The Effects of Message Length, L2 Proficiency and Cognitive Workload on Performance Accuracy and Speech Production in a Simulated Pilot Navigation Task

Candace Farris

A Thesis

In

The Department

Of

Education

Presented in partial fulfillment of the requirements for the degree of
Master of Arts (Applied Linguistics) at
Concordia University
Montreal, Quebec, Canada

August 2007

Your file Votre référence
ISBN: 978-0-494-34591-7
Our file Notre référence
ISBN: 978-0-494-34591-7

ABSTRACT

The Effects of Message Length, L2 Proficiency and Cognitive Workload on Performance Accuracy and Speech Production in a Simulated Pilot Navigation Task

Candace Farris

The goal of the present study was to investigate the influence of L2 proficiency in controller-pilot communications under conditions of varying cognitive workload. The study therefore examined the effects of L2 proficiency and cognitive workload on performance accuracy and speech production in a simulated pilot navigation task. Three groups of 20 participants (one native-English speaking and two native-Mandarin speaking of relatively high and low levels of proficiency in English) took part in the experiment. Participants listened to, repeated and responded to "controller" messages (in English) of varying lengths in two conditions: clear and workload. In the workload condition, participants performed a concurrent arithmetic task while repeating the controller message.

The dependent variables were divided into two sets: performance accuracy (message repetition and navigation accuracy), and speech production (accentedness, comprehensibility, fluency and confidence, as perceived by 10 native-English speaking raters). Based on the results of the present study, it is recommended that air traffic controllers limit their messages to a length of 2 commands when communicating with L1 pilots under high workload conditions. When communicating with L2 pilots of low or intermediate L2 proficiency, it is recommended that controllers limit the length of their messages to 1 command under high cognitive workload conditions. A significant detriment to speech production due to increased message length was observed for the L2 groups for all speech production measures. A significant detriment to speech production due to increased cognitive workload was observed for the Low L2-proficiency group for the perceived fluency measure.

DEDICATION

This thesis is dedicated to Chuck, Max, Nilah and Diane, whose support and understanding have made this work possible.

ACKNOWLEDGMENTS

I would like to acknowledge the valuable contributions of my committee members – Dr. Pavel Trofimovich, Dr. Elizabeth Gatbonton and Dr. Norman Segalowitz. Their guidance and wisdom have been greatly appreciated.

First and foremost, I extend my heartfelt gratitude to my supervisor, Dr. Pavel Trofimovich, for his inspiration and unwavering support. Dr. Trofimovich has been a wonderful teacher, role-model and friend, and I thank him for his significant contributions to this work and for having faith in me. I would also like to thank Dr. Elizabeth Gatbonton and Dr. Norman Segalowitz for their generosity and expert advice. I feel fortunate to have had the opportunity to learn from them. I would like to extend a special thank you to Randall Halter for his assistance in the lab, statistical advice and overall kindness. These wonderful people have made Concordia a second home.

I would like to extend my sincere gratitude to Dr. Immanuel Barshi of NASA-Ames Research Center for his friendship, wisdom and helpful comments. I also thank Dr. Barshi for kindly lending the software for the navigation task. I would also like to thank Dr. Alice Healy and Vivian Schneider of the University of Colorado at Boulder for their assistance and useful comments.

I would like to thank Andrea Kneeland and Tarek Nocharli for their friendship and insightful comments, and Graham Elliott of the Federal Aviation Administration for his kindness and help along the way.

I wish to acknowledge the 82 individuals who participated in this research project. Without them this work would not have been possible. Finally, I would like to thank all

of my teachers, friends and fellow students in the Master of Applied Linguistics program for a wonderful learning experience.

TABLE OF CONTENTS

LIST OF FIGURES

Figure Page

LIST OF TABLES

Table Page

CHAPTER 1

INTRODUCTION

It has been determined that air traffic controller (hereafter "controller")-pilot miscommunications are a threat to air safety (Billings & Cheaney, 1981; Billings & Reynard, 1984). Evidence for this has come from a number of sources, including incident/accident reports, tapes and transcripts of controller-pilot communications and simulator studies (see Prinzo & Britton, 1993, for a comprehensive review). Most research in this area has, however, been descriptive in nature, and there is a paucity of experimental research investigating possible reasons for controller-pilot miscommunications, including those related to language. The goal of the present study is therefore to systematically examine in an experimental setting some of these reasons—for example, factors that pertain to the nature of controller-pilot communications (message length and language proficiency), and those that relate to controller/pilot working conditions (cognitive workload and concurrent task management).

The following example of a major accident in which 583 passengers lost their lives illustrates the importance of effective air-ground communications and suggests that language proficiency, among other factors affecting controller/pilot performance, is of paramount importance in determining air safety. The 1977 accident which involved the collision of two Boeing 747's in Tenerife, Canary Islands, was caused in part by the pilot's misunderstanding of the controller's instructions. Whereas the pilot believed the aircraft was cleared for take off, the aircraft was actually awaiting take-off clearance and in its take-off attempt collided with another aircraft taxiing down the runway.

Communications between the native Dutch-speaking pilot and the native Spanish-speaking controller took place in English, and cross-linguistic differences between the native Dutch-speaking pilot's mother tongue (L1) and his second language (L2) may have contributed to the misunderstanding. A number of factors—such as increased workload for both the pilot (caused by the flight's diversion to Tenerife due to a bomb scare) and the controller (due to increased air traffic), chain of command protocol in the cockpit environment (preventing a subordinate from challenging the captain's decision to take off), and congested radio frequencies blocking an important transmission—may have interacted and contributed to the miscommunication. Nevertheless, this tragic accident has been instrumental in raising awareness in the aviation community as to the importance of proficiency in the commonly-used language of international civil aviation, namely English.

ICAO Language Proficiency Requirements

As the example provided above illustrates, L2 proficiency is an important factor in effective pilot-controller communications in aviation situations where the crew and ground do not share the same native language. In response to data from three accident report databases, including the International Civil Aviation Organization (ICAO) Accident/Incident Data Reporting System (ADREP), the U.S. National Transportation and Safety Board (NTSB) reports, and the United Kingdom's mandatory Occurrence Reporting Systems, ICAO has concluded that the role of language in aviation accidents is significant (ICAO Doc 9835 AN/453, 1-1). As a result, ICAO has introduced language proficiency requirements to ensure that all air traffic control personnel and flight crews are proficient in the language(s) used in controller-pilot communications.

The ICAO Language Proficiency Requirements are to be applied to all languages used in radiotelephony. These requirements stipulate that English be made available in situations where the crew and the ground do not share the same native language (e.g., an Arabic-speaking Moroccan pilot communicating with a Spanish-speaking controller in Mexico). However, in the event the pilot is proficient in the native ground language (e.g., a Spanish-speaking Moroccan pilot communicating with a Spanish-speaking controller in Mexico), it is the pilot's choice whether to use English or the native ground language (i.e., Spanish) in communications with the controller. Therefore, according to the ICAO Language Proficiency Requirements, all crews and controllers involved in flight operations where the crew and ground do not share the same native language and the use of English may be necessary are required to be proficient in conducting and comprehending radiotelephonic communications in the English language (ICAO Doc 9835 AN/453, iiv).

All ICAO member states[1] must comply with the new standards, as outlined in the manual, by March 5th, 2008 (ICAO Doc 9835 AN/453, 5-1). In order to satisfy the requirements, pilots and controllers are required to demonstrate Operational Level 4 language proficiency in the use of both ICAO phraseology and plain language (ICAO Doc 9835 AN/453, ix) in the language(s) of controller-pilot communications. The standards also stipulate that pilots and controllers who demonstrate language proficiency below Level 6 undergo recurrent testing. While ICAO's resolution to enforce a world-wide language-proficiency requirement clearly acknowledges the important role of L2 proficiency in effective air-ground communications, several important questions related

[1] There are currently 190 ICAO member states (http://www.icao.int/cgi/goto_m.pl?/cgi/statesDB4.pl?en, July 15, 2007)

to language proficiency remain unanswered. For example, it is not clear how and to what degree L2 proficiency corresponds to the communicative needs of pilots and controllers as they communicate under varying workload conditions or interact using messages of different length and complexity. The present study thus attempts to address at least some of these questions by investigating the effects of (1) language-related factors (pilot L2 proficiency, controller message length) and (2) factors related to controller/pilot workload (low vs. high) on native and non-native English-speaking pilots' task performance and speech production.

Discussed in the following chapter are theoretical and practical motivations for the study of controller-pilot communication, particularly in the international aviation context (i.e., situations where the use of L2s is required). First, a general description of the controller-pilot communicative environment is presented (see *Controller-Pilot Communicative Environment*). Second, language-related issues in controller-pilot radiotelephonic communications are discussed (see *Linguistic Factors in Controller-Pilot Communications*). Third, the role of workload in controller-pilot communication is described (see *Workload as a Factor in Controller-Pilot Communications*). Finally, the hypotheses of the present study are presented.

CHAPTER 2

LITERATURE REVIEW

Chapter Overview

This chapter provides a review of the literature relevant to the present study. The chapter is divided into three sections. In the first section, a general description of the controller-pilot communicative environment is presented. In the second section, linguistic factors in controller-pilot communications are identified. Within this section, the literature pertaining to message length is reviewed first. Next, the literature pertaining to L2 proficiency is reviewed. In the third section, workload as a factor in controller-pilot communications is discussed. Within this section, the literature pertaining to workload and task performance is reviewed first. Next, the literature pertaining to workload and speech production is reviewed. Finally, a chapter summary is provided, followed by the hypotheses of the present study.

Controller-Pilot Communicative Environment

Despite its decidedly important role in international air safety, proficiency in an L2 has been given little consideration as a factor in the descriptive studies of controller-pilot communications. It should be noted, however, that most of the studies in controller-pilot voice communications are based on data taken from the U.S. Air Traffic Management System (ATMS), and miscommunications due to L2 proficiency or adherence to standard phraseology may not be as salient for the U.S. national airspace system as for other geographical areas, such as Europe and Asia, where cross-linguistic communication is a more common and constant challenge.[2] Despite this possible sampling bias in reporting language-related miscommunications, the data from the American ATMS have been a valuable source of information for controller-pilot communication studies, particularly in identifying factors responsible for communication breakdowns leading to incidents and accidents.

Analyzing data from incident/accident reports, Billings and Cheaney (1981), for example, reported that information transfer problems were present in over 4,800 incident reports submitted to the Aviation Safety Reporting System (ASRS)[3] per year between 1978 and 1980. The study went on to identify factors commonly associated with information-transfer problems, including human behavior problems, such as (in order of frequency), distraction, forgetting, failure to monitor, non-standard or ad-hoc procedures or phraseology, and complacency. Systems factors (i.e., factors other than those related to human behavior), such as (in order of frequency), non-availability of traffic information, degraded information, ambiguous or (rarely) absent procedural guidance, environmental factors (noise, confusion), high workload, and equipment failure were also cited as contributors to information-transfer problems.

Although the data and analyses gleaned from these incident/accident reports are very valuable, there are certain problems associated with this method of data collection.

[2] This is not to say, however, that proficiency in the English language and adherence to standard phraseology is not a concern to the world's largest national aviation system, as international flights in and out of the U.S. are operated by pilots and crews from all parts of the world with various linguistic backgrounds and levels of proficiency in English. Even with regard to domestic air travel, the various regional dialects of English within the U.S. make adherence to the ICAO language proficiency requirements imperative.

[3] The ASRS was established in 1975 as a joint effort by the Federal Aviation Administration (FAA) and the National Aeronautics and Space Administration (NASA). The program is administered by NASA and funded by the FAA. Policies are set by NASA in consultation with the FAA and the aviation community. The purpose is to provide a non-punitive forum for the reporting of aviation incidents for the purpose of improving safety of the ATMS (for further information, see http://asrs.arc.nasa.gov/main_nf.htm).

First, the incident/accident reports reflect the perspective of the reporter, suggesting a certain bias in the interpretation of facts and events. Second, the transcripts provide limited information in that they are taken out of context. Third, the transcripts are not created by linguists for the purpose of studies in controller-pilot communications, and are therefore perhaps not sufficiently accurate and detailed for a full analysis of language-related miscommunication issues. Despite these limitations, however, these reports have been useful in highlighting the existence of information-transfer problems and in creating awareness as to their importance.

Other studies have examined controller-pilot communications through the analysis of tapes and transcripts of controller-pilot voice communications (Cardosi, 1993, 1996; Morrow, Lee & Rodvold, 1993; Morrow, Rodvold & Lee, 1994).These studies have looked at both routine (Morrow et al., 1993) and non-routine communications (Morrow et al., 1994) in different air traffic environments, such as TRACON[4] (e.g., Cardosi, 1996; Morrow et al., 1993; Morrow et al., 1994) and En Route[5] (e.g., Cardosi, 1993). The results of these studies have emphasized the importance of the collaborative, interactive nature of controller-pilot communications, defined by Morrow et al. (1993) as a collaborative scheme involving three phases:

a) **Initiate:** the pilot initiates communication by getting the controller's attention.

b) **Present:** the controller gives the pilot new information and/or instructions.

c) **Accept:** the pilot acknowledges and/or reads back the information to confirm

[4] The TRACON (Terminal Radar Approach Control) facility is usually located near an airport and controllers are able to monitor approaching aircraft on a radar screen and provide services once radio communication has been established on the appropriate frequency.

[5] En Route air traffic control services are provided for aircraft operating on instrument flight rules (IFR) between departure and terminal areas. Workload permitting, services may also be provided to pilots operating on visual flight rules (VFR).

mutual understanding and the controller "hears back" the pilot's readback. The following is a hypothetical example of a controller-pilot communication reflecting the collaborative scheme outlined above. In this example, the pilot is requesting clearance to transit Class C[6] airspace:

Pilot: *Montreal Terminal, Cessna GXMT.*

Controller: *Cessna GXMT, Montreal Terminal.*

Pilot: *Cessna GXMT, 15 miles south of Trudeau airport, 4,500 feet, enroute to Mirabel Airport. Request Clearance to transit Class Charlie airspace.*

Controller: *Cessna GXMT, squawk 1234 and ident.*[7]

Pilot: *Cessna GXMT, squawking 1234 and ident.*

Controller: *Cessna GXMT, radar contact 15 miles south of Trudeau airport, altimeter 30.03, cleared to enter Class Charlie airspace, maintain 4,500 feet heading 350.*

Pilot: *Cessna GXMT, altimeter 30.03, cleared to enter Class Charlie airspace. Maintain 4,500 feet and heading 350.*

The above example illustrates the importance of the collaborative scheme in ensuring that the controller and pilot are sharing the same situational awareness or mental model—that is, that they attain a mutual understanding of the situation (Clark & Schaefer, 1987). Morrow et al. (1994) cite non-standard collaborative practices, defined as deviations from this collaborative scheme, as a cause of non-routine communications. Non-routine communications can compromise the safety of ATMS by disrespecting the communication system's inherent accuracy and efficiency constraints, and controller-

[6] In order to enter Class C airspace, pilots are required to contact the controller and comply with any heading or altitude changes given.

[7] When asked to "squawk", the pilot is being asked to activate specific modes, codes or functions on the aircraft radar beacon transponder, which transmits and responds to radio interrogators on the ground. In this case the controller is asking the pilot to squawk so that the controller may identify the aircraft on the radar screen.

pilot communication must maintain a balance between accuracy and efficiency in order to ensure safe and expeditious flow of air traffic.

