Language Testing

http://ltj.sagepub.com

Context and content visuals and performance on listening comprehension stimuli

April Ginther

Language Testing 2002; 19; 133
DOI: 10.1191/0265532202lt225oa

The online version of this article can be found at:
http://ltj.sagepub.com/cgi/content/abstract/19/2/133

Published by:
SAGE
http://www.sagepublications.com

Context and content visuals and performance on listening comprehension stimuli

April Ginther *Purdue University*

The listening comprehension section of the TOEFL has traditionally involved audio presentations of language without accompanying visual stimuli. Now that TOEFL is computer based, listening comprehension items are being created that include both audio and visual information. A nested cross-over design (participants nested in proficiency, level and form) was used to examine the effects of visual condition (present or absent), type of stimuli (dialogues/short conversations, academic discussions and mini-talks) and language proficiency (high or low) on performance on CBT (Computer-based Test) listening comprehension items. Three two-way interactions were significant: proficiency by type of stimuli, type of stimuli by visual condition and type of stimuli by time. The interaction between type of stimuli by visual condition although weak, was perhaps the most interesting and indicated that the presence of visuals results in facilitation of performance when the visuals bear information that complements the audio portion of the stimulus.

1 Introduction

1 Background for the study

In an effort to use multimedia computer technology to develop a more realistic and valid test of communicative competence, the Listening Comprehension section of the Test of English as a Foreign Language (TOEFL) now includes visual accompaniments to verbal stimuli. The primary purpose of this study was to examine the effects of the presence or absence of different types of visual accompaniments (context or content) in association with different types of Listening Comprehension stimuli. The stimulus types were Dialogues/Short Conversations, Academic Discussions and Mini-talks.

Current test development efforts involve the production of Listening Comprehension items that include still photos, drawings and pictures. The decision to include visuals was made for several reasons.

Address for correspondence: April Ginther, Assistant Professor, Department of English, Purdue University, 1356 Heavilon Hall, West Lafayette, IN 47907, USA; email: aginther@purdue.edu

 10.1191/0265532202lt225oa

Presenting examinees with a blank screen in the CBT environment was considered inappropriate, and the introduction of visuals was intended to enhance the face validity of the test. Most importantly, item stimuli including visual accompaniments to the audio text are considered better representations of actual communicative situations, so the inclusion of visuals may enhance the measurement of the test-taker's listening comprehension. Although determining precisely how or why the inclusion of visuals affects the measurement of listening comprehension may be difficult, TOEFL CBT (Computer-based Test) Listening Comprehension stimuli are different from their pencil-and-paper counterparts, and performance on these items may involve previously untapped aspects of proficiency.

The most frequently used type of visual in the current CBT Listening Comprehension item pool provides information about the context in which the verbal exchanges occur. For example, a photo might portray the participants in a conversation in a particular setting. Such photos accompany all three types of Listening Comprehension stimuli. Broadly, the presence of these visuals serves two purposes: (1) to set the scene for the verbal exchange and (2) to cue examinees to a change in speakers in a conversation. As a group, these visuals are referred to as context visuals.

The second type of visual consists of a photo, graph or drawing that is related to the content of the verbal stimulus. As a group, these visuals are referred to as content visuals. Content visuals occur only as accompaniments to mini-talks; however, their presence is confounded in the TOEFL pool because content visuals are typically accompanied by one or more context visuals. Furthermore, as TOEFL test developers are reluctant to test 'visual comprehension', it may be safe to assume that even when the inclusion of a content visual seems reasonable, a visual depicting the context of the mini-talk is preferred.

2 Characteristics of the TOEFL CBT Listening Comprehension Pool

Four combinations of audio and visual conditions represent the CBT Listening Comprehension item pool at the present time:

1) dialogues with a context visual (a still photo of the speakers and setting);
2) short conversations with context visuals (a series of still photos of the speakers and setting);
3) academic discussions with context visuals (a series of still photos of the speakers and setting);
4) mini-talks with context visuals (a series of still photos of the

speaker and setting) *and/or* content visuals (photos, diagrams and/or drawings related to the content of the audio portion of the stimulus).

Confounds among types of visuals (context and content) and type of the audio stimulus – as in (4) above – are not entirely negative. Current trends in test development reflect efforts to capture the complexity of actual situations in which language is used and to exploit the technology available at the present time. Nevertheless, the purposes the visuals are intended to serve do not have one-to-one correspondence with particular item types, and there has been no empirically-based investigation of actual as opposed to intended effects.

The listening section is adaptive; however, each examinee is presented with a minimum of 30 items in the following configuration: 11 dialogue items, 4 short conversation items divided across two sets, 3 academic discussion items in one set and 12 mini-talk items in three sets. Given that dialogues, short conversations and academic discussions appear only with context visuals – along with test development's preference for context visuals in the mini-talk condition – the result is that the majority of the visuals in the pool are context related. Because pre-test items are also included in the Listening Comprehension section and that the adaptive algorithm may add additional sets, the actual number of items an examinee receives ranges from 30 to 50.

Stimuli differ with respect to the type of accompanying visuals, the length of the audio text, and the number of items presented. The tasks associated with items (most often multiple-choice questions) depend on comprehension of either the 'gist' or specific details presented in the verbal portion of the stimulus. When the visuals are related to the content of the message, the relevant information may also be present in the visual part of the message. Items can also be differentiated depending on whether the relevant information is explicitly stated in the verbal message or is non-explicit and must be inferred. Thus, a particular question might be gist/explicit, gist/non-explicit, detail/explicit or detail/non-explicit. In all cases, the presentation of the first visual occurs simultaneously with the beginning of the verbal text.

Dialogues and Short Conversations contain the shortest verbal stimuli, usually a total of five turns by two speakers who are always a male and a female. These stimuli are accompanied by a single still photo of the participants in the setting where the conversation is supposed to be taking place. Dialogues are accompanied by a single item while Short Conversations are accompanied by a set of two or three

items. The only information that the visual stimulus provides pertains to the setting of the dialogue or short conversation and the sex of the speakers.

Academic Discussions can be described as short conversations about a topic related to course material. Verbal stimuli are accompanied by visuals that contain information only indirectly related to the content in the text: still photos of 'speakers' are used to set the context of the communicative event and to mark turns in a conversation. The verbal portions are timed to be about 2 minutes in length and may involve 2 or 3 turns by different speakers. As many as five still photos may appear, but if the identity of the speaker is clearly marked by verbal qualities such as tone of voice, then only one photo may appear. This is the case even when there is more than a single speaker of the same sex. For example, if a male professor is speaking to two students (male and female), one still photo of the scene will suffice if the male voices are distinguished by 'maturity'.

The verbal stimulus materials classified as mini-talks are based on the broad content areas of physical sciences, social sciences, life sciences and the arts. The verbal portions of the stimuli last approximately 2. minutes and are accompanied by 1–4 photos. The function of the context visuals is to 'set the scene' as the visuals provide information about the setting in which the verbal portion of the stimulus is presented. Mini-talks may also be accompanied by 1–4 content visuals. The content visuals that accompany mini-talks (photos, illustrations and/or diagrams) complement information presented in the text. Mini-talks are followed by 4–6 items on the TOEFL. Listening Comprehension specifications for the TOEFL require that one item be a gist item and one be based on non-explicitly stated information. The remaining items can involve any combination of gist/detail and explicit/non-explicit information.

3 Research questions

This study was designed to provide preliminary answers to the following questions:

1) Do participants perform better on the test items when visuals accompany the audio text?
2) Does the presence of visuals interact with stimulus type? If visuals play different roles in different types of TOEFL Listening Comprehension stimuli, this result may help identify the functions that visuals can be intentionally designed to perform.
3) Is an effect on performance related to participants' English proficiency? Some theorists have argued that visuals should have the

strongest effects on the performance of less-skilled participants because they make the task easier. Others have argued that if the complexity of the stimulus is increased with the addition of visuals, less-skilled participants will be unable to take advantage of their presence, and their performance will be debilitated. Experimental results appear to support the former interpretation. However, 'skill' in this context refers to prior knowledge. Can the notion of skill be related to language proficiency?

The focus of this study involves a fundamental concern: does the presence of visuals affect performance on language examinations? The potential of types of visuals to interact with the texts and tasks typical of language examinations can be argued to be an unimportant concern if the presence of visuals produces no effect. If the presence of visuals produces a positive effect on performance, there are at least two possible interpretations:

1) a positive effect could provide the basis for an argument that visual representations of communicative situations play a part in the processing of audio information and should be included in high-stakes examination stimuli; alternatively
2) a positive effect might be considered representative of an artificially upward bias with respect to performances and therefore visuals should not be included in high-stakes examination stimuli.

Given that language users (barring visual impairment) typically process language within some visual context, the former appears more reasonable. However, it is important to remember that by pairing frozen still photos with continuous audio texts, the TOEFL Program has created a mixed-media item type that is unusual. If the presence of visuals can be associated with debilitation of performance, their presence in high-stakes examinations becomes more difficult to justify if we intend to 'bias for best' (Duran, 1989).

