graphs. The few-chunks graphs' organization minimizes the amount of information to be described, and is therefore predicted to elicit better descriptions. The second factor manipulated participants' attention to selected portions of the graphs. 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 superiority of the few-chunks graphs might be due to a particular descriptive strategy.

This alternative hypothesis suggests the possibility of drawing participants' attention to the fewer chunks even within a many-chunks graph (e.g., seeing the maximal values for the two age groups 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[5]) participated in the experiment. Ten students[6] were recruited from each of four universities in the U.S., and students participated at their local institution.[7] 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 (31%), math or science (26%), business (23%), humanities (10%), and social science (10%).

Each institution was asked to recruit eight nonnative English speakers and two native English speakers. Most of the participants were native speakers of a Chinese dialect ($n$ = 19); 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 requested two. Most of the students had been living in the United States for fewer than 2 years ($n$ = 22); the remaining students were evenly split between those that had lived in the United States 10 or more years ($n$ = 9) and between 2 and 10 years ($n$ = 8).

### *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 nonnative speakers of English. The data represented the interaction of two independent variables, with one variable having fewer levels (2–3) than the other (3–5). The variables with fewer levels were either years or age groups (as in Figure 1). The other variables were either nominal categories (e.g., work values) or intervals (e.g., hours in a day).

We created two graphs from each data set, for a total of eight graphs. One graph in a pair placed the 2–3 level variable along the x-axis and represented the other variable on the z dimension (the different shades of bars)—this organization created the few-chunks graphs. The many-chunks graph was created by switching the variables on the x and z dimensions.

***Design***

The independent variables of graph organization (few visual chunks vs. many visual chunks) and prompt directness (open vs. directed) were implemented in a completely within-subjects design: each participant received four graph items corresponding to both levels of both independent variables. For each participant, the organization type alternated, with half the participants receiving few-chunks graphs first and half receiving many-chunks graphs first. For the prompt directness variable, because of the possibility of one prompt type influencing the next, that variable was implemented using reverse counterbalancing, whereby each participant received both prompt types first in one order and then in the reverse order (sometimes called an ABBA design, where "A" and "B" refer to the two levels of the independent variable). Half the participants received an open-ended prompt first, and half received a directive prompt first.

Table 3 shows the full design of the experiment. As mentioned earlier, each participant received four graph items, each consisting of a different data set (and corresponding story line). Participants from each school responded to the items in the order shown in the table. Thus, participants at School 1 received the few-direct (graph organization and prompt directness, respectively) version of Data Set A first, then the many-open version of Data Set B, and so forth. Participants from Schools 3 and 4 received precisely the same graph items, as did participants from Schools 1 and 2, respectively, just in a different order. 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.[8]

**Table 3**

***Design of the Experiment***

| | Data Set A | | Data Set B | | Data Set C | | Data Set D | |
|---|---|---|---|---|---|---|---|---|
| School | Chunks | Prompt | Chunks | Prompt | Chunks | Prompt | Chunks | Prompt |
| 1 | Few | Direct | Many | Open | Few | Open | Many | Direct |
| 2 | Many | Open | Few | Direct | Many | Direct | Few | Open |
| | Data Set C | | Data Set D | | Data Set A | | Data Set B | |
| School | Chunks | Prompt | Chunks | Prompt | Chunks | Prompt | Chunks | Prompt |
| 3 | Few | Open | Many | Direct | Few | Direct | Many | Open |
| 4 | Many | Direct | Few | Open | Many | Open | Few | Direct |

***Procedure***

Each university conducted one data collection session of 10 students. Sessions were typically conducted in a language lab or similarly equipped facility. Besides a test booklet, 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 items were administered in two sets, with a short break between the sets; each set included nine nongraph items followed by two of the experimental items. After both sets were administered, students received a brief 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 taped response.

Highly experienced TSE raters scored each response using the TSE scoring rubric. Raters produced a *holistic score* using the identical procedures used to score actual TSE responses**.** To provide finer-grain scores than the five-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." This approach divides each 10-point score band into three sub-bands. Raters often discuss responses in this way, so producing this additional information was not difficult.

In converting these relative rankings into scores, "middle" scores were unadjusted to facilitate comparison between these scores and the typical score scale for the TSE. As a result, in the analyses presented below, a "high" score adds 3.3 (one third of the 10 point score band) 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).