The acceptance stage in the above-mentioned controller-pilot collaborative scheme has an important role in ensuring safe and expeditious flow of air traffic. More specifically, it is during this stage that the controller has the opportunity to detect possible communication breakdowns, indicated by either partial or erroneous pilot readbacks—that is, the pilot's repetition of the information received from the controller (Morrow et al., 1993a). These errors may be detected in the hearback phase (i.e., as the controller listens to the pilot's readback); however, failure to detect errors/omissions in the hearback phase can occur, particularly if the controller is preoccupied with other tasks. Under high controller workload conditions—for example, when a controller is tracking a large number of aircraft at a time—the controller may give the pilot a series of instructions in one long message, and then move on to the next aircraft, skipping the hearback phase in the interest of efficiency. On the other hand, if the pilot is preoccupied with other high-priority tasks, such as flying the plane and/or navigating, the workload of the pilot may be such that it is difficult to both process and act on the information contained in the long message and to read the information back correctly. In this case, the pilot may not read back the controller's instructions for verification and/or provide requested information, and air safety may as a result be compromised (see, e.g., Cardosi, 1993, 1996; Morrow et al., 1993a).

As these and other analyses of controller-pilot interaction indicate (e.g., Barshi, 1997; Cardosi, 1993, 1996; Morrow et al., 1993a), controller workload conditions (e.g., high vs. low), on the one hand, and message properties (e.g., long vs. short), on the other, appear to be important factors that likely determine the success of controller-pilot communications. The objective of the present study is thus to further investigate, in a simplified simulated environment, the effects of controller message length on native and non-native English-speaking pilots' navigation and readback accuracy as well as their speech production (i.e., perceived fluency, accentedness, comprehensibility, and confidence of pilot readbacks) under varying workload conditions. A detailed review of these factors and their effect on controller-pilot communications is presented in the following sections. The discussion of linguistic factors (controller message properties, pilot L2 proficiency) is presented first, followed by the discussion of workload as a factor in controller-pilot communications.

Linguistic Factors in Controller-Pilot Communications

Pilots and controllers have a mutual responsibility for the maintenance of air safety. Controllers' primary task is to ensure the safe and expeditious flow of air traffic within their sector, and the most important aspect of this task is ensuring aircraft separation by monitoring the position of aircraft within their sector. Pilots usually perform three primary tasks, which are outlined in the pilot task prioritization maxim "aviate-navigate-communicate". The hierarchical order of these three tasks is generally logical from a safety perspective; however, it should be noted that pilot task prioritization is dynamic and dependent on demands determined by flight phase and external as well as internal factors (Chou, Madhaven & Funk, 1996; Dismukes, Loukopoulos & Jobe, 2001; Loukopoulos, Dismukes, & Barshi, 2003). Both the pilot and the controller, from different perspectives, are required to maintain overall situational awareness, and to perform other tasks while maintaining the necessary level of communication. For the pilot, the tasks of flying the plane and navigating can, in many situations, be dependent on effective controller-pilot communications, and for the controller, maintaining aircraft separation is definitely dependent on effective communications with the pilot. Therefore, communication is an important aspect of pilot task complexity, and must be regarded as a factor that interacts not only with other pilot tasks, such as flying the plane and navigating, but with controller tasks as well (Barshi & Chute, 2001). It is therefore important to investigate the effect of linguistic factors on a pilot's ability to communicate, considering that English is the commonly-used language of international aviation and is an L2 of many of the world's pilots and controllers.

A number of important linguistic factors that may contribute to increased pilot task complexity have been identified in previous descriptive studies (e.g., Cardosi, 1993, 1996; Morrow et al., 1993a, 1994). These include controller/pilot L2 proficiency, message length, message composition, rate of speech, and degree of accentedness, among others (Cardosi, 1993). It should be noted that the data from these descriptive studies regarding the effect of interlocutors' rate of speech and degree of accentedness on pilot-controller communications are anecdotal, in that Cardosi (1993) claims that pilots and/or controllers often cite these as factors in miscommunications. Four of these factors (discussed in detail below), those regarding pilot L2 proficiency, message length, rate of speech, and prosody, have subsequently been investigated in controlled laboratory studies (Barshi, 1997; Barshi & Healy, 1998).

There has been no experimental research to date examining the effects of speaker accentedness, comprehensibility, fluency, and confidence in controller-pilot communications. One of the goals of the present study was therefore to explore these factors in controller-pilot communications. The relationship between *accentedness* and perceived *comprehensibility* is likely a complex one, subject to a high degree of variability between listeners (Derwing & Munro, 1997; Munro & Derwing, 1995a; 1995b), and possibly influenced by the listener's familiarity with the accent of the speaker as well as fluency-based characteristics of speech (Derwing & Munro, 1997). Munro and Derwing (1997) suggest that listeners may confound accentedness and speech rate in their perception of comprehensibility, in that listeners may perceive speech that is accented and spoken at a high rate as "too fast" because it may require more time to process, resulting in lower comprehensibility ratings (Munro & Derwing, 1995b). Previous research investigating listener perception of L2 speech suggests that the degree of listeners' perception of non-native speakers' accentedness is to a large extent determined by *fluency* characteristics of speech—such as speech rate and frequency and duration of pauses (Trofimovich & Baker, 2006). Finally, listeners' perception of speakers' *confidence* may be crucially dependent on specific aspects of the speech signal, most likely its fluency characteristics (Brennan & Williams, 1995). These speech production variables are discussed below in relation to the three main factors investigated in the present study: message length, L2 proficiency, and workload.

Message length

Morrow and Rodvold (1993), drawing on the Morrow et al. (1993) descriptive study of problems in routine controller-pilot communications, conducted a controlled simulator study in which they investigated the effects of message length and timing on pilot communications. "Message length" in this study was defined as the number of commands per message. The following is a hypothetical example of a controller message containing two commands and the subsequent pilot readback:

Controller: *Cessna GXMT. Turn right next taxiway. Contact ground 121.8.*

Pilot: *Cessna GXMT. Turn right next taxiway. Ground 121.8.*

The study was conducted in the context of a flight simulation involving 16 male native English-speaking aircraft pilots as participants. The results showed a statistically significant increase in errors (defined as transactions that interrupt routine communication such as partial or incorrect readbacks or repetitions/requests for clarification) with messages containing four commands. The results also indicated a statistically significant decrease in errors when the controller's commands were delivered in two messages, each containing two commands, and suggested that timing between the two messages was important in that with too little time, the second message interfered with pilots' memory of the first message.

Further evidence for the detrimental effects of message length (defined in terms of the number of propositions per message) on participants' performance in a simulated navigation task was obtained by Barshi (1997). Barshi (1997) varied the number of instructions per message between one and six and found a significant drop in performance, as measured by the participants' accuracy in navigation task performance between messages containing three and messages containing four propositions. Similar findings were obtained in similar studies when accuracy of readback, defined as the number of accurately repeated critical (i.e., non-redundant) words in a given command, was used as a measure of task performance accuracy (Barshi, 1998; Barshi & Healy, 1998, 2002; Mauro & Barshi, 1999; Schneider, Healy & Barshi, 2004). The above-mentioned results suggest that controllers should limit the number of propositions in a given message to three, a finding of particular interest because Morrow and Rodvold (1993) found that the delivery of two two-command messages as opposed to one four-command message generated an overall greater number of controller-pilot transactions, indicative of lower communication efficiency, despite a reduction in errors. Taken together, these findings suggest that when both the pilot and controller share the same native language, controller delivery of messages of three commands may provide an optimal balance of the accuracy and efficiency constraints inherent in the pilot-controller communication environment.

At least two studies have investigated the effects of message length on L1 speech production (Kleinow & Smith, 2000; Maner, Smith & Grayson, 2000). These studies yielded contrasting results. In an investigation of the effects of message length and complexity, Kleinow and Smith (2000) tested adults who stutter and adults who do not and found that increased message length and complexity resulted in less-stable lower lip movements, a characteristic of degraded speech. The results of this study are inconclusive, however, with respect to the effects of message length on speech production because message length and message complexity were confounded in the study. That is, longer messages were created by inserting an imbedded clausal statement in the control sentence, thus creating not only a longer but also a more complex message. Using similar labial-kinematic measures, Maner et al. (2000) separated the variables of message length and complexity in a study conducted with adults and children. They overall found no significant effect for longer messages, although more complex messages resulted in significantly less stable lip movements for both participant groups in relation to the control condition. These findings suggest that for normally fluent adult L1 speakers, the utterance of longer messages may not result in a detriment to speech production, at least as indicated by spatio-temporal lip movement measures.

L2 proficiency

In the multi-task controller-pilot work environment, the need for the pilot and/or controller to conduct communications in an L2 may place additional cognitive demands on one or both interlocutors. Barshi and Healy (1998) extended Barshi's original study to include non-native speakers of English in an investigation of the effects of fluency in an L2 (in this case, English), message length, speech rate and prosody on navigation task performance. Surprisingly, as in the previous study (Barshi, 1997), no significant effect of speech rate was obtained, even among the low-proficiency participants, such that task performance was not impaired overall for messages presented at slower or faster speech rates. As in the previous study (Barshi, 1997), speech rate in the samples was digitally manipulated in order to maintain intelligibility. The effect of message length also paralleled the results of the previous study with one striking exception: In the low proficiency group, a significant drop in navigation accuracy was observed between messages of two and three commands, as opposed to between messages of three and four commands in the native-speaker and high-proficiency groups (Barshi, 1997).

The findings of Barshi and Healy (1998) suggest that listeners' (i.e., pilots') processing capacity for messages delivered in an L2 is dependent on their level of L2 proficiency and that these additional cognitive demands placed on low-proficiency listeners in turn affect their comprehension and retention of the information presented, as reflected in their performance on the navigation task. Therefore, these findings reveal that messages of more than two propositions place increased processing demands on low-proficiency listeners. These findings, however, leave unanswered two important questions. First, they do not reveal if long messages would have a comparable (detrimental) effect on L1 and L2 "pilot" speech production, as reflected by measures of perceived comprehensibility, accentedness, fluency and confidence. Conceptualized as a partial replication of Barshi's original studies (Barshi, 1997, 1998; Barshi & Healy, 1998), the present study was designed to extend the results of these previous investigations by examining the effect of message length on native and non-native listeners' speech production, using measures of speech perceived comprehensibility, accentedness, fluency and confidence.

Second, previous studies have not investigated the interaction between additional cognitive workload and language proficiency, as it affects "pilots'" comprehension and retention of the information presented (measured by navigation task performance and readback or message repetition accuracy) and their speech production (as investigated using measures of perceived fluency, comprehensibility, accentedness and confidence). It is important to determine how workload affects pilots' navigation task performance and speech production because pilots' processing and repetition of controller messages draws on the same pool of cognitive resources used to perform other concurrent cognitive tasks (e.g., navigating, visual monitoring). Therefore, in the context of controller-pilot communications, the additional cognitive workload, characteristic of controllers' and pilots' multi-task work environment (see below), may interact with L2 proficiency and message length, detrimentally affecting pilot navigation performance and speech production. Thus, the present study was also designed to examine the effect of workload (high vs. low) on native and non-native listeners' navigation accuracy, repetition accuracy and speech production (using measures of accentedness, comprehensibility, fluency, and confidence), particularly as a function of pilot L2 proficiency and message length. In order to better understand how workload may affect L1 and L2 pilots, the role of workload in controller-pilot communications is discussed next.

Workload as a Factor in Controller-Pilot Communications

The cockpit may be described as a multi-task environment in which pilots are required to prioritize and perform concurrent tasks (Chou, Madhaven, & Funk, 1996; Dismukes, Young & Sumwalt, 1998; Dismukes, Loukopoulis & Jobe, 2001; Loukopoulis, Dismukes & Barshi, 2003; Raby & Wickens, 1994). As mentioned above, pilots are trained in the maxim "aviate-navigate-communicate". However, not all pilot tasks can be easily defined according to these categories, and this order of prioritization may not be appropriate in all situations. Therefore, an additional skill required of pilots is concurrent task management, which involves the prioritization and timing of task performance in a multi-task environment in order to achieve optimal performance, the first measure of which is, ideally, safety.

Workload and task performance

The cognitive load theory which suggests that humans are limited in working memory capacity when processing new information, particularly when the new information must be worked on in some respect (Sweller, 1994), has particular significance in the work environments of both controllers and pilots, where new information is given, received, and acted upon in a concurrent multi-task environment. The effect of concurrent tasks on native speakers' task performance in an aviation context is relatively well-documented (e.g., Loukopoulis et al., 2003; Raby & Wickens, 1994). In particular, concurrent tasks can produce a detrimental effect on performance in a variety of tasks, including high-priority tasks. An oft-cited example of the potentially fatal consequences of ineffective concurrent task management is the December 1972 accident in which 99 people died when the crew of the Lockheed L-1011, preoccupied with a landing gear light malfunction, failed to monitor the flight instruments and therefore did not notice that the plane's altitude had dropped significantly, resulting in a fatal crash. This tragedy has been instrumental in raising awareness as to the importance of effective workload management in the cockpit environment.

In three high-fidelity simulated landing approaches using student pilots, Raby and Wickens (1994) analyzed pilot performance in a concurrent multi-task environment, and found that participants were reasonably good at maintaining performance on high-priority tasks under high-workload conditions, but that tasks of lower priority were either shed or degraded in performance as workload increased. As workload increased, pilots' communications with controllers became shorter, and the duration of higher priority tasks became longer. This order of task prioritization corresponds with the "aviate-navigate-communicate" maxim and has relevance to the present study in that shorter communications, related to procedural deviations such as impartial or incomplete readbacks, are likely indicative of high-workload conditions. In another study investigating cockpit task management, Dismukes, Young and Sumwalt (1998) analyzed ASRS reports and found that nearly half the activities that distracted or preoccupied pilots fell under the broad category of communication (e.g., discussion among crew or radio communication).

The above-stated findings (e.g., Raby & Wickens, 1994) provide evidence warranting the investigation of the effects of workload in a controller-pilot communicative environment, and more specifically, on native and non-native participants' comprehension and retention of the information presented (measured by navigation task performance and readback accuracy) and their speech production (using measures of perceived fluency, comprehensibility, accentedness and confidence). It is hypothesized here that, due to increased processing demands, low-proficiency L2 participants will experience a relatively increased detriment in task performance when faced with high-workload conditions involving L2 communications. In addition, some studies (reviewed below) indicate that increased cognitive workload may have a detrimental effect on L1 speakers' speech production; therefore, in the section below, literature examining the effects of workload on L1 speakers' speech production is reviewed.

Workload and speech production

Whereas the effects of high workload on pilot task performance are relatively well-researched (Chou et al., 1996; Dismukes et al., 1998, 2001; Loukopoulis et al., 2003; Raby & Wickens, 1994), few (if any) studies have explicitly investigated the effect of increased workload or concurrent task performance on speech production in the context of pilot-controller communication. However, several have done so in a non-aviation context and the findings of many of these studies (Brenner & Shipp, 1987; Dromey & Benson, 2003; Hecker, Stevens, von Bismarck, & Williams, 1968; Jou & Harris, 1992; Lively, Pisoni, Van Summers & Bernacki, 1993; Mendoza & Carballo, 1998; Ooman & Postma, 2001; Williams & Stevens, 1972) are relevant to the present study.