II Related literature

It is well established in the literature that the effects of visual stimuli on the comprehension of verbal stimuli depend on several factors, including the task (Weidenmann, 1989), the kinds of visual materials used (Mastropieri *et al.*, 1987; Bauer and Johnson-Laird, 1993), the characteristics of the learners (Parkhurst and Dwyer, 1983; Hegarty and Just, 1993) and the interaction of these factors (Salomon, 1979; Dwyer and de Melo, 1984; Weidenmann, 1989; Moore and Dwyer, 1994). The consideration of possible interactions among content and

visual characteristics creates what Cronbach (1975: 119), in a discussion of the interpretation of interaction effects, referred to as 'a hall of mirrors that extends to infinity'.

As stated above, one of the advantages that computer-based testing is thought to offer TOEFL is the ability to represent and test communicative competence more effectively. In discussions of the nature of communicative competence, emphasis is consistently placed on the language learner's ability to use language effectively within and across different contexts. Nevertheless, in spite of at least 20 years' concern about the influence of context on language performance, no generally agreed-upon characterization of context has emerged. One well-known attempt to capture the potentially influential features of context on the performance of language users is provided by Hymes (1974) in the following mnemonic:

> **S**ituation: the meaning the participants attribute to the physical and temporal setting, psychological and cultural scene
> **P**articipants: the role(s) they think they should take in the interaction
> **E**nds: outcomes and goals they attribute to the exchange
> **A**ct sequence: message form and content they think they should attend to
> **K**ey: the tone and the manner they think appropriate
> **I**nstrumentalities: channels, codes and registers they think appropriate
> **N**orms: norms of interaction and interpretation they think are called for
> **G**enres: categories of speech events they think they are engaged in

Given this all-inclusive characterization of context as a starting point, it is easy to understand why no consensus has emerged. Researchers are free to emphasize the aspects of context most germane to their particular interests, and the diverse interests of researchers have lead to multiple terms and emphases. As Douglas (1997: 15) observes, 'Context has been defined in various ways by researchers and has been associated with such notions as "situation," "setting," "domain," "environment," "task," "content," "topic," "schema," "frames," "script," and "background knowledge".' Schegloff (1997: 165–66), in an address to the American Association of Applied Linguistics (AAAL), adds to the list of potentially relevant features:

> the ways of formulating the context within which something occurred are multiple . . . [For example] if one had to characterize what I am doing at the moment, one might say I am presenting a paper, introducing my remarks, reading a text, arguing a point of view, responding to our chair's invitation, gesticulating occasionally and suppressing gesticulation mostly, managing recurrent eye contact with members of the audience, and many others. And a similar stricture can be introduced here: none of these characterizations can get an adequate warrant by saying that it was employed because it is *true* – even though it *is* true. They are *all true.*

But if all aspects of context are, in some sense, true, which aspects of context should be emphasized in representations of those contexts?

TOEFL's solution has been to emphasize the first two features of Hyme's characterization of context: situation and participants. Representing these features of context can be argued to enhance the 'situational authenticity' (Bachman, 1991) of the exam by allowing the examinees to orient themselves to the appropriate domain of language use. It is assumed that having the proper orientation provides information that facilitates comprehension and interpretation. The frequency of context visuals as compared to content visuals in the TOEFL CBT item pool appears a reflection of the extent to which both applied linguists and test developers value the representation of context. In specific testing situations, however, the effects of such representations remain undetermined.

A substantial body of literature exists on the effects of the presence of visuals on written text comprehension (e.g., Willows and Houghton, 1987; Mandl and Levin, 1989; Winn, 1991); however, the focus is primarily on the effects of visuals that are directly related to the content, as opposed to the context, of written text. Similar to the literature involving investigations of context, the literature that focuses on the effects of content visuals has not led to a consensus or an overarching theoretical framework with respect to the interaction of graphic and textual sources of information (see Mayer, 1993a).

The absence of a clearly applicable theoretical perspective is directly related to the complexity of the combinatory effects of textual and visual stimuli. Salomon (1989) argues that one of the reasons that no clear theoretical perspective has emerged is because, too often, researchers have attempted to isolate effects, and he states that different effects are produced when different combinations of variables enter into a particular research design. Salomon (1989) criticizes the most common research paradigms, explaining that:

1) approaches based on empirical findings related to the surface features of materials do not offer a basis for explanation in and of themselves because such findings tend to be inconsistent;
2) approaches that conceptualize independent variables in terms of underlying cognitive processes invite circular reasoning because of the difficulty of distinguishing, for example, the 'complexity' of the stimulus from 'complexity' of the cognitive process invoked; and
3) approaches adopting a purely phenomenological stance based on person–situation interactions make objective classification of stimulus materials impossible.

Salomon (1989) proposes an approach that attempts to explain effects of pictures in terms of the interactions among surface features of the stimulus materials, underlying cognitive functions, and person

and task characteristics. This approach is based on three underlying assumptions:

> what figures in learning from text and pictures is not the surface features of stimuli or their combinations but the cognitive functions they accomplish . . . ; (2) texts and pictures can potentially accomplish a variety of cognitive functions as a result of their specific symbolic, informational, and configurational nature . . . However, whether different stimulus configurations accomplish these functions in actuality depends not only on the nature of the stimulus materials but also on person, situation, and task variables . . . ; and (3) a cognitive function accomplished in actuality does not guarantee better learning. [Again] a variety of person, situation, and task variables [are involved]. (p. 77)

He continues by stating that at least five clusters of variables are involved in the process of learning from text and pictures. The five clusters are:

> (a) Stimulus variables: the nature of the pictures, the texts, and their interrelations; (b) Cognitive variables: the processes that are called upon, are required, and become employed in actuality; (c) Person variables: the nature of the learner's relevant abilities, prior knowledge, motivations and perceptions of the stimuli, of self, of the situation and the task; (d) the psychological functions accomplished: summary, introduction, explication, memory pegs, supplantation, and the like; and (e) Task variables: the nature of the tasks to be accomplished. (p 74)

Salomon concludes that developing an explanatory model of picture and text comprehension should attempt to take all of these variables into account. Obviously, this could never be accomplished by a single study but rather would require a series of related investigations.

Nevertheless, Salomon does present an overarching theoretical concept for the integration of these variables: visual supplantation. He states:

> According to this formulation, explicit pictoral presentations of a process, or of an intermediate state in a process, may overtly model (that is – supplant) the kind of imagery that learners should have conjured up on their own, assuming of course that such imagery is necessary for the acquisition of the material to be learned. To the extent that learners cannot conjure up such imagery on their own, visual supplantation should facilitate their learning. (p. 77)

It is important to note that in order for visual supplantation to occur and to have a positive effect on performance, there must be a complementary relationship between information presented in the visual portion of the stimuli and the information presented in the audio portion of the stimuli. Thus, the crucial determinant of the effects of the visual stimulus is the relationship between the information contained in the visual and that contained in the verbal (oral or written) text.

Finally, what Salomon refers to as 'person' effects at this point must also be considered. Salomon (1989) states:

> if particular stimuli are capable of supplanting learners' imagery, then it follows that learners who, for whatever reason, cannot generate their own will benefit from these stimuli; not so for learners who can and do conjure up their own mediating images. Indeed, the results of a series of experiments (Salomon, 1979) consistently showed aptitude-by-treatment interactions: students with a poor mastery of the supplanted skills benefited from the visual supplanting treatment, whereas those with better mastery showed clear signs of debilitation, a possible result of interference. (p. 79)

Salomon's emphasis on the complexity of the possible interaction effects does not comprise a model *per se*, but it does provide a perspective from which a theoretical framework might eventually be developed.

A series of studies that attempt to unravel the complexities of the potential interactions among text, graphic and individual variables have been conducted by Mayer and his collaborators (Mayer, 1984; Mayer, 1989; Mayer and Gallini, 1990; Mayer and Anderson, 1991; Mayer and Anderson, 1992; Mayer, 1993a; 1993b; Mayer and Sims 1994; Mayer *et al*., 1995; Mayer *et al*., 1996). Common to these studies is the examination of methods to improve students' comprehension of scientific text. These studies develop a perspective called the 'generative theory of multimedia learning' which Mayer (1997) summarizes as follows:

> In a generative theory of multimedia learning, the learner is viewed as a knowledge constructor who actively selects and connects pieces of visual and verbal knowledge. The basic theme of a generative theory of multimedia learning is that the design of multimedia instruction affects the degree to which learners engage in the cognitive processes required for meaningful learning within the visual and verbal information processing systems. (p. 4)

Given the focus on understanding scientific explanations, the dependent measure in these studies involves the ability to demonstrate the transfer of learning to the solution of new problems, rather than on recall or comprehension. The success of a student to integrate visual and verbal sources of information is examined in the generation of acceptable and creative solutions in different but related domains.

A consistent finding is that learners benefit when the presentation of visual and verbal information is contiguous (the contiguity effect). Mayer's contiguity effect parallels Salomon's argument that the information presented in text and visual sources must be complementary in order for facilitation to occur. Facilitative effects of the presentation of visual information were reduced when visual information appeared on separate textbook pages or on separate computer screens. Given that the contiguity effect was consistent across text-based or computer-based presentations, Mayer argues that the common research emphasis on the medium of presentation needs to be

rethought. He asserts that the commonalities across media are, most likely, more important than the medium itself.