After providing the holistic scores for all of his or her assigned responses, each rater was asked to listen to each response again and provide a score for each of the *component competencies* in the TSE Score Band Chart, as described earlier. Thus, in addition to a holistic score, each response received a discourse, functional, sociolinguistic, and linguistic score. These scores were rated on the typical five-level (20–60) scale.

### *Results*

We look at the effects of graph organization (few or many visual chunks) and prompt type (open or directive prompt) 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 needed 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 holistic 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 2 x 2 repeated-measures ANOVA, 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 (see Table 4). There was a significant main effect of graph organization: participants spent less time inspecting the few-chunks graphs before responding ($M = 5.5$; $SD = 3.7$) compared to the many-chunks graphs ($M = 6.8$; $SD = 4.6$). However, the effect size measure ($\eta^2$, the proportion of the total variance that is attributed to an effect)

suggests that this statistically significant effect might not be practically significant; we address this issue in the discussion. The main effect of prompt type was not significant nor was the interaction of graph organization and prompt.

**Table 4**

***Response Latency ANOVA***

| Source | *df* | *F* | $\eta^2$ | *p* |
|---|---|---|---|---|
| Graph organization | 1 | 4.9* | 0.12 | 0.03 |
| Graph organization by subjects (within-group error) | 37[a] | (13.4) | | |
| Prompt type | 1 | 0.67 | 0.02 | 0.42 |
| Prompt type by subjects (within-group error) | 37 | (10.0) | | |
| Graph organization by prompt type | 1 | 0.03 | 0.00 | 0.87 |
| Graph organization by prompt type by subjects (within-group error) | 37 | (7.9) | | |

*Note*. Values enclosed in parentheses represent mean square errors.
[a]Due to technical difficulty, one participant's latency was not obtained.
*$p < 0.05$.

Similar results were obtained for holistic scores (see Table 5). An identical 2 x 2 repeated-measures ANOVA revealed a significant effect of graph organization: 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). Again, although statistically significant, the effect size was small. The main effect of prompt type was not significant nor was the interaction of graph organization and prompt.

The effects of graph organization on response latency and holistic scores were also observed in the subsample of seven native English speakers, albeit attenuated due to ceiling effects. Native speakers were quicker to respond to few-chunks graphs (3.6 sec) than to many-chunks graphs (4.2 sec) and produced better responses to those with few-chunks (60.7 versus 59.5). These trends are consistent with the idea that the effects of graph organization are not limited to nonnative speakers of English, and suggest a degree of generality of the results.

**Table 5**

***Holistic Score ANOVA***

| Source | *df* | *F* | $\eta^2$ | *p* |
|---|---|---|---|---|
| Graph organization | 1 | 8.08* | 0.18 | 0.007 |
| Graph organization by subjects (within-group error) | 38 | (12.1) | | |
| Prompt type | 1 | 0.34 | 0.01 | 0.56 |
| Prompt type by subjects (within-group error) | 38 | (16.9) | | |
| Graph organization by prompt type | 1 | 0.00 | 0.00 | 0.96 |
| Graph organization by prompt type by subjects (within-group error) | 38 | (27.5) | | |

*Note*. Values enclosed in parentheses represent mean square errors.
*$p < 0.01$.

Which components of participants' language ability tend to be affected by graph organization? Are 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 working memory load, thus, impinging on all language competencies.

Table 6 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 demonstrate their awareness of audience and situation (sociolinguistic competence), and in their pronunciation, grammar, and fluency as a whole (linguistic competence).

**Table 6**

***Mean (SD) Scores by Graph Type***

| Competence component | Graph organization | | $t(37)$ | $p$ | $d^a$ |
|---|---|---|---|---|---|
| | Few-chunks | Many-chunks | | | |
| Discourse | 47.1 | 45.3* | 2.2 | 0.03 | 0.19 |
| | (8.6) | (9.9) | | | |
| Functional | 47.1 | 45.8 | 1.7 | 0.11 | 0.14 |
| | (8.7) | (9.9) | | | |
| Sociolinguistic | 46.2 | 45.5 | 1.5 | 0.14 | 0.08 |
| | (8.8) | (9.1) | | | |
| Linguistic | 48.0 | 47.2 | 1.0 | 0.30 | 0.09 |
| | (8.8) | (8.5) | | | |

*Note*. Each graph type score is the mean of the two scores for each participant. $N = 38$ per cell because one participant's component scores were unavailable.
[a]Cohen's *d* is a measure of effect size calculated as the difference between the two means divided by the pool standard deviation. Cohen (1988) refers to effect sizes of 0.20 and below as "small."
*$p < 0.05$.