Previous experimental studies investigating the effects of cognitive workload on L1 speech production have used objective (acoustic, temporal, labio-kinematic) measures of speech production, and have found that high cognitive workload results in measurable changes to speech production in comparison to speech produced in a single-task or low cognitive workload paradigm. Hecker et al. (1968) measured the effects of increased cognitive workload on the acoustic speech signal in an attempt to determine whether or not increased workload produced changes in the participants' speech. The participants' speech was recorded while they performed an arithmetic task under the low-workload condition (with no time constraint) and under the high-workload condition (with a time constraint). Results indicated that the participants' speech under the high-workload (or time-constrained) condition was associated with changes in amplitude and formant frequency. The authors also mention articulatory changes, such as devoicing and aspiration at certain points of speech, less precise articulation, omission of speech sounds and slurred syllables; however, such changes may result from increased speech rate in the time-constrained condition. In addition, cognitive workload was likely confounded with psychological stress in the Hecket et al. (1968) experimental paradigm, given that participants were required to perform an arithmetic task under various degrees of time pressure and that there was monetary compensation offered for correct answers. A number of studies have investigated the effects of concurrent task performance on speech production in an experimental setting. Lively et al. (1993), for example, found that for some participants, significant acoustic changes (increased amplitude, amplitude variability, decreased spectral tilt) in the production of an excised vowel token occurred in a concurrent task paradigm in which participants were required to perform a visual monitoring task while repeating a sentence. Brenner and Shipp (1987) also noted significant acoustic changes to the speech signal (fundamental frequency, amplitude, speech rate) in response to a concurrent task in which participants were required to perform a visual monitoring task while counting aloud. Interestingly, although the experimental paradigm was not a simulation of an aviation scenario, the authors considered the cognitive workload imposed by the concurrent task paradigm to be analogous to the type and level of cognitive workload faced by a pilot in routine flight operations.

In another study using a concurrent task paradigm, where participants were required to perform a speech task concurrent with a manual shape-identification task, Ooman and Postma (2001) found that the production of dysfluencies (filled pauses and repetitions) increased. Jou and Harris (1992) likewise investigated the effects of divided attention on language production in a dual-task paradigm. This study is interesting in relation to the present study because participants were required to recall information learned and perform a mental arithmetic simultaneously. Relative to the control condition, speech produced in the divided-attention or concurrent-task paradigm was less fluent (characterized by more frequent and longer pauses), showed a detriment to clausal and sentential structures, and in addition, less information was recalled. Finally, Dromey and Benson (2003) found that three types of concurrent tasks (motor, linguistic, cognitive) had an effect on several labial-kinematic measures (measurements of lip movements). More specifically, the concurrent cognitive workload imposed by the concurrent task of counting backwards from 100 in increments of 7, while uttering the control sentence, resulted in decreased lip-movement stability as well as in faster speech rate (although this result was likely due to time-constraints imposed on participants in the high cognitive workload condition).

While all of the studies discussed above have measured the effects of cognitive workload on objective measures of L1 speech production (acoustic, temporal, labio-kinematic), only two of the studies (Hecker et al., 1968; Lively et al., 1993) have investigated the effects of workload on L1 listener perception of L1 speech. Hecker et al. (1968) found that, for some speakers, listeners could identify with 90 percent accuracy utterances that were produced in the high-workload condition, whereas for other speakers, listeners could identify the high-workload utterances only at chance level. Lively et al. (1993) found that perceptual identification of utterances produced in the high-workload condition paralleled the acoustic measurements, in that listeners were able to identify the utterances produced in the high-workload condition but only for speakers who showed robust changes in speech production. The results of both of these studies (Hecker et al., 1968; Lively et al, 1993) suggest a high degree of individual variability in the production of perceptual changes in the speech signal in conditions of high cognitive workload. Taken together, these findings suggest that such changes to speech production may be even more perceptible in the speech of non-native speakers, particularly in high workload conditions. Because there are no studies to date that have investigated the effect of cognitive workload on measures of L2 speech production (objective or subjective), one goal of the present study was therefore to investigate how cognitive workload affects the speech production of L2 speakers in the context of controller-pilot communications.

Summary and Conclusions

The above literature review indicates that L2 proficiency can be paramount to air safety, due to the necessity of accurate and efficient controller-pilot communications in a common language, which is often the L2 of one or both interlocutors. It has been determined that the jobs of both controllers and pilots involve concurrent task performance, that message length affects navigation task performance (Barshi 1997, 1998; Barshi & Healy, 1998, 2002, Mauro & Barshi, 1999; Schneider et al, 2004) and readback accuracy (Barshi, 1998; Schneider et al., 2004), and that the effects of message length on navigation task performance are greater for low- than for high-proficiency non-native speakers (Barshi & Healy, 1998). The present study therefore attempted to replicate these previous findings and to extend them by investigating whether message length similarly affects the speech production of native- and non-native-speaking pilots, as indicated by measures of perceived accentedness, comprehensibility, fluency, and confidence.

It has also been determined that (increased) workload affects native-speaking pilots' task performance (e.g., Raby & Wickens, 1994). However, there is a paucity of research investigating the effects of workload on both task performance and speech production in non-native speakers. The present study thus investigated the effects of workload (high vs. low) on L1 speakers and L2 speakers of varying levels of proficiency. To summarize, the present study examined the effects of message length and workload on navigation accuracy, message repetition accuracy and speech production (using measures of perceived accentedness, comprehensibility, fluency, and confidence) in a simulated pilot navigation task, testing participants divided into three groups representing native-

English speakers and speakers of relatively high and low levels of proficiency in the English language.

Hypotheses

The present study investigated the following hypotheses:

1. The length of the message would affect participants' task performance (navigation and readback accuracy) and speech production (accentedness, comprehensibility, fluency and confidence) in a simulated pilot navigation task in that longer messages would result in a performance detriment on all measures. The prediction here was that the effects of message length would differ for all groups and that the greatest drop in task performance and speech production would occur in the group with the lowest level of L2 proficiency.
2. Increased cognitive workload, that is, the addition of a concurrent arithmetic task, would adversely affect all participants' task performance (navigation and message repetition accuracy) and speech production (accentedness, comprehensibility, fluency and confidence). The prediction here was that the effects of increased cognitive workload would differ for all groups and that the greatest drop in task performance and speech production would occur in the group with the lowest level of L2 proficiency.

In the following chapter, the methodology for the present study is described.

CHAPTER 3

METHODOLOGY

Chapter Overview

This chapter provides an overview of the participants, materials and procedures used in the present study. First, participant-selection criteria are discussed and the participants' demographic and language-background characteristics are presented. Second, materials used in the present study are described. Next, the experimental procedure is outlined. Finally, the dependent variables and methods of data analysis are discussed.

Participants

Selection of Participants

Participants were 62 university students (47 males, 15 females) enrolled in university engineering programs (mean age: 27.2 years; range: 19-36; *SD*: 5.6). All were volunteers recruited from the engineering departments of English-medium Montreal universities and were paid $15 for their participation. Engineering students were selected in order to maintain relative uniformity in the participants' training background and to maintain similarities between their training and that of pilots, who often have an engineering or technical training background. Volunteers with pilot, air traffic control or other relevant aeronautical experience were not selected for the study in order to control for expertise effects. Data obtained from two participants (both male) were excluded from the final data analyses. The first participant was excluded due to a malfunctioning of the recording equipment which resulted in the loss of a large portion of the data. The second participant was excluded because he was unable to perform the primary experimental task and other tasks, to an extent indicative of a learning disability. The following description of the participants' language background is based on the data of the final 60 participants (45 male, 15 female).

Participants' Background

All participants completed a language background questionnaire (Appendix A). The 60 participants were divided into three groups of twenty. The first group was comprised of 20 native English speakers. This group will henceforth be referred to as "NS". The second and third groups were comprised each of 20 native speakers of a Chinese language (Mandarin or Cantonese), all of whom had learned English as their second language (L2). The participants in the second group had a higher level of proficiency in English than those in the third group. (The specific proficiency measures used to divide these participants into two proficiency groups are discussed later in this chapter.) These two groups will henceforth be referred to as "High" and "Low", respectively.

All NS participants had been raised in monolingual or bilingual homes with at least one native English-speaking parent/guardian, and had been exposed to English from birth. One of the 20 native English speakers was a balanced English/French bilingual. Of the 40 Chinese participants, 37 were native speakers of Mandarin, two were native speakers of Cantonese, and one was a balanced Mandarin-Cantonese bilingual. All 40 of the High and Low group participants were born and raised in a Mandarin and/or Cantonese speaking environment in China, and all were proficient speakers of Mandarin who continued to use Mandarin in their daily lives. Thirty nine of the 40 Mandarin-speaking participants reported that they currently spoke Mandarin at home, while one reported currently speaking English at home (due to residence with a non-Mandarin speaking roommate). Thirty nine were born and raised in Mainland China, while one was born and raised in Taiwan. All Mandarin-speaking participants arrived in Canada as adults to pursue post-secondary education (mean age: 28.0 years; range: 19-41; *SD*: 4.7), and all spoke English as an L2.

Participants' Language Proficiency Self-Evaluation

All participants rated their native language (L1) and L2 abilities on a 9-point Likert scale (1 = *very poor*, 9 = *extremely fluent*) for the following competencies: speaking, listening, reading and writing. The purpose of the L1 ratings was to verify that all participants were native speakers of English, Mandarin or Cantonese. Participants also indicated the percentage of time (0-100%) that they used their L1 and L2 each week (speaking, listening to media, reading, writing). Results of L1 self-ratings are reported in Table 1.

Table 1

L1 Self-ratings and Usage

Measure	Proficiency Groups (*n*=60)					
	NS (*n*=20)		High (*n*=20)		Low (*n*=20)	
	M	*SD*	*M*	*SD*	*M*	*SD*
L1 speaking[a]	8.8	0.41	8.8	0.44	8.7	0.57
L1 listening[a]	8.6	0.75	8.8	0.41	8.8	0.44
L1 reading[a]	8.7	0.49	8.8	0.38	8.8	0.64
L1 writing[a]	8.5	0.61	8.4	0.75	8.3	0.80
L1 use in speaking[b]	91.0	9.68	61.0	23.15	67.0	22.03
L1 use in listening to media[b]	91.0	12.52	35.5	31.67	39.0	35.67
L1 use in reading[b]	94.5	6.87	33.0	21.79	40.5	24.81
L1 use in writing[b]	97.0	5.71	35.0	31.71	39.0	29.36

[a] Measured on a 9-point Likert scale

[b] Measured on a 10-point scale (0-100%)

Mandarin-Speaking Participants' L2 Background

Only results of the language background questionnaire pertaining to the L2 background of the Mandarin-speaking participants (High and Low groups) will be reported as it is the L2 background of these groups that is of interest for the present study. The L2 background of the High and Low group participants was fairly homogeneous. All began learning English in late childhood or early adolescence (mean age: 11 years; range: 6-15; *SD*: 2), and all began their English language training in the classroom in China. All received some university education in China (in Mandarin) and then came to Canada to pursue further post-secondary education (in English). Between-group comparisons revealed no significant differences between groups in age of arrival in Canada, length of residence in Canada, age of exposure to English, years of formal English language instruction or hours per week of formal English language instruction ($p > .05$). A summary of Mandarin-speaking participants' L2 background appears in Table 2.

Table 2

L2 Background, Self-ratings and Usage for High and Low Groups

Measure	Proficiency Group			
	High (*n*=20)		Low (*n*=20)	
	M	*SD*	*M*	*SD*
Age of arrival in Canada (in years)	26.2	4.12	29.8	4.61
Length of residence in Canada (in years)	2.3	1.07	2.1	1.62
Age of exposure to English (in years)	10.7	1.92	12.2	1.82
Years of English language learning	12.9	4.18	13.5	6.03
Hours per week of English language learning	5.7	2.73	4.8	2.7
L2 speaking self-rating[a]	5.7	1.18	5.2	1.18
L2 listening self-rating[a]	5.9	1.46	5.5	1.47
L2 reading self-rating[a]	7.1	1.19	6.8	1.55
L2 writing self-rating[a]	5.7	1.21	5.5	1.67
Pct/week L2 speaking[b]	37.5	23.14	31.0	18.89
Pct/week L2 listening to media[b]	56.0	29.45	58.5	34.53
Pct/week L2 reading[b]	63.0	24.51	58.0	24.40
Pct/week L2 writing[b]	61.0	31.44	59.5	28.56

[a] Measured on a 9-point Likert scale
[b] Measured on a 10-point scale (0-100%)

Mandarin-Speaking Participants' L2Proficiency

The 40 Mandarin-speaking participants were divided into two experimental groups (High, Low) based on their level of proficiency in English. Three separate proficiency measures were used. Two of the measures (a listening comprehension score and a speaking accuracy score) were obtained from participants during the experimental session. The third measure was a set of listener-rated scores, obtained in a separate session from 10 native-English-speaking listeners recruited from McGill University in Montreal (an English-medium university). These listeners rated short excerpts from the participants' speech samples (approximately 30 seconds in duration) for accentedness, comprehensibility, and fluency.

Listening comprehension test. The purpose of the listening comprehension test was to determine participants' listening comprehension ability in English. A listening diagnostic pretest intended for TOEFL exam preparation (Phillips, 2005) was administered to all participants and consisted of one conversation and one lecture (in English). The conversation and lecture were presented on a personal computer and following each one, participants were presented with six multiple choice comprehension questions (on paper) pertaining to the conversation or the lecture they had just heard (see Appendix B for participant worksheets). Each participant obtained a listening comprehension score out of 12 for this test, with one score assigned for each of the 12 correctly answered comprehension questions.

Oral interview. The purpose of the oral interview was to obtain a speech sample from each participant for the purpose of evaluating oral-proficiency level. The task consisted of a brief interview in which the participant was asked to speak for approximately two minutes on a topic of personal interest or experience, for example, their last trip/vacation or the differences between university life in Canada and China. Participants' responses were digitally recorded directly onto a PC running *CoolEdit* software using a Plantronics DSP-300 headset microphone.

The recorded speech samples were first analyzed for speaking accuracy. A measure of speaking accuracy was obtained by calculating the number of errors in the recorded speech samples. Errors were defined as omitted words or phonemes, and any mistakes in sentence structure, morphology or syntax. Common pronunciation mistakes, particularly those typical of Chinese learners of English (e.g., /r/-/l/ substitutions, mispronounced word-final stops) were not considered errors. For each participant, a speaking accuracy score was defined as a proportion of errors, calculated by dividing the total number of errors (as described above) in a speech sample by the total number of words in it.

Listener ratings. The recorded speech samples were then presented to listeners for accentedness, comprehensibility and fluency rating in a native speaker rating task (Munro & Derwing, 1995a, 1995b). In this task, ten native-English-speaking students recruited from McGill University (mean age: 24 years; range: 19-33 years; *SD*: 4.3 years) rated 30-second samples of the recorded oral interviews for degree of accentedness, comprehensibility and fluency on a 9-point Likert scale. For accentedness (1 = *heavily accented*, 9 = *not accented at all*), the raters were told to estimate the degree of foreign accent in the participants' speech, disregarding acceptable pronunciations typical of native regional varieties of English. For comprehensibility (1 = *hard to understand*, 9 = *easy to understand*), raters were told to judge how difficult or easy it was to understand what the participants were saying. For fluency (1 = *not fluent at all*, 9 = *very fluent*), the raters were asked to judge the degree to which the participants' speech sounded fluent (i.e., spoken without undue pauses, filled pauses, hesitations, or dysfluencies such as false starts and repetitions). The samples were presented to the raters via a loud speaker in a quiet room and the raters recorded their scores for each speech sample on worksheets provided (see Appendix C). The raters received a small monetary remuneration for their participation.

Global proficiency score. In order to assign the Mandarin-speaking participants to the experimental groups based on *all* of the proficiency measures obtained (listening comprehension score, speaking accuracy score, and listener-rated judgments of accentedness, comprehensibility and fluency), the following formula was used to derive a single global proficiency score for each participant:

$$\frac{(accentedness + comprehensibility + fluency) \times listening\ \ comprehension}{speaking\ \ accuracy}.$$

The assumption here was that high accentedness, comprehensibility, fluency and listening comprehension scores and a low speaking accuracy score (i.e., a low proportion of errors) characterize a native-like proficiency.

The Mandarin-speaking participants were rank-ordered based on this global proficiency score and were divided into two equal groups (high and low), with a constraint that there should be an equal number of participants in each group who performed the main experimental task in one of two orders. (This condition was necessary to maintain task-order counterbalancing across groups.) The 20 native English speakers, whose global proficiency scores were all ranked higher than those of Mandarin-speaking participants, formed the comparison group. Independent-samples *t*-tests conducted to compare the global proficiency scores among the three proficiency-based groups yielded significant *t* values ($p \leq .001$). Mean global proficiency scores for the NS, High and Low groups appear in Table 3.