In further examinations of the contiguity effect, Mayer and Gallini (1990) and Mayer *et al*. (1995) uncover interactions with what Solomon identified as 'person effects'. In a series of three experiments investigating both the appropriateness and creativity of solutions to problems as related to different multimedia conditions (presentations either possessing or lacking contiguity), Mayer and Gallini (1990) found that the effects of multimedia were strong for those participants without prior knowledge but weak or nonexistent for participants with prior knowledge; that is, when multimedia presentations were contiguous, low prior knowledge participants benefited. When presentations were not contiguous, neither the low prior knowledge or high prior knowledge groups benefited. This means that without consideration of presentation effects, the potential facilitation of multimedia presentations might be missed – and misunderstood.

Mayer *et al*. (1995) uncover a related interaction effect involving spatial ability. In another series of experiments, Mayer *et al*. (1995) found that participants with high spatial ability were those who benefited from contiguous multimedia presentations. Low spatial ability participants did not benefit from contiguous multimedia presentations. Mayer (1997) explains this result as follows:

> Students who possess low levels of spatial ability may be less able than high spatial ability learners to take advantage of contiguous presentation of visual and verbal material because they have more difficulty in holding and manipulating the visual representation in memory, as is required to integrate the visual and verbal representations. In contrast, students who possess high levels of spatial ability may be more likely than low spatial ability learners from contiguous presentation because they are more facile at holding and manipulating visual representations in memory . . . (p. 16)

This series of studies demonstrates that the effects of contiguous presentations of visual and verbal information are more effective for students with low levels of prior knowledge and high levels of spatial ability.

The final experiment that will be discussed here is perhaps most germane to the present purposes. Mayer (1997) reports the findings of a pilot study which examined appropriateness and creativity of solutions to problems across two conditions: text with visuals vs. narration with visuals. In both conditions, the visuals were animated. Mayer cites Sweller *et al*. (1990), Chandler and Sweller (1991) and Baddeley (1992) with respect to split-attention theory and working memory to point out that verbal information may be processed differently when presented as text (visually) or as narration (acoustically). He explains:

> In particular, when text and animation are both presented visually, the learner's visual attention must be split between the animation and the text. When visual attention is overloaded, some of the information may be lost and the process of constructing connections between visual and verbal information will be disrupted. In contrast, when text is presented auditorily and the corresponding animation is presented visually, the learner can process the representation of text within an acoustic working memory and the representation of the animation within a visual working memory, which reduces the load on attention. This situation increases the chances that the learner will be able to construct connections between visual and verbal representations of the causal chain. (p. 17)

Findings of the pilot confirm this prediction. Participants who were exposed to the animation accompanied by narration produced approximately 50% more creative solutions than did the participants who were exposed to the animations accompanied by text. While Mayer argues that this finding suggests that narrations may be more effective than texts in multimedia programs, he also argues that researchers' lack of attention to differences in memory load across different types of tasks in multimedia learning environments is a critical lapse.

Finally, a note on the systems used to classify visuals in the studies reviewed above is important. The development of a theoretical framework that would allow consistent categorizations of graphically-presented information is a primary concern of the work of Salomon. The emphasis that his work shares with that of Mayer is that the classification system is of interest in order to explicate the relationship between visual and text-based sources of information and learning.

Because the work of Mayer and his collaborators addresses the relationship between visual- and text-based sources of information within a paradigm of contiguity, the concern is not so much with the specific role a visual might play but rather on the extent to which the sources of information are parallel. If the text involves an explication of the sequence of moves that a pump goes through in moving a liquid from one location to another, the visuals are designed to follow this information. Thus the function of the text determines the functions of the associated visuals.

The emphasis on instruction is problematic for investigations of the effects of visuals on TOEFL Listening Comprehension. It might be argued that, because the presentation of visuals occurs simultaneously with the presentation of the text-based information, their presence should not produce debilitation of performance. That is, the visuals in the TOEFL listening comprehension pool do, in some sense, meet Salomon's minimal requirements for visual supplantation and/or Meyer's minimal requirements for meaningful learning. Visual presentations are always, in some sense, contiguous; therefore, we should anticipate facilitation. While contiguity may be a necessary

condition for facilitation to occur, it is extremely important to keep in mind that a direct relationship between content-based information in a visual and in the text is always assumed in the studies of Salomon and Meyer. None of the studies reviewed considered the use of visuals to provide information about the context of the verbal exchange as opposed to content information.

The studies reviewed above suggest that participants may benefit from the presence of content-based information by using the information presented in a content visual to confirm the inferences they make as they read or listen to a text. Content-based information in a visual may facilitate the retrieval of related information when a particular content-based question is asked or a particular content-based performance is required (problem solving in extended domains). However, if the only information presented in the visual is context-based, it may be the case that the presentation of that information creates noise – a distraction – when content-based information is sought.

III Methodology

This section of the paper addresses the following aspects of the study: characteristics of the participants, characteristics of TOEFL Listening Comprehension stimuli, characteristics of the experimental stimuli, design of the study, administration of the experimental materials and observation during administration.

1 Participants

The number of participants participating in the study was 160. These participants were recruited from ESL programs associated with a large state university in the Midwest. Participation was voluntary, and each participant was paid US$40 after completion of the experiment. Students volunteered by responding to flyers and announcements. All of the participants completed the experiment during the spring semester of 1998.

Given the potential for proficiency to interact with the presence or absence of visuals, recruitment efforts targeted two levels of proficiency, high and low. All 80 participants classified as high proficiency had been admitted into the university. (The university requires a minimum total TOEFL score of 550 for admission to both undergraduate and graduate programs.) The high proficiency group consisted of 40 undergraduate and 40 graduate students.

The 80 participants classified as low proficiency were recruited from the Married Student Housing English as a Second Language

Program or were visiting scholars. None of these participants had been admitted into the university, and none had been in the USA for longer than two months. (It was assumed that these participants would be lower in English proficiency than the participants who had been admitted into regular university programs and the validity of this assumption was confirmed by the significance of the main effect for proficiency.) As virtually every person who was enrolled in the Married Student Housing ESL Program and who had been in the USA for less than two months was recruited and volunteered, there was a deficit of 14 participants at the low proficiency level. Fourteen additional participants who were visiting scholars were then accepted for participation at the low proficiency level.

The nested effect within proficiency – high (undergraduate or graduate) and low (married student housing or visiting scholar) – is referred to as status and was tested. Table 1, Proficiency by status by sex, presents participants' proficiency, the status of participants within proficiency groups and the sex of the participants within status. Of the 160 total participants, 73 were male and 87 were female. Within the low proficiency group, females predominated (see Table 1). These participants were primarily the female spouses of male graduate students. Within all other groups, males outnumbered females.

The range in age of the participants who were not enrolled was 17–52; median = 29. The range of age in the participants who were visiting scholars was 23–43; median = 33. The range in age of the undergraduate participants was 17–32; median = 20. The range in age of the graduate participants was 21–46; median = 28.

Participants were asked to indicate the highest degree they had obtained in their native countries, and these responses are reported in Table 2, Highest degree earned by proficiency by status. Of the participants who were recruited from the Married Student Housing

Table 1 Proficiency by status by sex

Proficiency								
Low = 80				High = 80				
Status								
Married student ESL program		Visiting scholar		Undergraduate		Graduate		
66		14		40		40		
Sex								
M	F	M	F	M	F	M	F	
17	49	8	6	25	15	23	17	

Table 2 Highest degree earned by proficiency by status

Highest degree earned in native country	Proficiency			
	Low		High	
	Status			
	Married Student ESL Program	Visiting scholar	Undergraduate	Graduate
Elementary	9			
High School	11	1	36	
Associate's	1			
BA/BS	26	6	3	26
MA/MS	14	5	1	12
PhD	3	2		2
MD	1			
Law	1			
Total	66	14	40	40

ESL Program, only 9 had not completed a secondary or high school degree in their native countries, and an additional 11 had not completed a college degree. (In the USA, a secondary or high school degree corresponds to the 9th–12th years of public or private schooling; students are typically 14–18 years of age.) The remaining 46 participants had completed, at least, a college degree, indicating that, in spite of their assumed lower English proficiency, these were highly educated participants.

As expected of the 40 undergraduates in the high proficiency group, the vast majority (36) reported that they had not completed an undergraduate degree in their native countries. All of the graduates in the high proficiency group had completed, at least, an undergraduate degree in their native countries.

The native languages of the participants in both the low and high proficiency groups are presented in Table 3, Native language by proficiency. Native Chinese-, Hindi- and Indonesian-speaking participants predominated in the high proficiency group; Native Korean-, Chinese- and Spanish-speaking participants predominated in the low proficiency group. A balanced sample with respect to native language would have been preferable; however, given the dearth of potential participants at the low proficiency level along with the university's prohibition against the recruitment of individuals with specific characteristics, acceptance of the volunteers was the only option available.