Thus far, the results are consistent with the hypothesis 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 9, 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 (see Table 7). 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 (the relative-height visual chunks) of the many-chunks graph ($\chi^2(1) = 11.8$, $p < 0.001$).

**Table 7**

***Graph Type by First Description***

| Graph type | Global comparison | Discrete comparison |
|---|---|---|
| Few-chunks (Figure 1) | 19 | 1 |
| Many-chunks (Figure 9) | 8 | 10 |

## *Discussion*

The results provide further 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 Carpenter and Shah's (1998) results. Graphs with fewer chunks also elicited descriptions of better overall 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. Interestingly, incorporating a directive prompt had no influence on participants' descriptions. Although it is dangerous to draw conclusions from null results, this lack of effect is consistent with the idea that visual chunks are a visual processing phenomenon and might not be influenced by directions on problem-solving strategy.

The results for graph organization (number of visual chunks) might seem modest compared with the results of the regression analysis presented earlier. While statistically significant, the effect sizes were small (all less than .20): one might reasonably ask whether a 1.3 second difference in latency and a 1.6 score difference reflects an effect that is significant for practice. However, in interpreting these results, one should keep in mind two facts regarding the experimental situation. First, the contrasts were between highly similar graphs—the typical TSE graphs differ more widely than those used in the current experiment. The regression analysis presented earlier showed a much wider range of visual chunks leading to large differences in item quality. Second, the experiment deliberately included controls (two versions of the same data set, within-subjects comparisons) that might minimize the effects. These controls helped rule out alternative explanations of the results, allowing us to focus on the main purpose: to test whether, consistent with the visual chunks hypothesis, the quality of responses could be affected by the graphs *independent of the relevant construct* (communicative competence).

Furthermore, these small average differences mask a good deal of variability. For example, of the 39 participants, approximately half (19) demonstrated superior performance on the fewer-chunks graphs, ranging from 1.7 to 11.7 with a mean of 4.3. Thus, to the extent that fewer chunks lead to better performance for an individual, we might expect on average half a score band difference. While this would not affect people in the middle range of a band, this difference is important to examinees on the lower end of each score band. To the extent there is an effect, the number of visual chunks is certainly enough to shift, for example, a low 40 response to the 30 score band level.

A limitation of this study should be noted. For logistical reasons noted early, the experimental design had each school receive a unique order of graph items (see Table 3) instead of spiraling the order of test items within a school. This latter design would have minimized the likelihood that any results were due to an interaction between the particular order of items and the characteristics of the potential test-takers at a school. Whereas the within-in subject design lessens the chances of such an effect (because each participant acts as his or her own control), such a possibility is not eliminated. However, further inspection of the data provides support for our intended interpretation of the results. Within the subsample of each school, the mean scores on few-chunks graph items are always higher than the scores on the many-chunks items, replicating the overall result reported earlier. Furthermore, at each school, approximately half

of the participants (4–6) performed better on the fewer-chunks graphs, consistent with the analysis presented in the previous paragraph. Thus, it appears unlikely that the effects reported in the Results section are due to unintended differences between participants at the different schools.

## General Discussion: Recommendations

The visual chunks hypothesis—fewer visual chunks leading to descriptions of higher communicative quality—has practical implications, suggesting desirable characteristics of graph description tasks for the Test of Spoken English. For example, two or three 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 mean limiting the number of bar-groups placed along the x-axis. Other recommendations include:

- **Line graphs**. Line graphs should contain no more than three visual chunks. These chunks may consist either of an individual line with one reversal or two nonparallel lines with no reversals.
- **Bar charts**. Bar charts come in two varieties based on the x-axis scale. If the x-axis is a continuous scale, then the bar chart may be treated as a line graph, following the same recommendations provided above. If the x-axis is a discrete scale (i.e., nominal categories), there should be no more than two or three categories, and each of those categories should contain a visual pattern that can be described simply. If each group contains only two bars, each pair should be either clearly equal or clearly unequal. If each group contains more than two bars, either the bars should form a trend (assuming the different bars are on a continuous scale, such as years) or one of the bars should be much greater or much lesser than the others in the group (creating a clear maximal or minimal value).