Table 3

Global Proficiency Scores for All Participant Groups

Participant Groups (*n* = 60)	Global Proficiency Scores	
	Mean	*SD*
NS (*n* = 20)	57547.5	276574.9
High (*n* = 20)	1019.0	587.4
Low (*n* = 20)	471.5	209.4

To ensure that the resulting three proficiency-based groups differed on all proficiency measures, the proficiency measures used in the calculation of the global proficiency score were compared among the groups using separate one-way analyses of variance (ANOVAs). These analyses yielded significant *F* ratios for listening comprehension scores, $F(2,57) = 27.35, p < .0001$, speaking scores, $F(2, 57) = 99.89, p < .0001$, as well as ratings of accentedness, $F(2, 57) = 347.48, p < .0001$, comprehensibility, $F(2, 57) = 157.18, p < .0001$, and fluency, $F(2, 57) = 225.26, p < .0001$). For all measures, Tukey HSD post-hoc tests revealed significant differences among the three proficiency groups ($p < .05$). Mean values for all proficiency scores are presented for each group in Table 4.

Table 4

Results of Proficiency Measures Contributing to Global Proficiency Score

Proficiency Measure	Participant groups (*n* = 60)					
	NS (*n* = 20)		High (*n* = 20)		Low (*n* = 20)	
	Mean	SD	Mean	SD	Mean	SD
TOEFL listening diagnostic pretest [a]	11.06	0.84	9.13	1.26	7.89	1.81
Lexical and morphosyntactic errors[b]	0.01	0.00	0.14	0.05	0.19	0.06
Accentedness[c]	8.53	0.37	3.65	1.14	2.55	0.57
Comprehensibility[c]	8.66	0.56	5.15	1.21	3.86	0.76
Fluency[c]	8.68	0.25	5.05	1.01	3.91	0.74

[a] Score /12

[b] Proportion of lexical and morphosyntactic errors in spontaneous speech samples from oral interview

[c] NS ratings of participant speech samples /10

Materials

Background Questionnaire

The background questionnaire (Appendix A) sought information regarding the participants' demographic and language background (i.e., age, gender, place of birth, parents' native language, age at the time of arrival in Canada, field of study, languages in which formal education was obtained from elementary school to university), as well as the participants' history of language learning and use (i.e., languages spoken at home from childhood to the present, nature and amount of L2 training, degree of L1/L2 use).

Main Experimental Task

Navigation task

The navigation task used was the pilot-navigation task created by Barshi (1997, 1998). Versions of this task have been used in several studies to simulate air traffic control messages in experimental settings (Barshi, 1997, 1998; Barshi & Healy, 1998, 2002; Mauro & Barshi, 1999; Schneider et al., 2004). Each participant performed two versions of the navigation task: the "clear" condition and the "workload" condition. (The order of the conditions was counterbalanced between participants in order to control for practice effects.) In order to create the workload condition, the task was modified for the purposes of the present study by the addition of a concurrent task (described below). The materials used in the experimental task consisted of a number of auditorily-presented messages that varied in length. The messages were comprised of commands or instructions to navigate in four 4 × 4 "navigation" grids (see task description below). Each command had the following basic structure: verb + directional particle + number of displacements (1 through 3). A combination of a unique verb and a directional particle corresponded to each of the three dimensions in the navigation space (Table 5).

Table 5

Partial Composition of Commands Used in the Navigation Task

Dimension	Verb	Directional particle
horizontal within the same grid	Turn	left/right
vertical between separate grids	Climb	up/down
vertical within the same grid	Move	forward/back

Within each dimension, the number of "displacements" from any current location in the grids was limited to 1-2 due to the starting point on the 4 × 4 grid (see Appendix E for a graphic representation of the grid and the starting point). This restriction ensured that the instructions given did not cause participants to "fall off" the grids (see Barshi, 1997). Each movement direction or dimension was designated by a unique "displacement" marker. The displacement markers used for each dimension appear in Table 6.

Table 6

Complete Composition of Commands Used in the Navigation Task

Dimension	Verb	Directional particle		Number	Displacement
horizontal within the same grid	Turn	right	left	1, 2 or 3	square(s)
vertical between separate grids	Climb	up	down	1, 2 or 3	level(s)
vertical within the same grid	Move	back	forward	1, 2 or 3	step(s)

The resulting messages varied in length between 1 and 3 commands. However, the order of commands in a sequence was always fixed: turn, climb, move, roughly corresponding to a fixed order of typical controller commands for pilots: heading, altitude and speed (Barshi, 1997). An example of a three-command message was thus: (a) *turn* right two squares, (b) *climb* up one level, (c) *move* forward one step. The messages were spoken by a male native speaker of a standard American dialect of English and digitized on a Macintosh SE30 computer using *SoundEditPro* (Barshi, 1997). The clearest sample of each word was selected and the messages were composed of the spliced words in order to allow for uniform pause duration between words as well as uniform word duration between messages. This process also maintained a natural stress pattern within each isolated word (Barshi, 1997). Participants were required to repeat the messages heard and click on a button labeled "DONE" before carrying out the instructions by clicking on the appropriate squares on the navigation grids. Once the instructions had been carried out and "navigation" was complete, participants clicked again on the button labeled "DONE" and in doing so automatically proceeded to the next message. For the present study, the task was presented on a Macintosh G4 iMac computer running *Hypercard* presentation software.

Concurrent or "workload" task

In the workload condition, participants performed a concurrent addition task. For this task, a number between 11 and 99 appeared randomly in one of six squares surrounding the navigation grids 0.5 seconds after the auditory presentation of the message, as described above. Participants were required to invert the number and mentally add the original and inverse number together *while* repeating the message. At the end of repeating the message, the participant was required to report the answer to the arithmetic problem by stating the following: "*Answer, (the number)*". As in the clear condition, participants then clicked on a button labeled "DONE", then carried out the instructions, or navigated, by clicking on the appropriate squares. Once navigation was complete, participants clicked on the button labeled "DONE" again, and by doing so automatically proceeded to the next message.

Procedure

Pre-Experimental Tasks

Once selected on the basis of the criteria described above (language and engineering background), participants were invited to individually attend an experimental session. The testing was conducted in the CSLP SAGE laboratory using a Macintosh G4 iMac computer running *Hypercard* presentation software, and a personal PC laptop computer running speech presentation software (Smith, 1997) and *CoolEdit* speech recorder. A Logitec USB desktop microphone (connected to both the MAC and the PC) was placed in front of the participant. The entire task (i.e., navigation messages, their repetitions by participants and, in the workload condition, participants' answers to the concurrent arithmetic task) was recorded directly onto the PC using *CoolEdit* software. Prior to performing the computer-based navigation task, the participants filled out consent forms and completed the language background questionnaire with the assistance of the experimenter.

Navigation Task

Navigation practice tasks

The primary experimental task, or "navigation task" (described above), was introduced to participants using a wall-poster visual of the experimental paradigm which illustrated the three movements (*turn, climb, move*) on the four 4 × 4 grids. After the introductory poster demonstration, participants were shown (on paper) an example of the navigation movements following a sample message (see Appendix E for example sheet).They then completed a similar practice on paper (see Appendix F for practice sheet) in which they were required to number the squares upon which they would click in response to the sample message.

For the actual computer-based practice, the participant sat at the MAC computer. Participants performed 12 practice trials in which they listened to recorded "controller" messages and practiced the navigation task as described above in the *Materials* section. The practice trials contained four samples of each command type and message length (1-3), and began with a one-command message. The practice phase preceded each of the conditions (which were counterbalanced between participants), and in the workload condition, the practice phase included the concurrent addition task described above. Participants were provided, through the software, with built-in feedback regarding their navigation accuracy (Barshi, 1997). Upon clicking the button labeled "DONE", a male voice indicated whether or not the participant had clicked on the correct squares by saying "Correct" or "Wrong". During the practice trials, the experimenter answered any questions the participant may have had regarding the task.

Navigation experimental task

Each of the experimental conditions of the navigation task consisted of 36 experimental trials divided into three 12-trial blocks. The trials (messages of 1-3 commands) were presented in pseudo-random order, with each block containing four samples of each message length. Each participant completed a total of six 12-trial blocks (3 blocks in each condition, clear and workload). Four randomized lists of the command stimuli were created in order to control for order effects, and these were counterbalanced across participants. Participants were presented with a different list for each condition in order to control for practice effects.

The procedure in the experimental phase was as follows: Each trial started with the presentation of a message followed by a beep. Following the beep, the participant repeated the instructions heard. In the workload condition, the participant simultaneously calculated the addition problem and provided the answer at the end of the message repetition. Once the message was repeated, the participant clicked on the button labeled DONE, and then carried out the instructions by clicking in the appropriate places on the grid, starting from the highlighted square. For example, in the clear condition, the participant may hear "Turn left one square. Climb up one level. Move forward one step." (three-command message) followed by the beep. They would then attempt to repeat (read back) the commands verbatim, click DONE upon completion, then carry out the instructions by clicking on the appropriate spaces on the grid, and clicking the DONE button again upon completion. Each participant took approximately 30 minutes to complete the clear condition of this task.

In the workload condition the participant performed the task exactly as outlined above, except that the message repetition was followed by the answer to the addition problem. For example, in response to the example provided above, if the number "12" were randomly presented following the presentation of the message, the participant would say: "Turn left one square. Climb up one level. Move forward one step. Answer, thirty three". The participants were required to click on each square they passed, so that their "navigational path" could be traced in order to measure navigation accuracy. Each participant took approximately 30 minutes to complete the workload condition of this task.

Think-aloud questions

Once both versions of the navigation task (clear, workload) had been completed, the experimenter asked the participant to indicate which version of the task they found more difficult and to explain why. The participant was also asked to describe any strategies they may have employed to facilitate task performance. The purpose of these questions was to inform the experimenter of performance strategies employed by the participants. At this point in the experimental session, participants were given a 10-minute break.

Listening comprehension task

Upon returning from the break, the participant completed the listening diagnostic pre-test described in the *Materials* section above. The participant answered the six multiple choice questions pertaining to each conversation and lecture (presented on a PC using Windows Media Player) on a piece of paper provided by the experimenter following each conversation/lecture. The participant answered a total of twelve questions. This task lasted approximately 15 minutes.

Oral proficiency interview

For this task the participant wore a Plantronics DSP-300 microphone headset and the monologue was recorded onto a PC using *CoolEdit* software.

Addition task

The participant completed the addition task described above. The participant was instructed to complete as many of the arithmetic problems as possible in one minute. The experimenter timed the participant with a stopwatch and the participant was instructed to stop writing when one minute was up.

Other tasks

The participants also performed three other attention and memory tasks which are not relevant to the research questions investigated here.

Design

Research design

A 3 × 2 × 3 mixed factorial design was used in the present study. The between-participant variable—language proficiency—had three levels, as described above in the *Participant* section (NS, High, Low). The within-participant variables were condition (clear, workload) and message length (1, 2, 3). Within each level of the between-participant variable (language proficiency), the participants were tested in both conditions (clear, workload) using messages of all lengths (1, 2, 3). The order of conditions (clear-workload, workload-clear) was counterbalanced across participants.

Dependent variables

There were two sets of dependent variables. The first set included two measures of performance accuracy. The first was the participants' accuracy on the navigation task, measured as the number of trials (for each message length) in which the participant navigated accurately following all the commands in a given message. The second included the content accuracy of the repeated instructions, henceforth referred to as repetition accuracy. Repetition accuracy was measured as the number of trials (for each message length) that the participant accurately repeated all of the critical (i.e., non-redundant) words in a given message (Schneider et al, 2004). For example, if the message "Turn left one square. Climb up one level" was presented, then the participant had to repeat the critical words *left one* and *up one*" in order to obtain a score of 1/1 for that particular message. Any errors in the repetition of the critical words resulted in a score of 0/1 for that particular message. Errors in either the verb or displacement marker did not result in a deduction of errors as these errors were not likely to contribute to navigation errors. This strict scoring method was used for consistency with the scoring method for navigation accuracy. Participants obtained a score out of 12 for each message length (1, 2, 3).

The second set of dependent variables included four speech-production measures based on the participants' recorded repetitions of the messages. These measures of speech production were listener-based measures: *perceived* accentedness, comprehensibility, fluency and confidence. These measures were obtained from the ratings of 10 native English speaking raters (2 males, 8 females) who were university students and had no language teaching experience (mean age: 24.5 years; range: 20-35; *SD*: 4.3). The raters who participated in this rating were different individuals from those who participated in the language proficiency rating described above in the *Materials* section.

Ratings based on four criteria (accentedness, comprehensibility, fluency and confidence) were obtained for speech samples for each of the 60 participants in each condition (clear, workload). (See Appendix G for rater worksheet.) The raters were asked to rate each of the samples on four 9-point scales (accentedness: 1 = *heavily accented*, 9 = *not accented at all*; comprehensibility: 1 = *hard to understand*, 9 = *easy to understand*; fluency: 1 = *not fluent at all*, 9 = *very fluent*; confidence: 1 = *not confident at all*, 9 = *very confident).* For the accentedness, comprehensibility and fluency measures, the raters were instructed to rate the speech samples according to the criteria described above (see *Oral Interview* in the *Materials* section). For the confidence measure, the raters were instructed to consider the confidence level of the speaker, that is, whether the speaker conveyed confidence in the accuracy of the content of the message repetition.

The speech samples presented were messages excised from the message repetition recordings of each participant. For each participant, two randomly selected samples of each message length (1-3) were chosen from the recordings of each condition (clear, workload) and were saved as separate audio files using *CoolEdit* software. A total of 12 samples (6 per condition) per participant were spliced from the recordings for a total of 720 samples. All samples were normalized to ensure uniformity in perceived loudness. The 720 samples were randomized and presented to raters in eight blocks of 90 samples each. Raters were given a 10-minute break after every second block in order to minimize fatigue. The ratings were obtained in separate sessions using two presentation orders. The presentation order was counterbalanced between sessions to control for both experience and rater fatigue effects. Five raters rated the 720 samples presented in one order, and the remaining five rated the 720 samples presented in the other order.

Data analysis

Within each level of the between-participant variable (language proficiency), the data were tabulated separately by condition (clear, workload) and message length (1-3). The data were then submitted to three-way ANOVAs with language proficiency (NS, high, low) as a between-subjects variable and condition (clear, workload) and message length (1-3) as within-subjects variables. Significant main effects and interactions were further explored in two-way ANOVAs for the clear condition and two-way ANCOVAs (with the results of the addition task entered as a covariate) for the workload condition. Effects of workload were explored in two-way ANCOVAs for each message length, with condition as a within-subjects variable and proficiency as a between-subjects variable. Further comparisons were explored using one-way ANOVAs or Bonferroni tests. Alpha level was set at .05 for all statistical comparisons. The analyses described above were conducted for each of the dependent variables: (a) navigation accuracy, (b) repetition accuracy, (c) accentedness, (d) comprehensibility, (e) fluency, and (f) confidence.

Chapter Summary

This chapter provided a description of the participants, materials, and procedures used in the present study. In selecting the participants and experimental materials, careful consideration was taken to approximate (insofar as possible) the controller-pilot communicative environment. The findings of the present study are described in the following three chapters.

CHAPTER 4

RESULTS I

Chapter Overview

This chapter summarizes the results of between-group statistical analyses of the navigation accuracy and repetition accuracy data. First, the objectives and hypotheses related to these variables are recapitulated. Next, results of statistical analyses are presented. For each of the dependent variables, a summary of findings is provided at the end of each section (message length and workload). Findings related to each of the hypotheses are summarized at the end of the chapter.