Table 3 Native language by proficiency

Native language	Proficiency	
	Low	High
Arabic	5	1
Bambara	1	
Bengali		1
Bulgarian		1
Chinese	19	21
Dutch		1
Farsi	1	
German		4
Greek		1
Hindi	2	9
Indonesian	2	9
Italian	1	1
Japanese	4	4
Korean	22	6
Luo		1
Malay	1	4
Memde		1
Ndebele		1
Persian	2	
Polish	1	
Portuguese	2	2
Romanian		1
Russian	2	
Serbian		1
Setswana		2
Spanish	13	4
Telugu	1	
Turkish		1
Ukranian	1	
Urdu		1
Total	80	80

2 *Experimental materials*

In the case of the TOEFL CBT Listening Comprehension items currently in use, uncertainty about the identification of features which might critically affect performance along with the complexity of the interaction effects might encourage a conservative experimental approach. Variables of interest could be systematically manipulated and all external variables controlled as carefully as possible. While attractive, this approach would require considerable time and expense, and given that visuals are already in use, the results would be belated at best. An alternative approach is to study the effects of visuals in the current pools of CBT items, and to identify, albeit broadly, the

theoretically important variables in these items. This is the approach that has been adopted for this initial study.

Fortunately, the four types of Listening Comprehension items in the TOEFL CBT item pool are accompanied by visuals that provide different kinds of information, thus enabling a quasi-experimental, post-hoc manipulation of the information contained in the accompanying visual stimuli. First of all, dialogues and short conversations were combined as they comprised a single relatively short stimuli type accompanied by a single visual. Secondly, in order to disambiguate the effects of the type of visual that accompanies mini-talks, separate sets of mini-talks were created: mini-talks with context visuals and mini-talks with content visuals. This differentiation resulted in four types of experimental stimuli:

1) Dialogues/Short Conversations with context visuals (a still photo of the speakers);
2) Academic Discussions with context visuals (still photo(s) of the speakers);
3) Mini-talks with context visuals (still photo(s) of the speaker);
4) Mini-talks with content visuals (photos, diagrams, and/or drawings related to and contiguously presented with the content of the audio portion of the stimulus).

For the study, the four sets of experimental stimuli described above were divided into two subsets of stimuli and items so that the sets could be presented with or without visual accompaniments. Two of the subsets were Dialogues/Short Conversations (with or without context visuals), two were Academic Discussions (with or without context visuals), two were Mini-talks (with or without context visuals) and two were Mini-talks (with or without content visuals).

Table 4, Design of the study, has been provided to aid in the comprehension of the characteristics of the stimulus sets as well as in the design of the study. Column 1 of Table 4 identifies the stimulus type, subset and type of visual and is labelled 'Stimulus type/Subset/Type of visual'. Notice that there are four types of experimental stimuli (Dialogues/Short Conversations, Academic Discussions and two types of Mini-talks, i.e., Mini-talks with context visuals and Mini-talks with content visuals). Within each stimulus type, there are two subsets of stimuli and associated items, e.g., Dialogue/Short Conversation Subset 1 and Dialogue/Short Conversation Subset 2. Type of visual identifies whether the visual(s) represented context or content-related information when visuals were present.

Column 2 of Table 4, labelled 'Item/topic' identifies the subset, associated items and item topics. Note that there are five items associated with Dialogue/Short Conversation Subset 1, and these items

Table 4 Design of the study

Stimulus type (Subset: *type of visual*)	Item/topic	Visual condition			
		Forms: 1, 2, 5, 6, 9, 10, 13, 14 ($n = 80$)		Forms: 3, 4, 7, 8, 11, 12, 15, 16 ($n = 80$)	
		Proficiency			
		LP ($n = 40$)	HP ($n = 40$)	LP ($n = 40$)	HP ($n = 40$)
Dialogue/Short Conversations (Subset 1; *Context*)	D/SC 1.1: The wind D/SC 1.2: Biology class D/SC 1.3: Library card D/SC 1.4: Library card D/SC 1.5: Library card	V	V	NV	NV
Dialogue/Short Conversations (Subset 2; *Context*)	D/SC 2.1: Stat Course D/SC 2.2: Borrow book D/SC 2.3: Dinner party D/SC 2.4: Dinner party D/SC 2.5: Dinner party	NV	NV	V	V
Academic Discussion (Subset 1; *Context*)	AD 1.1: Art History AD 1.2: Art History AD 1.3: Art History AD 1.4: Art History AD 1.5: Art History	V	V	NV	NV
Academic Discussion (Subset 2; *Context*)	AD 2.1: Engineering AD 2.2: Engineering AD 2.3: Engineering AD 2.4: Engineering AD 2.5: Engineering	NV	NV	V	V
Mini-talk (Subset 1; *Context*)	MTX 1.1: Prairie Dogs MTX 1.2: Prairie Dogs MTX 1.3: Prairie Dogs MTX 1.4: Prairie Dogs MTX 1.5: Prairie Dogs	V	V	NV	NV
Mini-talk (Subset 2; *Context*)	MTX 2.1: Anthropology MTX 2.2: Anthropology MTX 2.3: Anthropology MTX 2.4: Anthropology MTX 2.5: Anthropology	NV	NV	V	V
Mini-talk (Subset 1; *Content*)	MTN 1.1: Aquifers MTN 1.2: Aquifers MTN 1.3: Aquifers MTN 1.4: Aquifers MTN 1.5: Aquifers	V	V	NV	NV
Mini-talk (Subset 2; *Content*)	MTN 2.1: Amber MTN 2.2: Amber MTN 2.3: Amber MTN 2.4: Amber MTN 2.5: Amber	NV	NV	V	V

Notes: V = Visual condition; NV = No visual condition.

are labelled D/SC 1.1 – D/SC 1.5. There are five items associated with Dialogue/Short Conversation Subset 2, and these items are labelled D/SC 2.1 – D/SC 2.5. Reading down column 2 of Table 1, note that there are two subsets of Dialogues/Short Conversations, D/SC1 and D/SC2; two subsets of Academic Discussions, AD1 and AD2; two subsets of Mini-talks with context-related information, MTX1 and MTX2; two subsets of Mini-talks with content-related information, MTN1 and MTN2. Each of the eight stimulus subsets was followed by five items. In the case of the Mini-talks, one non-standard item was included. These items required the examinee to order the presented information in the proper sequence. All of the remaining items associated with the mini-talks and other types of stimuli were standard multiple-choice items.

3 Design

A nested cross-over design was used. The design of this study may also be familiar to readers as a partially-crossed Latin Square (see Cochran and Cox, 1957). Participants were nested in form, proficiency and status. The partially-crossed and nested effects can be understood by considering the number of participants within the divisions associated with form, proficiency and status.

Each participant was administered 1 of 16 forms. The use of 16 forms allowed the effects associated with the order of presentation of stimulus types and visual conditions to be counterbalanced. Column 3 of Table 4, Visual condition, is divided first by form and then by proficiency. The 80 participants who were administered forms 1, 2, 5, 6, 9, 10, 13 or 14 (the first column under Visual condition) were presented with Dialogue/Short Conversation Subset 1 in the visual condition (V) and Dialogue/Short Conversation Subset 2 in the No visual (NV) condition. The 80 participants who were administered forms 3, 4, 7, 8, 11, 12, 15 or 16 (the second column under Visual condition) were presented with Dialogue/Short Conversations Subset 1 in the No Visual (NV) condition and Dialogue/Short Conversations Subset 2 in the Visual (V) condition. Thus, the presence or absence of visuals is partially crossed.

Within the visual condition, participants were nested in proficiency; i.e., 40 of the 80 participants who were administered forms 1, 2, 5, 6, 9, 10, 13 or 14 were low proficiency (LP) and 40 were high proficiency (HP): the first and second columns under 'Proficiency'. Correspondingly, 40 of the participants who were administered forms 3, 4, 7, 8, 11, 12, 15 or 16 were low proficiency (LP) and 40 were high proficiency: the third and fourth columns under 'Proficiency'.

Within each Type of Stimulus, 5 items were accompanied by visuals and 5 were not. Thus, each participant was administered 10 items in association with each of the four stimulus types ($10 \times 4 = 40$ items). Each participant was administered all of the 40 items associated with each stimulus, but 20 of the items were accompanied by visuals and 20 were not. For example, if low proficiency participant 1 is administered form 1, then he or she will answer the 40 items in the visual conditions presented in the first column under 'Proficiency' in Table 4. Low proficiency participant 2, administered form 3, will answer the 40 items in the visual conditions presented in the third column under 'Proficiency' in Table 4. The presence or absence of visuals in accompaniment to the stimulus types for any individual participant was determined by the assigned form (see Appendix 1, Test forms, for the actual configuration of each test form). Finally, the stimulus subsets within each type of stimulus were intended to have comparable content and difficulty; however, the items were not pre-tested and the assumption of comparable difficulty was based on the intuitions of the test developers who created the stimuli.

Using Cronbach's alpha, the estimated reliabilities for the scores on the 10 items associated with each type of stimuli are as follows: Dialogues/Short Conversations = .80; Academic Discussions = .72; Mini-talk (context) = .76; Mini-talk (content) = .70. Reliabilities were calculated collapsing across visual conditions.