Note that these specific, empirically supported recommendations leave open a variety of visual design elements in creating graphs. Within these guidelines, graphs may have whatever content that test development guidelines deem appropriate for the examinee population. Different types of graphs might be combined in a single item, such as presenting both a (one chunk) bar graph and a (one chunk) line graph as was done once in the past. Thus, the recommendations, while simply stated, are not as restricted as their brevity might imply.

To avoid discrepant items, test development should follow the prescriptions of the visual chunks theory, creating graphs that present only two or three chunks of information. This approach should lead to graph items that uniformly contribute to the assessment of communicative competence.

## References


Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford, England: Oxford University Press.

Carpenter, P. A., & Shah, P. (1998). A model of the perceptual and conceptual processes in graph comprehension. *Journal of Experimental Psychology: Applied, 4*(2), 75–100.

Carswell, C. M. (1993). Stimulus complexity and information integration in the spontaneous interpretations of line graphs. *Applied Cognitive Psychology, 7*(4), 341–357.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Mahwah, NJ: Erlbaum.

Douglas, D., & Smith, J. (1997). *Theoretical underpinings of the Test of Spoken English revision project* (ETS RM-97-02). Princeton, NJ: ETS.

ETS. (2001). *TSE and SPEAK score user guide*. Princeton, NJ: Author.

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*(2), 124–158.

Kosslyn, S. M. (1989). Understanding charts and graphs. *Applied Cognitive Psychology, 3*(3), 185–225.

Lewandowsky, S., & Behrens, J. T. (1999). Statistical graphs and maps. In F. T. Durso (Ed.), *Handbook of applied cognition* (pp. 513–549). Chichester, England: John Wiley & Sons Ltd.

Lohse, G. L. (1993). A cognitive model for understanding graphical perception. *Human-Computer Interaction, 8*(4), 353–388.

Myford, C. M., & Wolfe, E. W. (2000). *Monitoring sources of variability within the Test of Spoken English assessment system* (TOEFL Research Report. No. 65, ETS RR-00-6). Princeton, NJ: ETS.

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

Powers, D. E., Schedl, M. A., Leung, S. W., & Butler, F. A. (1999). Validating the revised Test of Spoken English against a criterion of communicative success. *Language Testing, 16*(4), 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*(4), 690–702.

Shah, P., & Hoeffner, J. (2002). Review of graph comprehension research: Implications for instruction. *Educational Psychology Review, 14*(1), 47–69.

Tufte, E. R. (1983). *The visual display of quantitative information*. Cheshire, CT: Graphics Press.

Wainer, H. (2000). *Visual revelations: Graphical tales of fate and deception from Napoleon Bonaparte to Ross Perot.* Mahwah, NJ: Erlbaum.

Wang, X., Bradlow, E. T., & Wainer, H. (2002). *A general Bayesian model for testlets: Theory and applications* (GRE Board Professional Report No. 99-01P, ETS RR-02-02). Princeton, NJ: ETS.

## Notes

[1] This graph was created expressly for the experiment presented later in this report. To preserve item confidentiality, none of the graphs in this report are actual TSE items. Some figures contain graphs that have been used previously on the TSE, but for the report all content information (e.g., axes labels) were removed.

[2] The three samples in this discussion are portions of actual responses collected during the experiment described later. As will be explained below, all of the responses in the experiment were scored with respect to the four communicative competencies individually, in addition to the usually holistic scores. The samples were selected to illustrate each competence based on these component scores.

[3] As noted earlier, to protect the confidentiality of TSE items, the figures in this section contain no content related to the original TSE item on which they are based. Instead of "Year 1" or "Category 1" would be an actual year or category, such as the name of a college.

[4] We thank Hunter Breland for conceiving the notion of discrepancy analysis.

[5] Two students chose not to list their gender.

[6] Due to technical difficulties, one participant's data were lost, so one school contributed only nine students.

[7] The local institutions comprised Drexel University, The Ohio State University, University of Buffalo, Wayne State University.

[8] A stronger design would have been to spiral the four orders of items within each school. However, this improvement was not logistically feasible because the experiment was conducted in the context of a pilot for an ETS operational test.

**Test of English as a Foreign Language**
**PO Box 6155**
**Princeton, NJ 08541-6155**
**USA**

To obtain more information about TOEFL programs and services, use one of the following:

**Phone: 609-771-7100**
**Email: toefl@ets.org**
**Web site: www.ets.org/toefl**

73313-035289 • Y24E.05 • Printed in U.S.A.
I.N. 990114