Main Objective and Hypotheses

One main objective of the present study was to determine whether and to what extent high- and low-proficiency L2 speakers are affected in their task performance by high workload involving L2 communication. This objective was accomplished by measuring the participants' comprehension and retention of information, using measures of navigation accuracy and repetition accuracy. Two hypotheses were proposed. The first hypothesis was that the length of the message would affect the participants' performance in a simulated navigation task in that longer messages would lead to a performance detriment on all measures. The prediction here was that the effects of message length would differ for all groups and that the Low group would demonstrate the greatest drop in task performance. The second hypothesis was that increased cognitive workload, that is, the addition of the concurrent arithmetic task, would adversely affect all participants' task performance. The prediction here was that the effects of increased cognitive workload would differ for all groups and that the Low group would again demonstrate the greatest drop in task performance.

Individual Differences in Mental Arithmetic Ability

Because workload was defined in the present study as the participants' ability to carry out simple arithmetic problems, it was important to determine that there were no individual differences among the participants in this ability or that such differences (if present) were distributed equally across the three participant groups. The first analysis, therefore, compared the participants' scores on the one-minute addition task. In this task, the participants were asked to mentally solve as many simple arithmetic problems as possible within one minute. The number of arithmetic problems solved was the dependent variable in this analysis.

The participants' scores were submitted to a one-way ANOVA which yielded a significant F ratio, $F(2, 59) = 4.29, p = .018$. Post-hoc Tukey HSD tests ($\alpha = .05$) used to explore this significant effect revealed three findings. First, the High group solved, on average, significantly more arithmetic problems than the NS group did (20.9 vs. 15.1, $p = .019$). Second, the Low group showed a similar tendency, although this difference failed to reach statistical significance (19.6 vs. 15.1, $p = .09$). Finally, both groups of Chinese participants (High and Low) solved, on average, a similar number of arithmetic problems (20.9 vs. 19.6, $p = .80$).

Taken together, these findings suggested that the effect of workload in the navigation task may have differed among the three participant groups because these groups varied in their ability to perform mental arithmetic problems. It is likely, for example, that the effect of workload may have been more pronounced for the NS group than it otherwise should have been because these participants, in comparison to the Chinese participants in the High and Low groups, had more difficulty solving mental arithmetic problems. Because, as these analyses suggested, workload was confounded with the participants' ability to solve mental arithmetic problems, analyses of covariance (ANCOVAs) were used in all further between-group comparisons exploring the effect of workload, with the participants' performance on the addition task entered as a covariate. Tests of homogeneity of regression slopes, carried out prior to conducting all ANCOVAs reported in this and the following chapters, revealed that the covariate did not interact with proficiency (between-subjects variable). This suggested that the effect of the covariate was comparable at all three levels of the proficiency factor, an important ANCOVA assumption.

Navigation Accuracy

Overall Analysis

The navigation accuracy data were submitted to a three-way repeated-measures ANOVA in order to determine if the three groups differed in navigation accuracy and if such differences interacted with message length and workload condition. In this analysis, proficiency served as a between-subjects factor (NS, High, Low); condition (clear, workload) and length (1, 2, 3) served as within-subjects factors. First, this analysis yielded a significant main effect of proficiency, $F(2, 57) = 3.65, p = .032$. Overall (regardless of message length and workload condition), the NS group performed more accurately than the Low group ($p = .011$, $\alpha = .017$) and the NS and the High groups did not differ in their performance. Second, this analysis also yielded a significant main effect of condition, $F(1, 57) = 43.23, p < .001$. All groups performed more accurately in the clear than in the workload condition ($p < .001$, α = .025). Finally, this analysis also yielded a significant main effect of length, $F(2, 114) = 80.98$, $p < .001$. Overall navigation accuracy was more accurate for message length 1 than for message lengths 2 and 3, and for message length 2 than for message length 3 ($p < .001$, α = .017). This omnibus ANOVA yielded no significant interaction effects although the condition × length interaction approached significance, $F(2, 114) = 2.95$, $p = .056$. Mean navigation accuracy scores are presented in Table 7.

Table 7

Mean Navigation Accuracy Scores (out of 12) and their Standard Deviations for the Three Proficiency Groups in the Clear and Workload Conditions

Condition	Length	Proficiency Groups					
		NS (*n* = 20)		High (*n* = 20)		Low (*n* = 20)	
		M	*SD*	*M*	*SD*	*M*	*SD*
Clear	1	11.90	.308	11.85	.366	11.70	1.13
	2	11.45	.826	10.70	1.72	10.70	1.49
	3	10.00	2.55	8.30	2.98	8.45	2.86
Workload	1	11.10	1.45	11.00	1.17	10.95	1.32
	2	10.75	1.02	9.30	2.34	9.35	2.08
	3	8.15	2.85	7.60	3.49	5.90	3.13

Despite the absence significant interactions in this omnibus analysis, further analyses (reported below) were conducted in order to explore the significant main effect of proficiency. This main effect of proficiency was a consistent trend in the analyses of all of the dependent variables reported in the present study; therefore, in order to investigate this trend, further analyses, such as the ones reported below for navigation accuracy, were conducted for each of the dependent variables.

Message Length

A series of analyses were carried out to explore the effect of message length on the participants' navigation accuracy in each condition. For these analyses, the data for the clear condition were submitted to a two-way repeated-measures ANOVA with proficiency (NS, High, Low) as a between-subjects factor and message length (1, 2, 3) as a within-subjects factor. The data for the workload condition were submitted to a two-way repeated-measures ANCOVA with proficiency (NS, High, Low) as a between-subjects factor, message length (1, 2, 3) as a within-subjects factor, and the addition task as a covariate.

Clear condition. The ANOVA comparing navigation accuracy data in the clear condition yielded significant main effects of length, $F(2, 14) = 45.31, p < .001$, and proficiency, $F(1, 57) = 3.14, p = .05$. Pairwise comparisons ($\alpha = .006$) carried out to explore these significant effects revealed that navigation accuracy depended on message length. The NS group performed significantly more accurately on message lengths 1 and 2 than on message length 3 ($p = .003$). By contrast, both the High and the Low groups performed significantly more accurately on message length 1 than on message lengths 2 and 3 and on message length 2 than on message length 3 ($p \leq .001$). Navigation accuracy scores for each message length in the clear condition are illustrated in Figure 1.

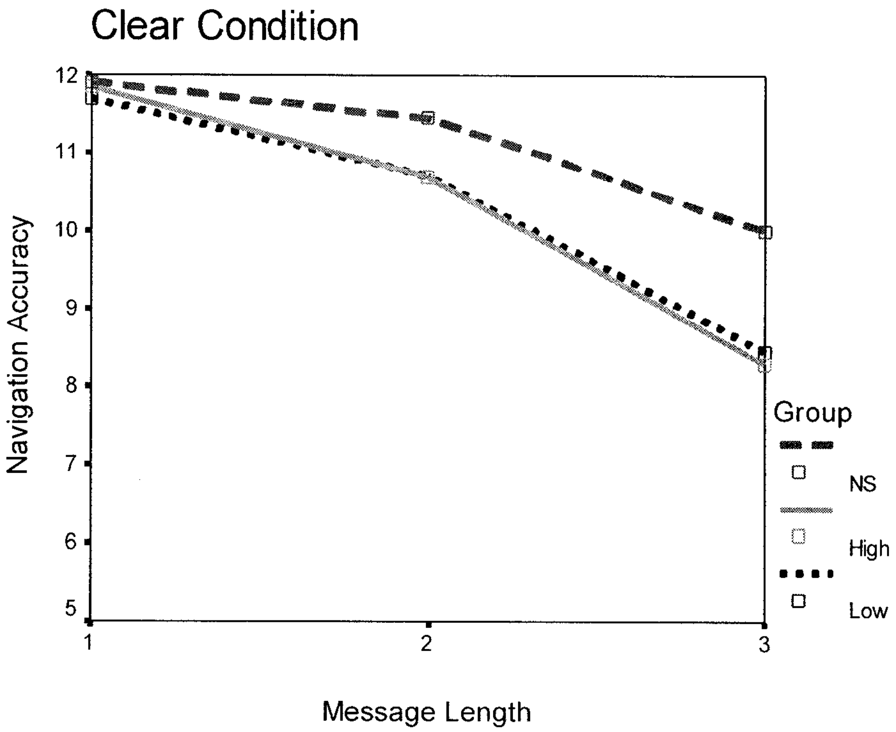


Figure 1. Navigation accuracy scores for each participant group (NS, High, Low) for each message length in the clear condition.

Workload condition. The ANCOVA comparing navigation accuracy data in the workload condition yielded significant main effects of length, $F(2, 112) = 19.11, p < .001$, proficiency, $F(2, 56) = 5.86, p = .005$, and the addition task, $F(1, 56) = 9.90, p < .003$. This analysis also revealed a significant length × proficiency interaction, $F(4, 112) = 3.11$, $p = .018$. Pairwise comparisons carried out to explore the significant interaction revealed two findings. First, navigation accuracy depended on participants' proficiency. At message length 1, all groups navigated with about equal accuracy; that is, there were no significant between-group differences in navigation accuracy. By contrast, at message length 2, the NS group navigated significantly more accurately than the High *and* Low groups ($p \leq .004$), and at message length 3, the NS navigated significantly more accurately than the Low group ($p < .003$). The second finding was that navigation accuracy depended on message length. The NS group navigated significantly more accurately for messages of lengths 1 and 2 than for message of length 3 ($p \leq .001$). By contrast, the High and Low groups displayed a significant decline in navigation accuracy with each increasing message length ($p \leq .006$). Navigation accuracy scores for each message length in the workload condition are illustrated in Figure 2.

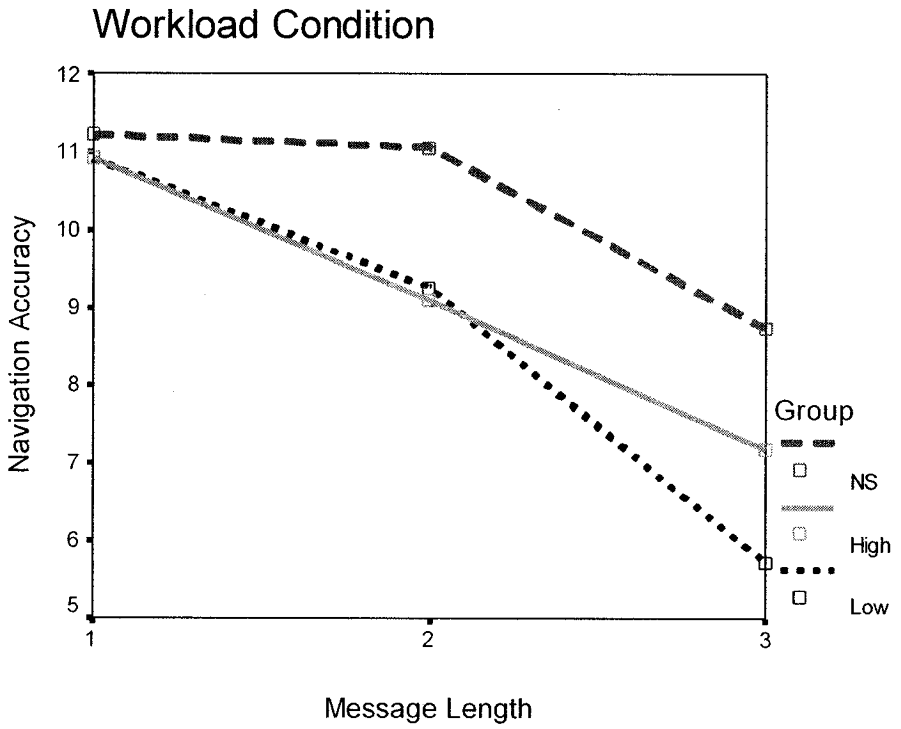


Figure 2. Navigation accuracy scores in the workload condition for each participant group (NS, High, Low) at each message length.

Message length effects. The extent to which message length affects navigation accuracy can also be expressed as a mean difference in navigation accuracy for messages of each consecutive length. This measure (message length effect) is summarized in Table 8 for the clear and workload conditions. In order to determine which group displayed the *greatest* effect of message length, a simple subtraction analyses was conducted. Message length effect was defined as a mean difference in navigation accuracy between the participants' performance on a shorter message and their performance on a subsequent longer message. For each condition, two values of message length effects were calculated: one for the difference in navigation accuracy between message length 1 and 2, the other for the difference in navigation accuracy between message length 2 and 3. A higher value indicates a greater effect of message length and, therefore, a greater performance detriment as a function of increasing message length. In the clear condition, the greatest effect of message length was observed in the High group between messages of length 2 and 3. In the workload condition, the greatest effect was observed in the Low group between messages of length 2 and 3.

Table 8

Effect of Message Length on Navigation Accuracy for the Three Proficiency Groups in the Clear and Workload Conditions

Proficiency Groups	Condition			
	Clear		Workload	
	1-2	2-3	1-2	2-3
NS (*n* = 20)	.45	1.45	.17	2.33
High (*n* = 20)	1.15	2.40	1.83	1.90
Low (*n* = 20)	1.00	2.25	1.66	3.53

Note. Means for the workload condition have been adjusted with the one-minute addition task scores entered as a covariate.

Summary. Analysis of the navigation accuracy data revealed that the length of the message did affect participants' navigation performance. However, the extent of this effect depended on proficiency and workload. At all message lengths in the clear condition *and* at message length 1 in the workload condition, all groups navigated with equal accuracy. By contrast, in the workload condition, the NS group navigated with greater accuracy than both L2 groups (High and Low) at message length 2, and with greater accuracy than the Low group at message length 3. In both the clear and workload conditions, the NS group navigated significantly less accurately when responding to messages of length 3 than when responding to shorter messages. By contrast, the High and Low groups navigated significantly less accurately when responding to messages of lengths 2 and 3 than of length 1, and to messages of length 3 than of length 2. The greatest drop in performance accuracy due to message length was observed for the Low group in the workload condition.

Workload

A series of analyses were carried out to explore the effect of workload on participants' navigation accuracy for each message length. For these analyses, the data for each message length (1, 2, 3) were submitted to three separate two-way repeated-measures ANCOVAs with proficiency group (NS, High, Low) as a between-subjects factor, condition (clear, workload) as a within-subjects factor, and the participants' score on the addition task as a covariate.

Message length 1. The ANCOVA comparing navigation accuracy data for message length 1 yielded a significant main effect of condition, $F(1, 56) = 7.49, p = .008$. Pairwise comparisons ($\alpha = .017$) carried out to further investigate this significant effect revealed that both the High and Low groups performed less accurately in the workload condition than in the clear condition ($p \leq .007$). By contrast, the NS group navigated with equal accuracy in both conditions. The effects of workload for message length 1 are illustrated in Figure 3.