All analyses of these data were conducted through an application of the General Linear Model on SAS (Statistical Analysis System, 1998).

4 Observations

Participants were observed as they worked through the experimental materials, and any irregularities were noted. One participant was excluded from the analysis because he fell asleep before he was able to complete the experiment. No other irregularities were noted. The time the participants required to complete the entire experiment ranged from 59–114 minutes.

IV Results

The model that was tested is presented in Table 5. The dependent variable was participants' scores on the 5 items presented with each stimulus subset (refer to Table 4). The order in which these effects are introduced follows the order of presentation in Table 5. The complexity of the design requires that Numerator Mean Squares and Denominator Mean Squares be presented in the source table (columns

Table 5 Model

Source	Num. SS	Num. Df	Num. MS	Den. MS	F	*p*	η^2	Partial η^2
Proficiency	325.50	1	325.50	4.92	66.15	.00		
Status (proficiency)	11.53	2	5.77	4.92	1.71	.31		
Form	49.11	15	3.27	4.92	.67	.82		
Subjects (prof*status*form)	693.80	141	4.92	.76	6.51	.00		
Stimulus Type	258.42	3	86.14	.76	113.92	.00		
Time	39.55	1	39.55	.76	52.31	.00		
Visual	.60	1	.57	.76	.75	.39		
Prof*StimType	8.12	3	2.71	.76	3.58	.01	.003	.009
Prof*visual	.09	1	.09	.76	.12	.72		
StimType*Visual	7.09	3	2.36	.76	3.12	.03	.003	.008
StimType*Time	62.11	3	20.70	.76	27.40	.00	.025	.06
Prof*StimType*Visual	1.15	3	.38	.76	.51	.68		
Error	833.27	1102						
Corrected total	2487.96	1279						

Notes: Num. SS = Number Sum of Squares; Num. Df = Number Degrees of Freedom; Num. MS = Numerator Mean Square; Den. MS = Denominator Mean Square.

4 and 5). In addition, eta square (η^2) and partial eta square have been included as measures of effect size or the proportion of variance accounted for by population membership (see Cohen, 1988).

The significance of the main effect for proficiency generally confirms assumptions about the proficiency levels of participants recruited from the Married Student Housing ESL Program as compared to the proficiency of undergraduate and graduate participants enrolled in the university. However, as stated above, there were different kinds of participants within the low and high proficiency groups; that is, the high proficiency group consisted of university-enrolled undergraduate and graduate participants, and the low proficiency group consisted of not-enrolled ESL students and not-enrolled visiting scholars. This nested effect, referred to as status, was not significant. Because proficiency was involved in a significant interaction with stimulus type, this effect will be discussed further only in the context of the interaction.

The main effect for form was not significant, indicating that the counterbalancing afforded by the use of 16 forms was adequate.

The significance of the main effect for participants indicates that a considerable amount of variation – in fact, the bulk of the variance accounted for by the model – was not accounted for by the other significant main effects and their interactions and thus remains unexplained.

The main effect of stimulus type (dialogues/short conversations, academic discussions, mini-talks with context visuals and mini-talks

with content visuals) was significant but was involved in a significant interaction with proficiency.

The main effect for time was significant and tested whether there was a practice or fatigue effect associated with order of presentation within type of stimulus. While the effects of presentation associated with visual accompaniments (visual vs. no visual) and subset (subset 1 followed by subset 2 vs. subset 2 followed by subset 1) were counterbalanced, the integrity of the stimulus type was not violated. That is, subsets within each type of stimulus (whether 1–2 or 2–1) were presented consecutively. (See Table 4 and Appendix 1.) Thus, the presence of a practice/fatigue effect within each type of stimulus could be tested. Again, because the effect of time is involved in a significant interaction with type of stimulus, these effects will be discussed in terms of their interaction.

The main effect for the presence or absence of visuals was not significant. Because the focus of the study was on the possible interaction among levels of proficiency, stimulus type and the presence or absence of visuals, the single three-way interaction that was tested was this interaction. This three-way interaction was not significant. However, given the patterns of results, it may have been more germane to test the significance of the three-way interaction among type of stimuli, visual condition and time. As the effect of time was not anticipated and not included in the original research questions, this interaction remains untested. However, the potential of certain types of visuals to ameliorate the effects of longer stimuli is an interesting possibility and will be briefly discussed in the following section.

While it may appear that the tested model is incomplete because of the absence of the interactions time by proficiency and time by visual condition, these interactions are confounded due to the partially-crossed nature of the design. That is, at time 1, half of the participants were low proficiency and half were high proficiency; correspondingly, at time 1, half of the participants were presented with stimuli accompanied by visuals and half were not. Because of the confounds, these effects were not tested.

The two-way interaction, stimulus type by proficiency, which averages across visual conditions, was significant. A graph of the interaction is presented in Figure 1. Reading Figure 1 from the top down, note that all of the sets were easier for the high proficiency group and that the rank of difficulty for both low and high proficiency groups remains the same: dialogues and short conversations are the least difficult (low proficiency, $\bar{x} = 3.78$, sd = 1.38; high proficiency, $\bar{x} = 4.87$, sd = .39), followed by mini-talks with content visuals (low proficiency, $\bar{x} = 3.32$, sd = 1.37; high proficiency, $\bar{x} = 4.51$, sd = .85), academic discussions (low proficiency, $\bar{x} = 2.79$, sd = 1.39; high

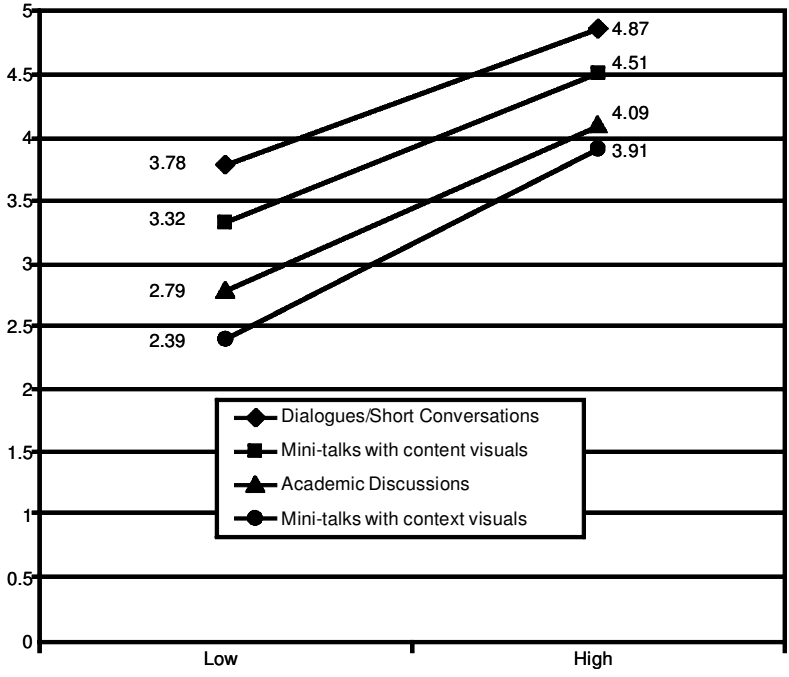

Figure 1 Stimulus type by proficiency

proficiency, $\bar{x} = 4.09$, sd = .94) and then mini-talks with context visuals (low proficiency, $\bar{x} = 2.39$, sd = 1.41; high proficiency, $\bar{x} = 3.91$, sd = 1.12). However, the differences in the means for academic discussions as compared to mini-talks with context visuals is less pronounced for the high proficiency group than for the low proficiency group. If the level of proficiency of the high group were extended, the rank order of Academic Discussions and Mini-talks with Context visuals would eventually reverse. Another way to consider the source of the interaction is simply to note that the lines associated with Academic Discussions and Mini-talks with context visuals are not parallel. This difference can be considered the source of the significant interaction.