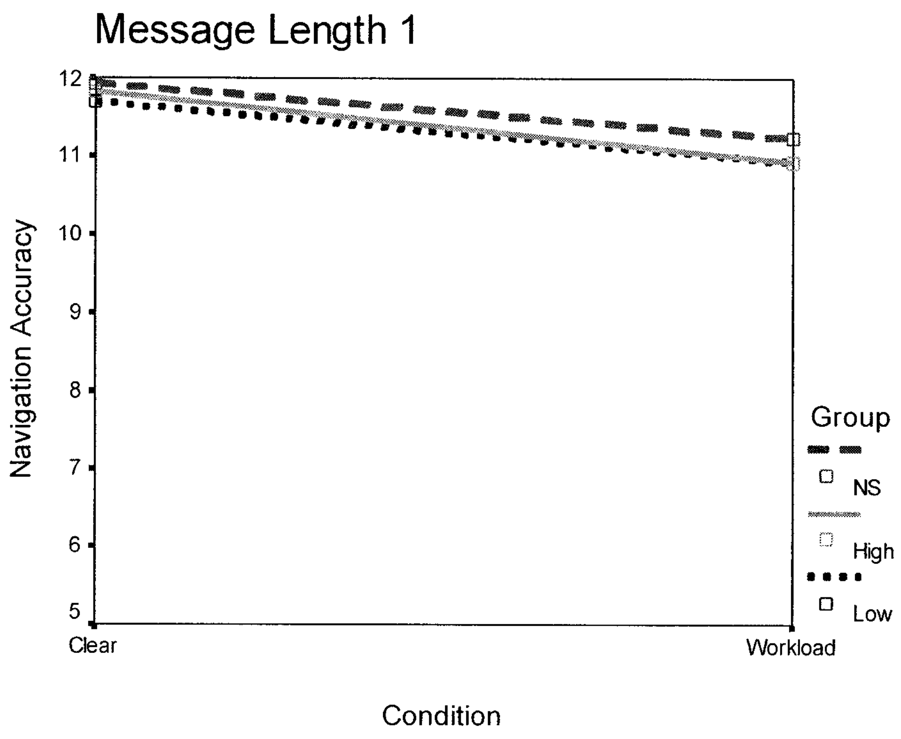


Figure 3. Navigation accuracy scores for message length 1 in the clear and workload conditions.

Message length 2. The ANCOVA comparing navigation accuracy data for message length 2 yielded significant main effects of condition, $F(1, 56) = 13.14, p = .001$, and proficiency, $F(1, 56) = 6.33, p = .003$, and a significant condition × addition task interaction, $F(2, 56) = 1.88, p = .02$. Pairwise comparisons carried out to explore these significant effects revealed that the High and Low groups navigated significantly less accurately in the workload condition than in the clear condition ($p \leq .004$). By contrast, the NS group navigated with equal accuracy in both conditions. The effects of workload for message length 2 are illustrated in Figure 4.

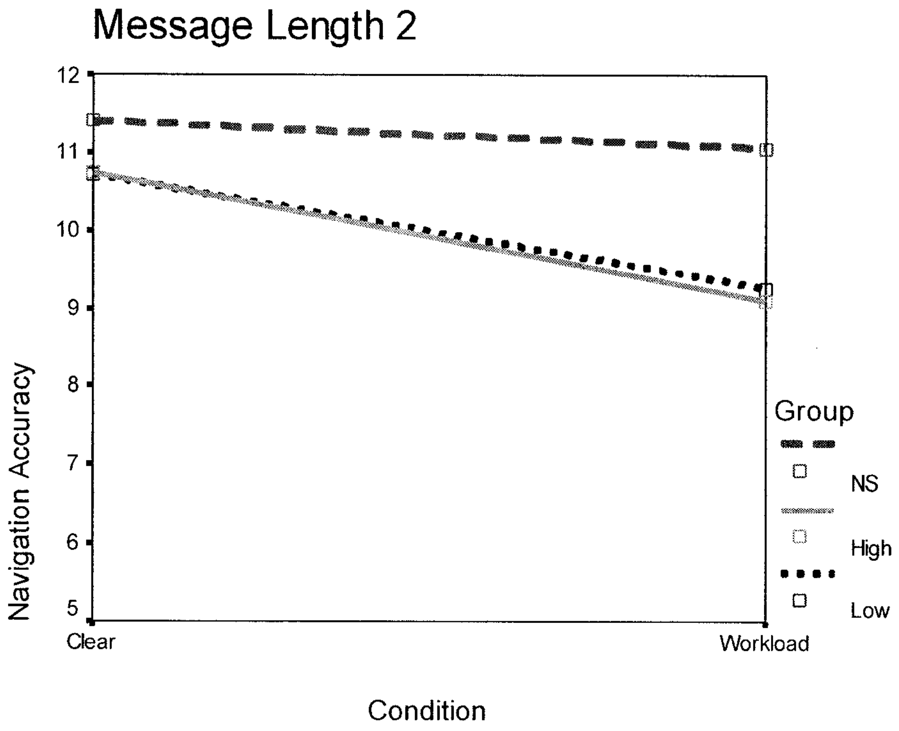


Figure 4. Navigation accuracy scores for message length 2 in the clear and workload conditions.

Message length 3. The ANCOVA comparing the navigation accuracy data for message length 3 yielded significant main effects of condition, *F*(1, 56) = 5.01, *p* = .029, and proficiency, *F*(2, 56) = 5.28, *p* = .008. Pairwise comparisons carried out to explore the effect of condition revealed that the NS and Low groups navigated significantly less accurately in the workload than in the clear condition (*p* ≤ .013). By contrast, the High group navigated with equal accuracy in both conditions. The effects of workload for message length 3 are illustrated in Figure 5.

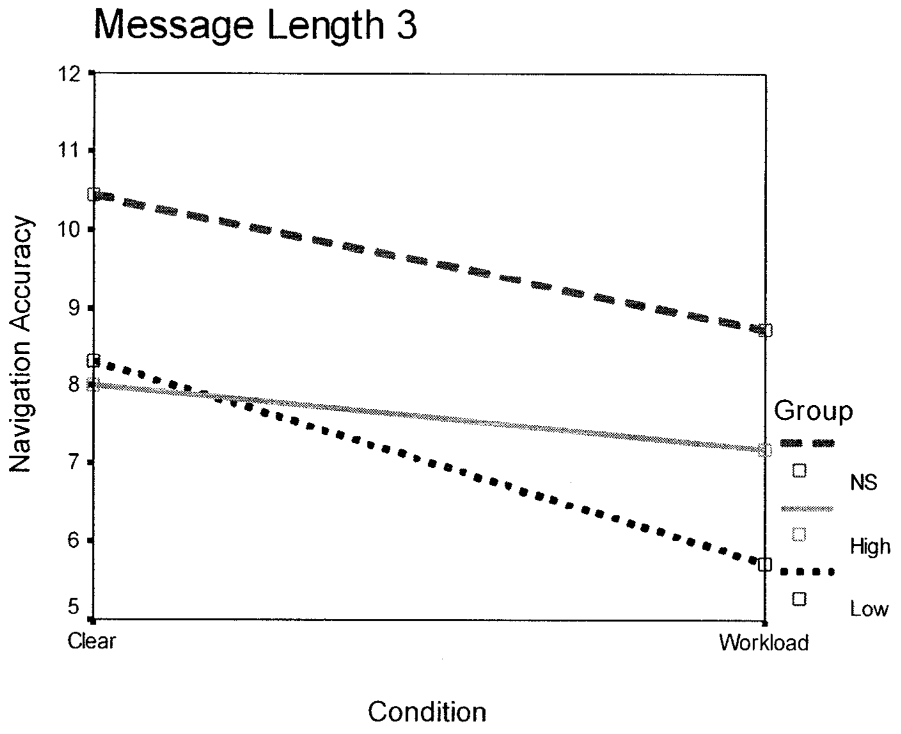


Figure 5. Navigation accuracy scores for message length 3 in the clear and workload conditions.

Workload effects. The extent to which cognitive workload affects navigation accuracy can be expressed as the mean difference between the clear and workload conditions. This simple subtraction analysis was conducted in order to determine which group displayed the *greatest* effect of workload in cases where there were significant effects of workload for more than one group. This measure (workload effect) is summarized in Table 9 for messages of different lengths. Workload effect was defined as a mean difference in navigation accuracy between the participants' performance in the workload condition and their performance in the clear condition. A higher value indicates a greater effect of workload and, therefore, a greater performance detriment as a function of increased workload. As can be seen in Table 9, the greatest effect of workload was observed in the Low group for messages of length 3.

Table 9

Effects of workload on navigation accuracy for messages of different lengths

Proficiency Groups	Length		
	1	2	3
NS (*n* = 20)	.70	.35	1.70
High (*n* = 20)	.91	1.64	.81
Low (*n* = 20)	.78	1.46	2.60

Note. Means for the workload condition have been adjusted with the one-minute addition task scores entered as a covariate.

Summary. Analyses of the navigation accuracy data revealed that navigation was less accurate in the workload condition than in the clear condition for all groups. For the NS group, there was an effect of workload for length 3 only. For the High group, there was an effect of workload for lengths 1 and 2. For the Low group, there was an effect of workload for all message lengths. The greatest detriment to navigation accuracy due to workload was observed in the Low group for messages of length 3.

Repetition Accuracy

Overall Analysis

The repetition accuracy data were submitted to a three-way repeated-measures ANOVA in order to determine if the three groups differed in repetition accuracy of the critical words in the message and if such differences interacted with workload condition and message length. In this analysis, proficiency served as a between-subjects factor (NS, High, Low); condition (clear, workload) and length (1, 2, 3) served as within-subjects factors. This analysis yielded significant main effects of proficiency, $F(2, 57) = 9.11, p < .001$, condition, $F(1, 57) = 13.19, p = .001$, and length, $F(2, 114) = 173.55, p < .001$. This analysis also yielded significant interaction effects of length × proficiency, $F(4, 114)$

= 4.24, p = .003, and of condition × length, F(2, 114) = 10.86, p < .001. Mean repetition accuracy scores are displayed in Table 10.

Table 10

Mean repetition accuracy scores (out of 12) and their standard deviations for the three proficiency groups in the clear and workload conditions

Condition	Length	Proficiency Groups					
		NS (n = 20)		High (n = 20)		Low (n = 20)	
		M	*SD*	*M*	*SD*	*M*	*SD*
Clear	1	12.00	.00	11.90	.31	11.85	.49
	2	11.85	.37	10.70	1.59	10.90	1.20
	3	9.50	2.69	8.05	2.72	8.13	2.05
Workload	1	11.95	.22	11.89	.31	12.00	.32
	2	11.80	.41	10.28	2.07	10.15	1.73
	3	8.75	2.67	6.78	2.95	5.65	2.72

Message Length

Clear condition. The two-way (proficiency × length) ANOVA comparing repetition accuracy data in the clear condition yielded significant main effects of length, F(2, 114) = 87.26, p < .001, and proficiency, F(2, 114) = 4.26, p = .019. Pairwise comparisons (α = .006) carried out to explore these effects yielded two findings. First, repetition accuracy depended on participants' proficiency only for message length 2, for which the NS group repeated messages more accurately than the High group. Second, repetition accuracy depended on message length. The NS group was less accurate in repeating messages of length 3 than messages of lengths 1 and 2 (p < .001). By contrast, both the High and Low groups displayed a significant decrease in repetition accuracy with each increasing message length (p < .001). Repetition accuracy scores for each message length in the clear condition are illustrated in Figure 6.

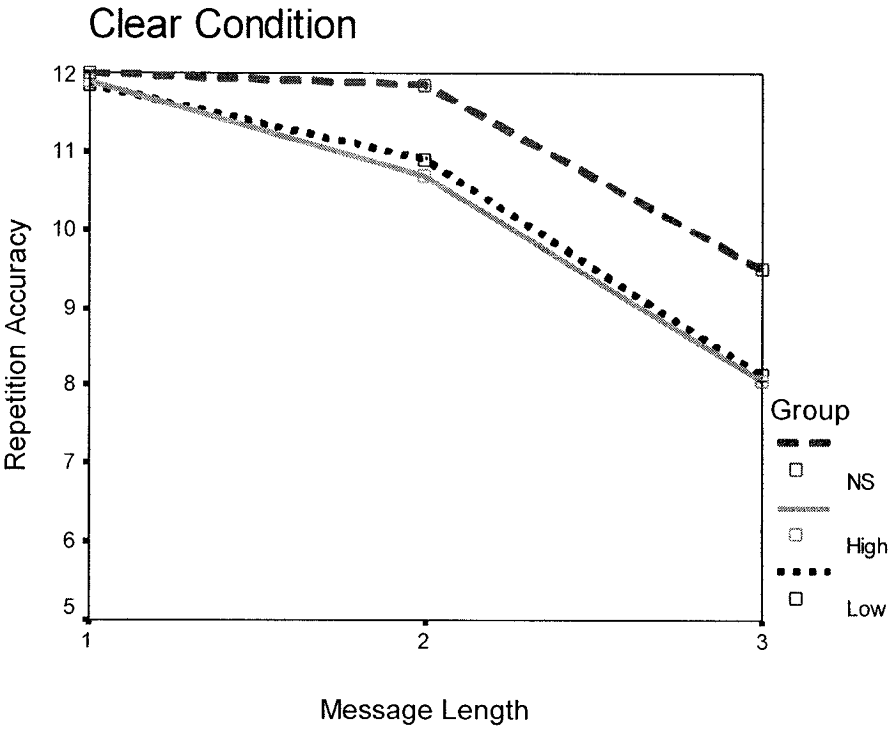


Figure 6. Repetition accuracy scores for each participant group (NS, High, Low) at each message length.

Workload condition. The two-way (proficiency × length) ANCOVA comparing repetition accuracy data in the workload condition yielded significant main effects of proficiency, $F(2, 56) = 15.69, p < .001$, the addition task, $F(1, 56) = 14.77, p < .001$, and length, $F(2, 112) = 42.41, p < .001$. This analysis also yielded significant length × proficiency, $F(4, 112) = 7.51, p < .001$, and length × addition task, $F(2, 112) = 7.12, p = .001$, interactions. Pairwise comparisons carried out to explore these significant effects yielded two findings. First, repetition accuracy depended on participants' proficiency at message lengths 2 and 3. At message length 1, all groups repeated messages with equal accuracy. For messages of lengths 2 and 3, the NS group repeated messages more accurately than the High and Low groups ($p \leq .001$). Second, repetition accuracy depended on message length. The NS group repeated messages of length 3 less accurately than messages of lengths 1 and 2 ($p < .001$). By contrast, the High and Low groups displayed a significant decline in repetition accuracy with each increasing message length. Repetition accuracy scores for each message length in the clear condition are illustrated in Figure 7.

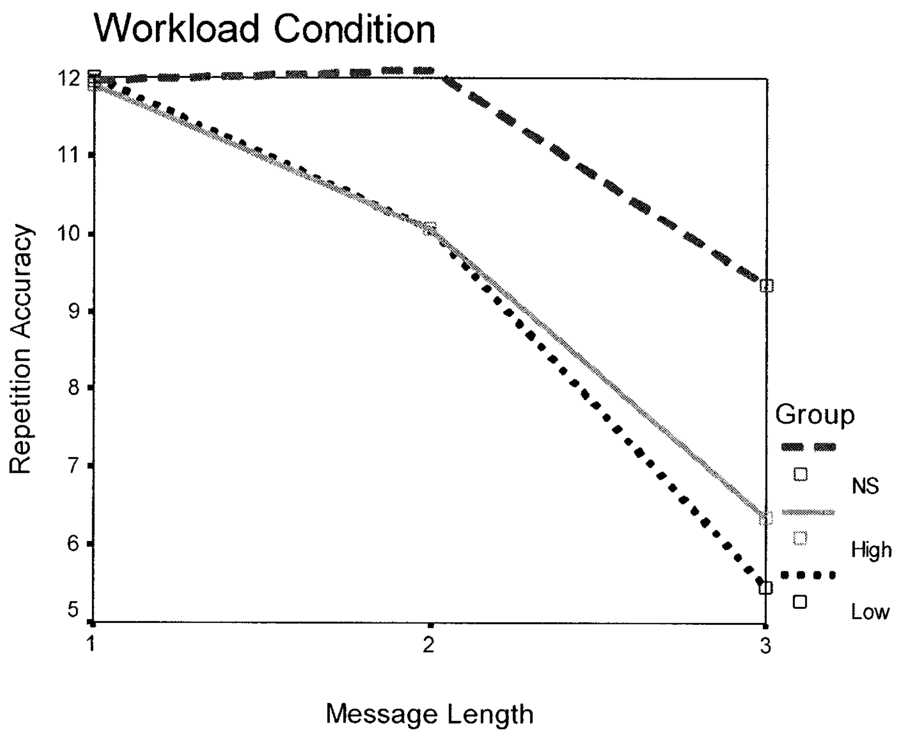


Figure 7. Repetition accuracy scores for each message length in the workload condition

Message length effects. Message length effects, representing mean differences in repetition accuracy for messages of two consecutive lengths, are summarized in Table 11 for both the clear and workload conditions. As with the analyses for navigation accuracy, a higher value indicates a greater effect of message length. In both conditions, the greatest effect of message length was observed in the Low group between messages of length 2 and 3.