The two-way interaction, stimulus type by visual condition, which averages across levels of proficiency, was significant. A graph of the interaction is presented in Figure 2. The presence or absence of visual accompaniments had virtually no effect when the content of the stimuli materials involved Dialogues/Short Conversations (no visuals, $\bar{x} = 4.31$, sd = 1.15; visuals, $\bar{x} = 4.34$, sd = 1.15). Both the Mini-talks with content visuals (no visuals, $\bar{x} = 3.79$, sd = 1.35; visuals, $\bar{x} = 4.03$, sd = 1.21) and Academic Discussions (no visuals, $\bar{x} = 3.39$, sd

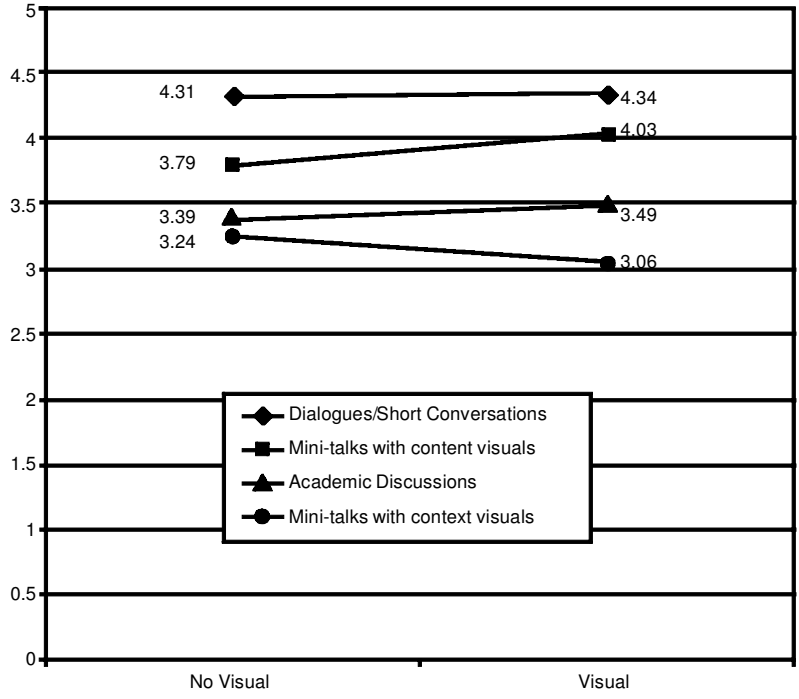

Figure 2 Stimulus type by visual condition

= 1.33; visuals, $\bar{x}$ = 3.49, sd = 1.36) were slightly less difficult in the visual condition. Mini-talks with context visuals (no visuals, $\bar{x}$ = 3.24, sd = 1.44; visuals, $\bar{x}$ = 3.06, sd = 1.52) were slightly more difficult when accompanied by visuals.

The two-way interaction, stimulus type by time was significant. This interaction collapses proficiency and visual conditions. Time was included in the model because participants were always presented with two subsets of stimuli within each set of stimuli, i.e., within each type of stimulus. The effect of time was tested to determine whether a practice effect was associated with order of presentation. A graph of the interaction is presented in Figure 3. A practice effect appears associated only with Dialogues/Short Conversations, the shortest stimuli type. When the Dialogue/Short Conversation sets were presented to participants, they performed better on the second set (1st presentation, $\bar{x}$ = 4.18, sd = 1.33; 2nd presentation, $\bar{x}$ = 4.48, sd = .90). The reverse is true for all of the longer stimuli sets. When the Mini-talks with content visuals were presented to participants, they performed better on the first set (1st presentation, $\bar{x}$ = 4.32, sd = .95; 2nd presentation, $\bar{x}$ = 3.49, sd = 1.43). When Academic Discussions were presented to participants, they performed better on the

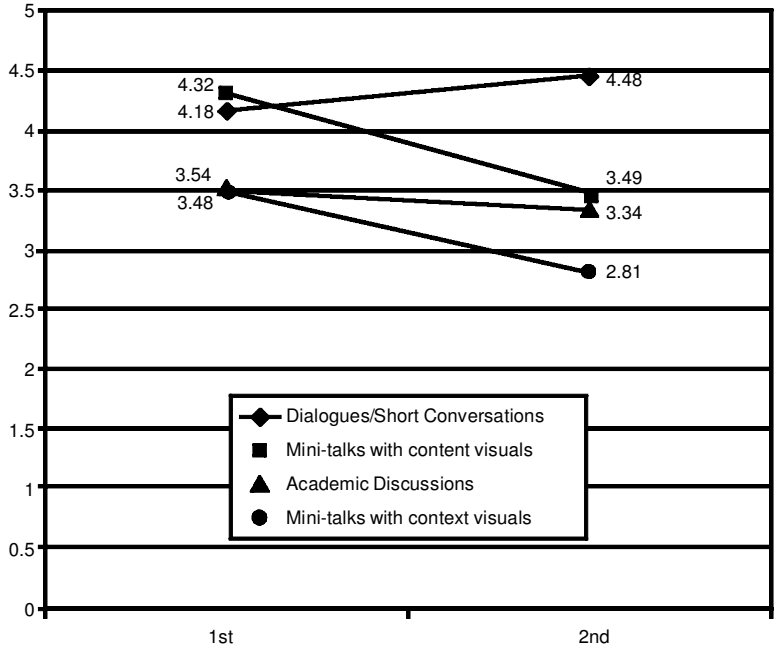

Figure 3 Stimulus type by time

first set (1st presentation, $\bar{x} = 3.54$, sd = 1.30; 2nd presentation, $\bar{x} = 3.34$, sd = 1.39). When Mini-talks with context visuals were presented to participants, they performed better on the first set (1st presentation, $\bar{x} = 3.48$, sd = 1.29; 2nd presentation, $\bar{x} = 2.81$, sd = 1.59).

V Discussion

While the inclusion of effect size is helpful in interpreting significance statistics, their interpretation remains somewhat arbitrary. In order for an effect to meet the criterion for serious consideration, the effect should be at least small; that is, the values associated with partial eta square should be at least .01 (see Cohen, 1988). Accepting this criterion means that the effects associated with the interactions stimulus type by proficiency (Figure 1) and stimulus type by visual condition (Figure 2) are small while the effect associated with stimulus type by time (Figure 3) is relatively large (see Table 5). However, it does not follow that small effects should be considered unimportant. In discussions of the interpretation of small effects both Abelson (1985) and Prentice and Miller (1992) argue that small effects may have cumulative effects when considered outside of experimental contexts,

for example, in actual testing situations. Given that TOEFL examinees are presented with more than a single mini-talk – indeed, they are presented with at least three mini-talks and an associated 12–18 items – the preference for context visuals in this condition requires careful consideration.

1 Stimulus type by proficiency

The significance of the two-way interaction between stimulus type and proficiency shows that while the order of difficulty associated with the stimulus types remains the same for both groups, the gap in performance between Academic Discussions and Mini-talks with context visuals is less pronounced for the high proficiency group.

Other observations are worth noting. The consistency of the high proficiency participants' ability to perform at higher levels than their low-proficiency counterparts does provide evidence that the selection procedure, although indirect, was adequate in distinguishing high and low proficiency groups. It should be noted that the high proficiency group performed close to the top of the scale across all of the stimuli sets, suggesting that the sets and items were relatively easy for these participants, all of whom were enrolled in regular academic courses and had already scored at least 550 on the pencil-and-paper TOEFL.

Although the texts were intended to be comparable with respect to content and difficulty, the mini-talks Aquifers and Amber (MTX1 and MTX2) functioned differently from mini-talks Prairie Dogs and Anthropology (MTN1 and MTN2). Prairie Dogs (a discussion of animal communication) and Anthropology (a discussion of early Polynesian trade routes) were more difficult than Aquifers (an explication of geological strata and water tables) and Amber (a discussion of the characteristics of amber and amber inclusions).

2 Stimulus type by visual condition

The pattern of results associated with the significant two-way interaction between stimulus type and visual condition confirms the predictions in the literature about the facilitative effects of visuals when those visuals can be argued to perform the function of supplanting participants' visual imagery. The stimulus conditions which have the greatest potential for complementary mapping of visual and audio portions of the stimulus materials are those in which the means of the participants were higher. This effect is most obvious in association with the Mini-talks. When content visuals accompanied Mini-talks, the effect was slightly facilitative; when context visuals accompanied Mini-talks, the effect was slightly debilitating.

The effect associated with the presence or absence of visuals with Mini-talks provides evidence for the idea that audio portions of stimuli interact differently with different types of visual information. Mini-talks with content visuals appeared to have the greatest potential for complementary mapping of information between the verbal and visual portions of the stimuli. This mapping was expected to result in visual supplantation and stronger performance on these items, and it did. While the presence of visuals bearing content information might be thought to enhance the complexity of the stimulus, all of the participants, regardless of level of proficiency, were apparently able to use the information presented in the content visual to help create a mental representation of the text. The content visuals that accompanied Mini-talks may have enhanced participants' representations of the auditorily-presented information by making those representations more concrete and/or memorable. It seems likely that, given a still photo of a mosquito in a piece of amber, it might be easier to answer a question concerning the types of animals typically found in amber inclusions than if the participant had only heard a verbal description.

When Mini-talks were presented with visuals bearing only context information, the presence of the visuals had a slightly debilitating effect on performance. It may be the case that when participants are presented with a short lecture that involves academic content, they are distracted by the presence of visuals that have little or no bearing on the content of the talk. This is the only condition in which participants actually performed better when viewing a blank screen. The information in the context visuals that accompanied Mini-talks, like that contained in the visuals accompanying Dialogues/Short Conversations and Academic Discussions, bore generic information about the context in which the communicative event occurred, i.e., information that was, at best, only tangentially related to the content of the text. For example, when the participant is presented with an audio text discussing Polynesian trade routes, a map depicting those routes would be expected to make the direction or location of the routes easier to remember. If presented with a visual depicting the lecturer, the relationship between the content information in both sources is indirect or nonexistent. Given that the visual information provided by context visuals in association with Mini-talks still requires processing but provides no directly-related complementary information, the processing load may be increased and the ability to recall a particular piece of information may be decreased. Given the interaction effect across each of the stimulus types, this scenario appears likely.

It might have been the case that the presence of a visual bearing

contextual information would activate certain scripts. While understanding that the verbal information is presented in an academic setting is unlikely to allow the examinee to predict anything about the content of the message, such understanding might allow the examinee to form expectations concerning the structure of the information. If such expectations were formed on the part of the participants, they did not produce a positive effect. Furthermore, it is extremely important to remember that examinees are tested on the content of the stimuli and not on the context or, directly, on the structure of their presentation.

As in the case of the context visuals that accompanied Mini-talks, the function of the visuals accompanying Dialogues/Short Conversations has been identified by TOEFL test developers as 'setting the scene'. However, in most cases, the content of the dialogues is only marginally related to the dialogue's location. The dialogues usually occur in a school-related setting like a dining hall, a corridor or a classroom, and typically involve topics such as selecting classes, studying for exams or going to dinner at a friend's house. While it should be noted that the topics of conversation can be broadly described as 'academic' or as related to an academic milieu, there are no contextual restrictions that prohibit speakers from engaging in a discussion about any topic in any of the contexts depicted. That is, viewing a scene in a hallway does not allow an examinee to predict anything about the content of the conversation. Given the very generic nature of both the visuals and the verbal information presented in TOEFL Dialogues/Short Conversations, it is hardly surprising that the effect of visuals in this condition was virtually nonexistent.

Like the visuals associated with Dialogues/Short Conversations, the visuals that accompanied Academic Discussions did not present contextual information that was directly related to the content of the verbal portion of the message; however, their presence did serve to facilitate comprehension. One reason that participants might have an advantage when visuals are present in association with Academic Discussions is related to the format of the stimuli. Because Academic Discussions involve a series of turns among several speakers who are often of the same sex, and because they are longer than the Dialogues/Short Conversations, changing still photographs from speaker to speaker in association with changes in turns may have made the discussions easier to parse. Given the similarity of the visuals that accompany Academic Discussions with the visuals that accompany Discussions/Short Conversations, it is likely that the visuals with Academic Discussions are slightly facilitative because they mark turns in the conversations rather than because they 'set the scene'.

It might be possible to create context visuals that produce facilitation, but the context would have to bear a more direct relationship with the content of the audio portion of the stimuli than is present in the current generic settings that are depicted. That is, the information provided by both sources would have to be closely matched or mismatched, e.g., talking about the Australian Crawl while standing in a swimming pool as opposed to talking about jumping hurdles when standing in a swimming pool. However, TOEFL items are not presently designed to take advantage of potential facilitation associated with tightly bound contextual information, and it is hardly clear that they should be. Although context, in some cases, may lead to the activation of particular scripts or aid in the interpretation of a particular word or phrase, the actual topic of conversation in normal conversation usually supersedes context. That is, understanding that a particular communicative event is taking place in a classroom does little that would allow an examinee to predict the content of a lecture; all lectures, covering myriad topics, take place in classrooms. If the content of conversation were bound specifically to context, our topics would be extremely limited. Indeed, a hallmark of human language as opposed to animal 'communication' is the extent to which human language is not determined or limited by context.

3 Stimulus type by time

Given that the experimental context was not a 'high stakes' testing situation, the drop in performance in association with the longer sets of stimuli may have occurred because of boredom or fatigue. Participants may not have been sufficiently motivated to overcome these effects in the experimental context. If participants had been told that they would receive only partial payment if they were unable to perform at a predetermined level, these effects might have disappeared. An alternative explanation is that the effect is the product of memory limitations and attentional deficits that would occur in association with longer stimuli sets in any case.

Mayer (1997) argues that one variable that has been consistently overlooked in the literature concerning the effects of visuals in instructional contexts is memory load. As TOEFL Listening Comprehension materials are consistently being lengthened in an attempt to reflect more accurately the kinds of materials to which students are exposed in academic settings, the significance of this interaction suggests that the effect of the length of Listening Comprehension stimuli deserves special attention.

4 Proficiency by visual condition

The interaction between proficiency and visual condition was not significant. While not much can be made of the absence of a significant interaction (that is, the absence of a significant interaction should not be considered evidence that a relationship does not exist), some speculative remarks are offered in the following section.

5 Proficiency by stimulus type by visual condition

The single three-way interaction that was tested – proficiency by stimulus type by visual condition – was not significant. The two-way interaction between proficiency and visual condition and the three-way interaction between proficiency, stimulus type and visual condition are the interactions that directly relate to the third research question addressed by the study (see Section I.3 above). Interestingly, the significance of the interaction between stimulus type and visual condition (which collapses proficiency levels) suggests that proficiency did not affect the potential of facilitation to occur. In addition, the absence of an interaction associated with proficiency by visual condition also supports this conclusion. That is, the performance of both groups was facilitated when the presence of visuals complemented the audio portions of the stimuli (effects associated with Academic Discussions and Mini-talks with content visuals). As Mayer and Galinni (1990) suggest, the potential for facilitation to occur may be the function of prior knowledge rather than of language proficiency. It is important to note that, in spite of different levels of proficiency, the majority of the participants in this study were highly educated and may be assumed to have requisite levels of background knowledge with respect to the topics presented in TOEFL stimuli. It may be the case that the only way to obtain differential facilitation with respect to proficiency would be to increase the complexity of the information presented in both the audio and visual portions of the stimuli. In that case, perhaps only highly proficient participants would be able to comprehend enough of the audio portions of the texts to make sense and use of the visuals.

The facilitative effects of visuals bearing context information may also be the result of presentation. Because visual and text information were complementary, presentation of visual information was contiguous, and audio texts rather than written texts were used when content visuals were involved, the test designers may have created the most felicitous conditions with respect to the potential for facilitation, and both groups benefited.

However, another effect needs to be considered: the debilitating effect

of context visuals in association with Mini-talks. Given that the gap between the high proficiency participants' performance on Academic Discussions and Mini-talks with context visuals is smaller than that of the low proficiency participants (see Figure 1), it may be the case that high proficiency participants were better able to overcome the presence of context visuals in association with the mini-talks. The graph of the interaction suggests this possibility and the absence of the three-way effect cannot rule it out. The design may simply not have been powerful enough to detect a three-way interaction.

VI Directions for future study

If we ever hope to fully explicate the effects of different types of visuals on participants' performance, it would be extremely useful to develop a model for the difficulty of TOEFL Listening Comprehension items independent of the effects of visuals. If the features affecting difficulty were more clearly understood, the characteristics of stimuli could be more carefully controlled. This would allow the interactions with visual sources of information to be examined in a more systematic fashion.

It is difficult to imagine an accurate model of Listening Comprehension item difficulty that did not take into account the memory requirements associated with different topics, discourse characteristics, item characteristics and the memory limitations associated with varying lengths of audio texts. Given the different ranks of difficulty in association with the different types of stimuli used in this article, clearly much remains unexplained and unexplored. If a model of difficulty for Listening Comprehension stimuli and items were available, it might be possible to devise more precise quasi-experimental manipulations of visual information. In the absence of such models, the actual function of visual accompaniments to audio stimuli will remain speculative.

Nevertheless, there seem to be reasonable explanations for the slight facilitation in performance associated with the content visuals that accompanied Mini-talks and with context visuals that accompanied academic discussions. Facilitative effects occur when the presentation of visuals is contiguous and, most importantly, when it is directly related to the content of the information presented in the audio portion of the stimuli or it marks a turn in the conversation. Thus, theoretical discussions with respect to contiguity are supported and slightly extended to include a facilitative presentation effect of visuals used to mark turn-taking – a kind of conversational contiguity.

It may be disappointing to some that the presence of context visuals as accompaniments to Dialogues/Short Conversations or as

accompaniments to Mini-talks produced either no effect or a slightly debilitating effect, given the value placed on context in discussions of communicative competence. Furthermore, if their presence improves the face validity of the test for both examinees and language specialists, it does not appear unreasonable to include context visuals if their presence does no harm. Given the slight debilitation of performance when context visuals were presented in association with Mini-talks, the preference for context visuals as opposed to content visuals in this condition is problematic, especially if the effect is cumulative. Furthermore, it may not be reasonable to emphasize the importance of representations of context if those representations cannot be associated with an improvement of the measurement of the underlying construct that goes beyond increased face validity.

If the presence of visuals is intended to 'bias for best', the characteristics of content visuals require continued investigation. Their potential to ameliorate the effects of increased difficulty that may occur with the use of longer audio stimuli is an interesting possibility. While content visuals appeared only with Mini-talks, they could easily be provided with Academic Discussions as well. An examination of the effects of context related visuals marking turns in the conversation as opposed to content related visuals presenting complementary visual information might lead to better understandings not only of how content visuals function but also of when and why contextual information can be used to facilitate comprehension in testing situations moving towards the use of longer, more 'authentic' stimuli. Academic Discussions and Mini-talks with content visuals lend themselves well to these types of continued investigations.

Acknowledgements

This study could not have been completed without the support and assistance of many individuals both at Educational Testing Service and at Purdue University. The TOEFL Research Committee's approval of the proposal and financial commitment to the project made the study possible.