Table 11

Effects of Message Length on Repetition Accuracy for the Three Proficiency Groups in the Clear and Workload Conditions

Proficiency Groups	Condition			
	Clear		Workload	
	1-2	2-3	1-2	2-3
NS (*n* = 20)	.15	2.35	-.14	2.76
High (*n* = 20)	1.20	2.65	1.82	3.70
Low (*n* = 20)	.95	2.77	1.94	4.59

Note. Means for the workload condition have been adjusted with the one-minute addition task scores entered as a covariate.

Summary. Analysis of the repetition accuracy data revealed that the length of the message did affect participants' repetition accuracy. In both conditions (clear, workload), longer messages resulted in less accurate repetitions for all groups; however, this decline in repetition accuracy depended on proficiency and workload. The NS group was less accurate repeating messages of length 3 in both conditions. By contrast, the High and Low groups repeated messages less accurately with each increasing length in both conditions. In other words, in both conditions, while the NS group displayed a significant decline in repetition accuracy for messages in length 3 in comparison to shorter messages, the L2 groups displayed a significant decline for messages of lengths 2 and 3 in comparison to messages of length 1. Overall, the Low group was the most affected by message length, demonstrating the greatest drop in repetition accuracy between message lengths 2 and 3.

Workload

Message length 1. The two-way (proficiency ×condition) ANCOVA comparing the repetition accuracy data for message length 1 yielded no significant effects. This finding suggested that repetition accuracy did not differ as a function of workload condition for any of the groups for messages of length 1. The effects of workload for message length 1 are illustrated in Figure 8.

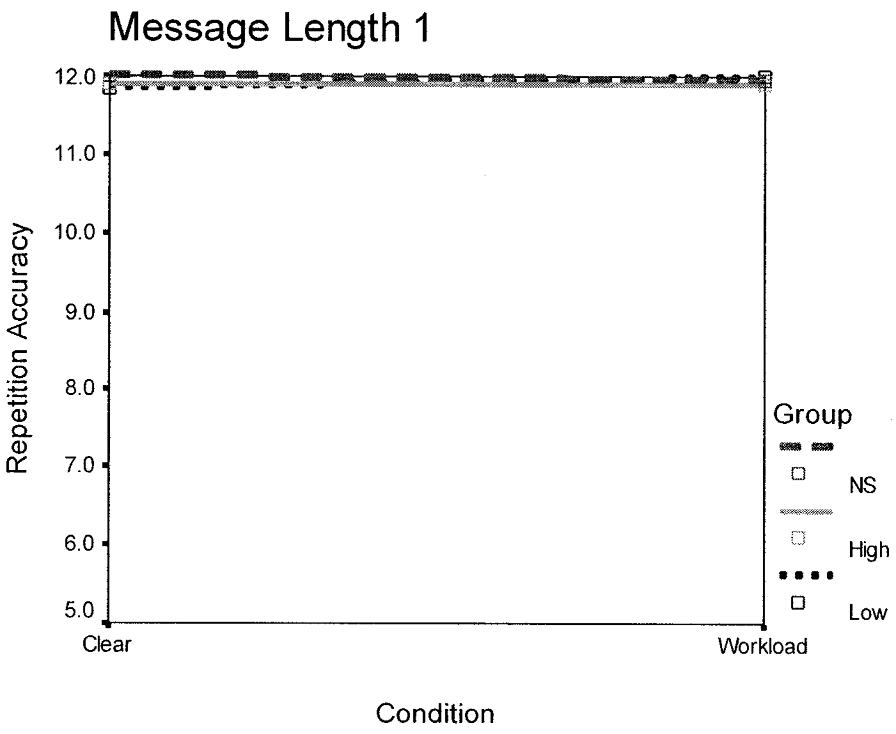


Figure 8. Repetition accuracy scores for message length 1 in the clear and workload conditions.

Message length 2. The two-way (proficiency ×condition) ANCOVA comparing the repetition accuracy data for message length 2 yielded significant effects of condition, $F(1, 56) = 6.08, p = .017$, proficiency, $F(2, 56) = 13.29, p < .001$, and the addition task, $F(1, 56) = 6.70, p = .012$. Pairwise comparisons ($\alpha = .017$) carried out to explore the effect of condition revealed that there were no significant differences in the repetition accuracy in the clear versus the workload condition for any of the groups. It is noteworthy, however, that the effect of workload approached significance for the Low group ($p = .029$). The effects of workload for message length 2 are illustrated in Figure 9.

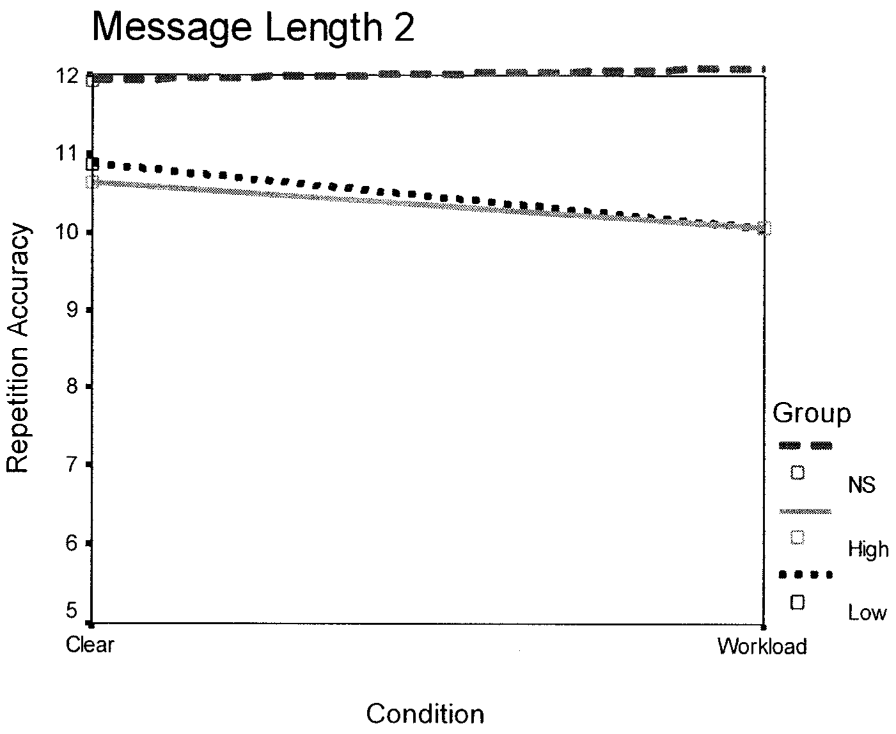


Figure 9. Repetition accuracy scores for message length 2 in the clear and workload conditions.

Message length 3. The two-way (proficiency ×condition) ANCOVA comparing the repetition accuracy data for message length 3 yielded a significant effect of condition, $F(1, 56) = 8.68, p = .005$, proficiency, $F(2, 56) = 9.63, p < .001$, and the addition task, $F(1, 56) = 8.44, p = .005$. Pairwise comparisons ($\alpha = .017$) carried out to explore the effect of condition revealed that the Low group repeated messages of length 3 significantly less accurately in the workload condition than in the clear condition ($p < .001$). It is noteworthy that the effect of workload approached significance for the High group ($p = .03$). The effects of workload for message length 3 are illustrated in Figure 10.

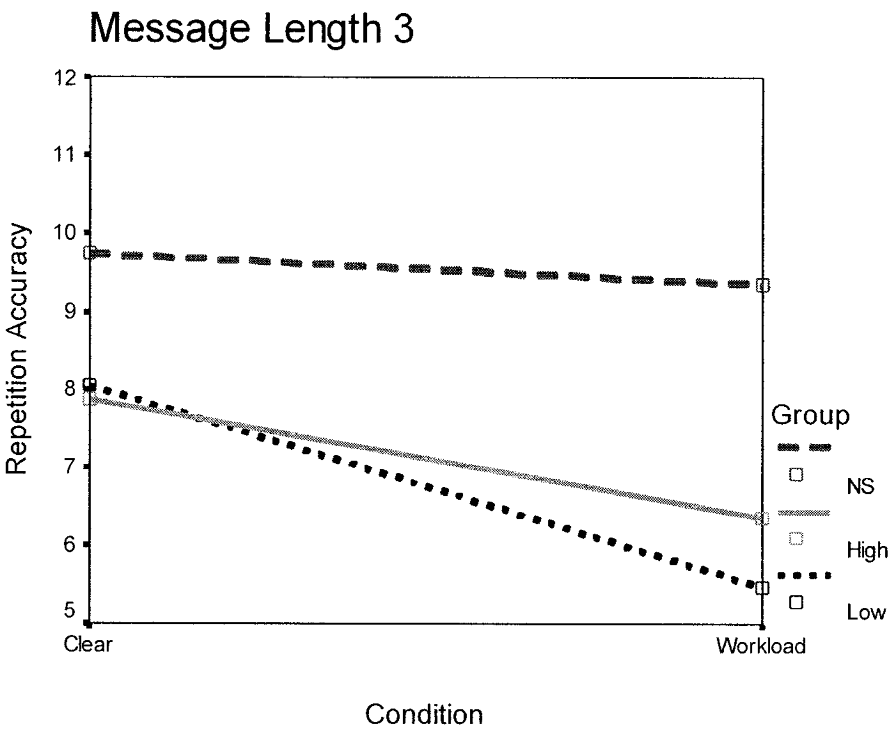


Figure 10. Repetition accuracy scores for message length 3 in the clear and workload conditions.

Workload effects. Workload effects, representing mean differences in repetition accuracy between the workload and the clear conditions, are summarized in Table 12 for messages of different lengths. This table illustrates that the greatest effect of workload, that is, the greatest difference in accuracy between the clear and workload conditions, was observed for the Low group at message length 3.

Table 12

Effect of Workload on Repetition Accuracy for Messages of Different Lengths

Proficiency Groups	Length		
	1	2	3
NS (*n* = 20)	.06	-.17	.38
High (*n* = 20)	.01	.57	1.53
Low (*n* = 20)	-.15	.82	2.59

Note. Means for the workload condition have been adjusted with the one-minute addition task scores entered as a covariate.

Summary. Analyses of the repetition accuracy data revealed that there was a significant effect of workload for the Low group only for messages of length 3. The effect of workload approached significance for the Low group for messages of length 2, and for the High group for messages of length 3. The low group appeared the most affected by workload effects; this group showed the most detriment in repetition accuracy due to workload.

Chapter Summary

Message Length

The length of the message did affect all groups' navigation and repetition accuracy. As hypothesized, the effects of message length differed as a function of proficiency and workload; these effects were similar for navigation and repetition accuracy. In the clear condition, for both navigation and repetition accuracy, all groups performed with about equal accuracy at each message length (with the exception of message length 2 for repetition accuracy where the NS group was more accurate than the High group). By contrast, in the workload condition, the NS group was more accurate than the L2 groups for both navigation and repetition accuracy. More specifically, the NS group was significantly more accurate than the Low group when navigating and repeating messages of lengths 2 and 3. The NS group was significantly more accurate than the High group when navigating and repeating messages of length 2, and when repeating messages of length 3. The two L2 groups did not differ significantly in navigation or repetition accuracy for any message length in either condition. Therefore, message length affected not only the Low group, as hypothesized, but also the High group. For both navigation and repetition of messages, the greatest drop in accuracy due to message length occurred in the Low group in the workload condition between messages of length 2 and 3.

Workload

Overall significant effects of workload were obtained for all groups for both navigation and repetition accuracy. As hypothesized, the effects of workload depended on proficiency and message length. These effects differed for navigation accuracy and repetition accuracy.

For navigation accuracy, participants' overall performance was significantly more accurate in the clear than in the workload condition for all message lengths. However, these effects of workload differed between the participant groups. The NS group was affected by workload at message length 3 only. By contrast, the Low group was affected by workload at *all* message lengths, and the High group was affected by workload at message lengths 1 and 2. In terms of navigation accuracy, therefore, both L2 groups (High and Low) were affected by workload, with the greatest drop in navigation accuracy due to additional cognitive workload observed in the Low group at message length 3.

For repetition accuracy, a significant effect of workload was found only for the Low group and only for messages of length 3. For messages of length 3, the Low group's repetitions were significantly less accurate in the workload than in the clear condition.

Therefore, as hypothesized, the greatest effect of workload on repetition accuracy was observed in the Low group.

CHAPTER 5

RESULTS II

Chapter Overview

This chapter summarizes the results of between-group statistical analyses of the speech production measures (accentedness, comprehensibility, fluency and confidence). First, the objectives and hypotheses related to these dependent variables are recapitulated. Next, results of statistical analyses are presented. For each of the dependent variables, a summary of findings is provided at the end of each section (message length and workload). Findings related to each set of hypotheses (message length and workload) are summarized at the end of the chapter.

Main Objective and Hypotheses

One objective of the present study was to determine whether and to what extent the speech production of high- and low-proficiency L2 speakers was affected by high workload involving L2 communication. This objective was accomplished by analyzing speech samples obtained in a simulated pilot navigation task. The speech samples were rated by native-English speakers based on four criteria: accentedness, comprehensibility, fluency and confidence. Two hypotheses were proposed. The first hypothesis was that the length of the message would affect L2 participants' speech production in a simulated navigation task in that longer messages would result in a detriment to speech production on all measures. The prediction here was that the effects of message length would differ for all groups and the greatest detriment to speech production due to increased message length would occur in the Low group. The second hypothesis was that increased cognitive workload, that is, the addition of the concurrent arithmetic task, would adversely affect L2 participants' speech production. The prediction here was that the effects of increased cognitive workload would differ for all groups and the greatest detriment to speech production on all measures would occur in the Low group.

Individual Differences in Mental Arithmetic Ability

As discussed in the previous chapter, workload was confounded with the participants' ability to solve mental arithmetic problems; therefore, analyses of covariance (ANCOVAs) were used in all between-group comparisons exploring the effect of workload, with the participants' performance on the addition task entered as a covariate.

Accentedness

Overall Analysis

The accentedness data were submitted to a three-way repeated-measures ANOVA in order to determine if the speech of the three groups differed in accentedness and if such differences interacted with message length and workload condition. In this analysis, proficiency served as a between-subjects factor (NS, High, Low); message length (1, 2, 3) and condition (clear, workload) served as within-subjects factors. First, this analysis yielded a significant main effect of condition, $F(1, 57) = 8.17, p = .006$, suggesting that overall, participants' speech was more accented in the workload condition than in the clear condition. This analysis also yielded significant main effects of proficiency, $F(2, 57) = 350.87, p < .001$, and length, $F(2, 114) = 49.30, p < .001$, and a significant length × proficiency interaction, $F(4, 114) = 17.52, p < .001$. Mean accentedness ratings for each participant group are presented in Table 13. Higher ratings represent less accented speech.