Some deserve special mention here because of their special contributions. In particular, Louis Mang and Mike Ecker at Educational Testing Service authored and programmed the experimental stimuli. Susan Nissan, Joyce Blanchette and Mark Tolo, of Test Development at Educational Testing Service wrote and reviewed the experimental stimuli. Kentaro Yamamoto at Educational Testing Service provided the experimental design. Kathleen Sheehan of Educational Testing Service reviewed earlier versions of the report. My thanks as well to

Phil Oltman, Ken Sheppard and Larry Stricker for their helpful comments and assistance throughout the process.

At Purdue University, Heather Allen, Genick Blaise, Jennifer Gerrity, Krishna Prasad and Alan Redmon – graduate students in the English Language and Linguistics Program – ran participants. My thanks, of course, to the participants who volunteered to participate in the study. Diana Woollen, of the English as a Second Language Program, provided administrative assistance. James O'Malley, of the Statistical Consulting Service, provided invaluable consulting expertise and conducted the statistical analyses. My colleague at Purdue, Mary Niepokuj, provided helpful comments and discussion on the final version of the article. I would also like to thank anonymous reviewers for *Language Testing* for their excellent suggestions and remarks. Any remaining flaws are mine alone.

VII References

Abelson, R.P. 1985: A variance explanation paradox: when a little is a lot. *Psychological Bulletin* 97, 129–33.

Bachman, L.F. 1991: What does language testing have to offer? *TESOL Quarterly* 25, 671–704.

Baddeley, A. 1992: Working memory. *Science* 255, 556–59.

Bauer, M.I. and **Johnson-Laird, P.N.** 1993: How diagrams can improve reasoning. *Psychological Science* 4, 372–78.

Chandler, P. and **Sweller, J.** 1991: Cognitive load theory and the format of instruction. *Cognition and Instruction* 8, 293–332.

Cochran, W.G. and **Cox, G.M.** 1957: *Experimental designs.* New York: John Wiley.

Cohen, J. 1988: *Statistical power analysis for the behavioral sciences. 2nd edition.* Hillsdale, NJ: Lawrence Erlbaum.

Cronbach, L.J. 1975: Beyond the two disciplines of scientific psychology. *American Psychologist* 30, 116–26.

Douglas, D. 1997: *Testing speaking ability in academic contexts: theoretical considerations.* TOEFL Monograph Series, RM-97–1. Princeton, NJ: Educational Testing Service.

Duran, R.P. 1989: Testing of linguistic minorities. In Linn, R., editor, *Educational measurement. 3rd edition.* New York: American Council on Education; Macmillan Publishing, 573–87.

Dwyer, F.M. and **de Melo, H.** 1984: Effects of mode of instruction, testing, order of testing, and cued recall on student achievement. *Journal of Experimental Education* 52, 86–94.

Hegarty, M. and **Just, M.A.** 1993: Constructing mental models of machines from text and diagrams. *Journal of Memory and Language* 32, 717–42.

Hymes, D. 1974: *Foundations in sociolinguistics: an ethnographic approach*. Philadelphia, PA: University of Pennsylvania Press.

Mandl, H. and **Levin, J.R.**, editors, 1989: *Knowledge acquisition from text and pictures*. Amsterdam: North-Holland.

Mastropieri, M.A., Scruggs, T.E. and **Levin, J.R.** 1987: Learning disables students' memory for expository prose: mnemonic versus nonmnemonic pictures. *American Journal of Educational Research* 24, 505–19.

Mayer, R.E. 1984: Aids to prose comprehension. *Educational Psychologist* 19, 30–42.

—— 1989: Models for understanding. *Review of Educational Research* 59, 43–64.

—— 1993a: Comprehension of graphics in text: an overview. *Learning and Instruction* 3, 239–46.

—— 1993b: Illustrations that instruct. In Glaser, R., editor, *Advances in instructional psychology, Vol. 5*. Hillsdale, NJ: Lawrence Erlbaum, 253–84.

—— 1997: Multimedia learning: are we asking the right questions? *Educational Psychologist* 32, 1–19.

Mayer, R.E. and **Anderson, R.B.** 1991: Animations need narrations: an experimental test of a dual-coding hypothesis. *Journal of Educational Psychology* 84, 444–52.

Mayer, R.E. and **Anderson, R.B.** 1992: The instructive animation: helping students build connections between words and pictures in multimedia learning. *Journal of Educational Psychology* 84, 444–52.

Mayer, R.E., Bove. W., Bryman, A., Mars, R. and **Tapangco, L.** 1996: When less is more: meaningful learning from visual and verbal summaries of science textbook lessons. *Journal of Educational Psychology* 88, 64–73.

Mayer, R.E. and **Gallini, J.K.** 1990: When is an illustration worth ten thousand words? *Journal of Educational Psychology* 82, 715–26.

Mayer, R.E. and **Sims, V.K.** 1994: For whom is a picture worth a thousand words? Extensions of a dual-coding theory of multimedia learning. *Journal of Educational Psychology* 86, 389–401.

Mayer, R.E., Steinhoff, K., Bower, G. and **Mars, R.** 1995: A generative theory of textbook design: using annotated illustrations to foster meaningful learning of science text. *Educational Technology Research and Development* 43, 31–44.

Moore, D.M. and **Dwyer, F.M.** 1994: Effect of cognitive style on test type (visual or verbal) and color coding. *Perceptual and Motor Skills* 79, 1532–34.

Parkhurst, P.E. and **Dwyer, F.M.** 1983: An experimental assessment of students' IQ level and their ability to profit from visualized instruction. *Journal of Instructional Psychology* 10, 9–20.

Prentice, D.A. and **Miller, D.T.** 1992: When small effects are impressive. *Psychological Bulletin* 112, 160–64.

Salomon, G. 1979: *Interaction of media cognition and learning*. San Francisco, CA: Jossey-Bass.

Salomon, G. 1989: Learning from texts and pictures: reflections on a meta-level. In Mandl, H. and Levin, J.R., editors, *Knowledge acquisition from text and pictures*. Amsterdam: North-Holland, 73–82.

Schegloff, E.A. 1997: Whose text? Whose context? *Discourse and Society* 8, 165–87.

Sweller, J., Chandler, P., Tierney, P. and **Cooper, M.** 1990: Cognitive load as a factor in the structure of technical material. *Journal of Experimental Psychology: General* 119, 176–92.

Weidenmann, B. 1989: When good pictures fail: an information processing approach to the effect of illustrations. In Mandl, H. and Levin, J.R., editors, *Knowledge acquisition from text and pictures*. Amsterdam: Elsevier Science.

Willows, D.M. and **Houghton, H.A.**, editors, 1987: *The Psychology of illustration: Volume 2. Instructional Issues.* New York: Springer-Verlag.

Winn, W. 1991: Learning from maps and diagrams. *Educational Psychology Review* 3, 211–47.

Appendix 1 List of test forms

Section	Form 1	Form 2	Form 3	Form 4
1	D/SC1V	D/SC2NV	D/SC1NV	D/SC2V
	D/SC2NV	D/SC1V	D/SC2V	D/SC1NV
2	AD1V	AD2NV	AD1NV	AD2V
	AD2NV	AD1V	AD2V	AD1NV
3	MTX1V	MTX2NV	MTX1NV	MTX2V
	MTX2NV	MTX1V	MTX2V	MTX1NV
4	MTN1V	MTN2NV	MTN1NV	MTN2V
	MTN2NV	MTN1V	MTN2V	MTN1NV

Section	Form 5	Form 6	Form 7	Form 8
2	AD1V	AD2NV	AD1NV	AD2V
	AD2NV	AD1V	AD2V	AD1NV
3	MTX1V	MTX2NV	MTX1NV	MTX2V
	MTX2NV	MTX1V	MTX2V	MTX1NV
4	MTN1V	MTN2NV	MTN1NV	MTN2V
	MTN2NV	MTN1V	MTN2V	MTN1NV
1	D/SC1V	D/SC2NV	D/SC1NV	D/SC2V
	D/SC2NV	D/SC1V	D/SC2V	D/SC1NV

Section	Form 9	Form 10	Form 11	Form 12
3	MTX1V	MTX2NV	MTX1NV	MTX2V
	MTX2NV	MTX1V	MTX2V	MTX1NV
4	MTN1V	MTN2NV	MTN1NV	MTN2V
	MTN2NV	MTN1V	MTN2V	MTN1NV
1	D/SC1V	D/SC2NV	D/SC1NV	D/SC2V
	D/SC2NV	D/SC1V	D/SC2V	D/SC1NV
2	AD1V	AD2NV	AD1NV	AD2V
	AD2NV	AD1V	AD2V	AD1NV

Section	Form 13	Form 14	Form 15	Form 16
4	MTN1V	MTN2NV	MTN1NV	MTN2V
	MTN2NV	MTN1V	MTN2V	MTN1NV
1	D/SC1V	D/SC2NV	D/SC1NV	D/SC2V
	D/SC2NV	D/SC1V	D/SC2V	D/SC1NV
2	AD1V	AD2NV	AD1NV	AD2V
	AD2NV	AD1V	AD2V	AD1NV
3	MTX1V	MTX2NV	MTX1NV	MTX2V
	MTX2NV	MTX1V	MTX2V	MTX1NV

Notes: V = visual present; NV = no visual present.