Table 13

Mean Accentedness Ratings (out of 9) and their Standard Deviations for the Three Proficiency Groups in the Clear and Workload Conditions

Condition	Length	Proficiency Groups					
		NS (n = 20)		High (n = 20)		Low (n = 20)	
		M	*SD*	*M*	*SD*	*M*	*SD*
Clear	1	8.39	.35	4.16	1.31	3.27	.64
	2	8.40	.31	3.45	1.04	2.70	.43
	3	8.41	.50	3.46	1.19	2.65	.59
Workload	1	8.27	.49	4.22	1.31	3.03	.53
	2	8.33	.63	3.22	1.27	2.47	.32
	3	8.44	.41	3.39	1.16	2.58	2.71

Message Length

A series of analyses were carried out to explore the effect of message length on the participants' accentedness in each condition. For these analyses, the data for the clear condition were submitted to a two-way repeated-measures ANOVA with proficiency (NS, High, Low) as a between-subjects factor and message length (1, 2, 3) as a within-subjects factor. The data for the workload condition were submitted to a two-way repeated-measures ANCOVA with proficiency (NS, High, Low) as a between-subjects factor, message length (1, 2, 3) as a within-subjects factor, and the participants' scores on the addition task as a covariate.

Clear condition. The ANOVA comparing accentedness ratings in the clear condition yielded significant main effects of proficiency, $F(2, 57) = 338.15$, $p < .001$, and length $F(2, 114) = 25.20$, $p < .001$, and a significant length × proficiency interaction, $F(4, 114) = 6.77$, $p < .001$. Pairwise comparisons carried out to explore these significant effects ($\alpha = .006$) yielded two findings. First, accentedness ratings depended on participants' proficiency. For all message lengths, the Low group was more accented than the High group and the High group was more accented than the NS group ($p \leq .003$). Second, accentedness rating depended on message length. Both the High and Low groups sounded significantly more accented when repeating messages of lengths 2 and 3 than messages of length 1 ($p < .001$). As expected, the NS group's speech was equally non-accented at all message lengths. Mean accentedness ratings for each group are illustrated in Figure 11 (higher ratings represent less accented speech).

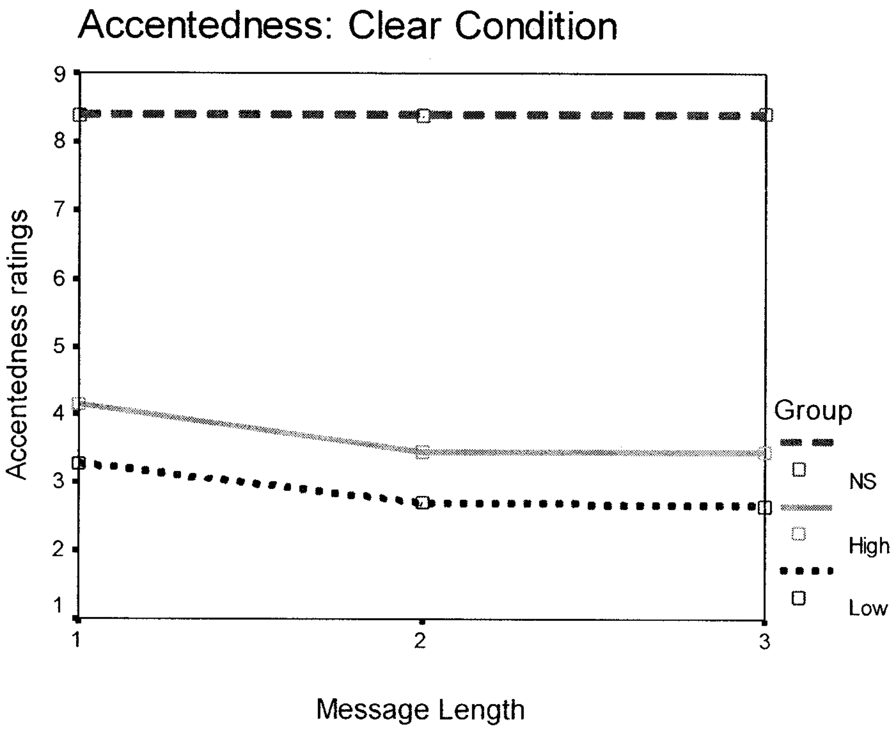


Figure 11. Mean accentedness ratings in the clear condition for each participant group (NS, High, Low) for each message length.

Workload Condition. The ANCOVA comparing accentedness ratings in the workload condition yielded a significant main effect of proficiency, $F(2, 56) = 307.73, p < .001$, and a significant length × proficiency interaction, $F(4, 112) = 8.54, p < .001$. Pairwise comparisons carried out to explore these significant effects ($\alpha = .006$) yielded two findings. First, the Low group was significantly more accented than the High group at all message lengths ($p \leq .001$), with the exception of message length 2 where the difference approached significance ($p = .008$). Second, both the High and Low groups sounded more accented when they repeated messages of lengths 2 and 3 than messages of length 1 ($p \leq .006$). As in the clear condition, the NS group received equal accentedness ratings at all message lengths. Mean accentedness ratings for each group are illustrated in Figure 12.

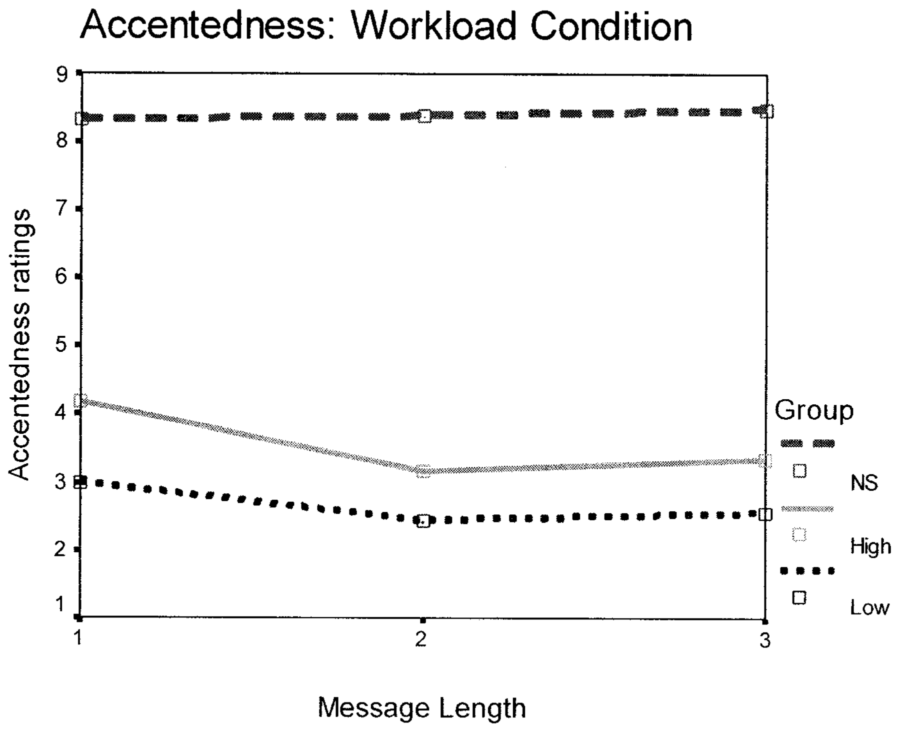


Figure 12. Mean accentedness ratings in the workload condition for each participant group (NS, High, Low) for each message length.

Message length effects. The extent to which message length affected accentedness ratings can also be expressed as a mean difference in accentedness ratings for messages of each consecutive length. This measure (message length effect) indicates which group displayed the greatest detriment in accentedness due to increased message length. Message length effects are summarized in Table 14 for the clear and workload conditions.

Message length effect was defined as a mean difference in accentedness between the participants' performance on a shorter message and their performance on a subsequent longer message. For each condition, two values of message length effects were calculated: one for the difference in accentedness between message length 1 and 2, the other for the difference in accentedness between message length 2 and 3. A higher value indicates a greater effect of message length and, therefore, a greater speech production detriment as a function of increasing message length. In both the clear and workload conditions, the greatest effect of message length was observed in the High group between messages of length 1 and 2.

Table 14

Message Length Effects in the Clear and Workload Conditions

Proficiency Groups	Condition			
	Clear		Workload	
	1-2	2-3	1-2	2-3
NS (*n* = 20)	-.00	-.02	-.05	-.09
High (*n* = 20)	.70	-.00	1.00	-.18
Low (*n* = 20)	.57	.05	.56	-.11

Note. Means for the workload condition have been adjusted with the one-minute addition task scores entered as a covariate.

Message length summary. Analysis of the accentedness data revealed that the length of the message did affect the accentedness of speech for L2 participants. As expected, the NS group did not sound more accented when repeating messages of increasing lengths. This was not the case for the L2 participants, however. In the clear and workload conditions, the Low group sounded more accented than the High group, and both groups sounded more accented when repeating messages of lengths 2 and 3 than messages of length 1. In both the clear and workload conditions, the greatest effect of message length, that is, the greatest detriment to accentedness due to increased message length, was observed for the High group between messages of lengths 1 and 2.

Workload

A series of analyses were carried out to explore the effect of workload on the participants' accentedness for each message length. For these analyses, the data for each message length (1, 2, 3) were submitted to three separate two-way repeated-measures ANCOVAs with proficiency (NS, High, Low) as a between-subjects factor, condition (clear, workload) as a within-subjects factor, and the participants' score on the addition task as a covariate.

Message length 1. The ANCOVA comparing the accentedness data for message length 1 yielded only a significant effect of proficiency, $F(2, 56) = 209.45, p < .0001$. This finding suggested that accentedness did not differ as a function of workload condition for any of the groups for messages of length 1.

Message length 2. The ANCOVA comparing the accentedness data for message length 2 yielded significant effects of proficiency, $F(2, 56) = 340.13, p < .0001$, and condition, $F(1, 56) = 5.49, p = .023$. Pairwise comparisons carried out to explore the effect of condition revealed that the High and Low groups sounded more accented in the workload condition than in the clear condition; however, these differences did not reach statistical significance for either group ($p \leq .029$). The effects of workload for message length 2 are illustrated in Figure 13.

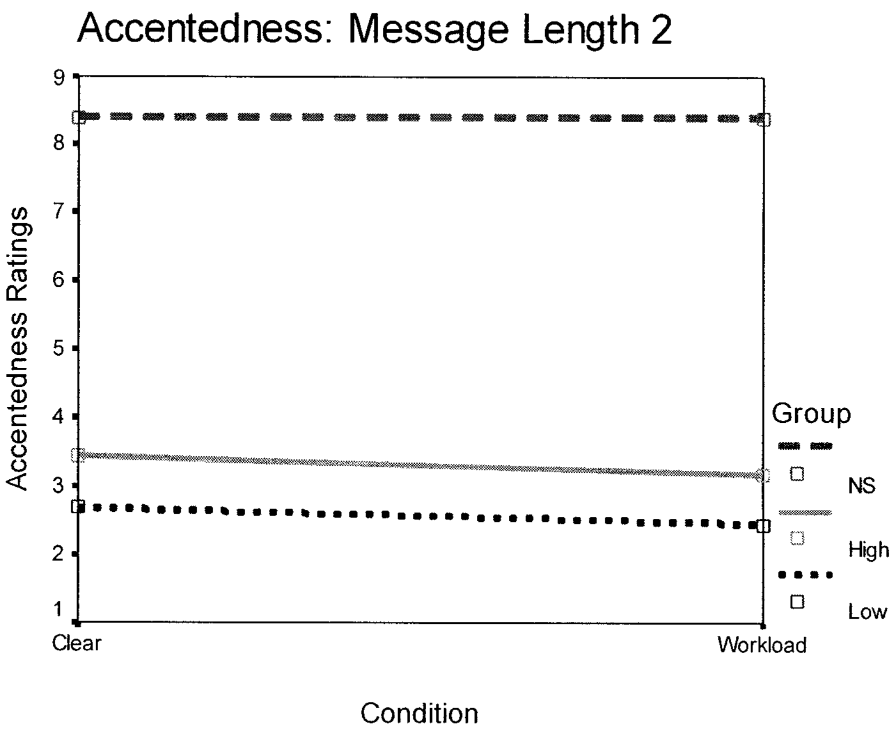


Figure 13. Accentedness ratings for message length 2 in the clear and workload conditions.

Message length 3. The ANCOVA comparing accentedness data for message length 3 yielded only a significant effect of proficiency, $F(2, 56) = 332.77, p < .001$. Thus, accentedness did not differ as a function of workload condition for any of the groups for messages of length 3.

Workload effects. The extent to which workload affects accentedness can also be expressed as a mean difference in accentedness between clear and workload conditions. This measure (workload effect) determines which group displayed the greatest detriment to accentedness due to increased cognitive workload. Workload effect was defined as a mean difference in accentedness ratings between the participants' accentedness scores in the clear condition and their accentedness scores in the workload condition. A higher value indicates a greater effect of workload and, therefore, a greater detriment to speech production as a function of increased workload. As can be seen in Table 15, the greatest effect of workload was observed in the High group for messages of length 2.

Table 15

Effect of Workload on Accentedness Ratings for Messages of Different Lengths

Proficiency Groups	Length		
	1	2	3
NS (*n* = 20)	.10	.01	.03
High (*n* = 20)	.05	.27	.07
Low (*n* = 20)	.25	.24	.08

Note. Means for the workload condition have been adjusted with the one-minute addition task scores entered as a covariate.

Workload summary. There was no significant effect of workload for any of the groups. However, for messages of length 2, the High and Low groups demonstrated a tendency to sound more accented in the workload than in the clear condition, although this difference did not reach statistical significance. The greatest detriment to accentedness due to increased cognitive workload was observed in the High group for messages of length 2.

Comprehensibility

Overall Analysis

The comprehensibility data were submitted to a three-way repeated-measures ANOVA in order to determine if the three groups differed in comprehensibility and if such differences interacted with message length and workload condition. As in the previous overall analyses, proficiency served as a between-subjects factor (NS, High, Low); message length (1, 2, 3) and condition (clear, workload) served as within-subjects factors. This analysis yielded a significant effect of condition, $F(1, 57) = 7.65, p = .008$, suggesting that overall participants were significantly more comprehensible in the clear condition than in the workload condition. This analysis also yielded significant main effects of proficiency, *F*(2, 57) = 251.59, *p* < .001, and length, *F*(2, 114) = 102.20, *p* < .001 and a significant length × proficiency interaction, *F*(4,114) = 19.78, *p* < .001. Mean comprehensibility ratings for each participant group are presented in Table 16. Higher ratings represent more comprehensible speech.

Table 16

Mean Comprehensibility Ratings (out of 9) and their Standard Deviations for the Three Proficiency Groups in the Clear and Workload Conditions

Condition	Length	Proficiency Groups					
		NS (*n* = 20)		High (*n* = 20)		Low (*n* = 20)	
		M	*SD*	*M*	*SD*	*M*	*SD*
Clear	1	8.43	.25	5.53	.94	4.42	.78
	2	8.37	.32	4.39	1.13	3.64	.91
	3	8.30	.48	4.37	1.20	3.18	.94
Workload	1	8.38	.37	5.44	1.09	4.21	.92
	2	8.27	.60	4.21	1.12	3.29	.69
	3	8.24	.55	4.17	.99	3.13	.57

Message Length

Clear condition. The ANOVA comparing comprehensibility ratings in the clear condition yielded significant main effects of proficiency, *F*(2, 57) = 205.83, *p* < .001 and length, *F*(2, 114) = 61.36, *p* < .001, and a significant length × proficiency interaction, *F*(4, 114) = 12.62, *p* < .001. Pairwise comparisons carried out to explore these significant effects (α = .006) revealed two findings. First, comprehensibility ratings depended on participants' proficiency. For all message lengths, the NS group was more comprehensible than the High and Low groups (*p* < .001). The High group was more comprehensible than the Low group at message lengths 1 and 3 (*p* < .001); at message length 2 this difference did not reach statistical significance (*p* = .007). Second, comprehensibility ratings depended on message length, but only for L2 participants. The High group's speech was significantly more comprehensible at message length 1 than at message lengths 2 and 3 ($p < .001$). The Low group's comprehensibility decreased significantly with each increasing message length ($p < .001$). Comprehensibility ratings in the clear condition are illustrated in Figure 14 (again, higher ratings represent more comprehensible speech).

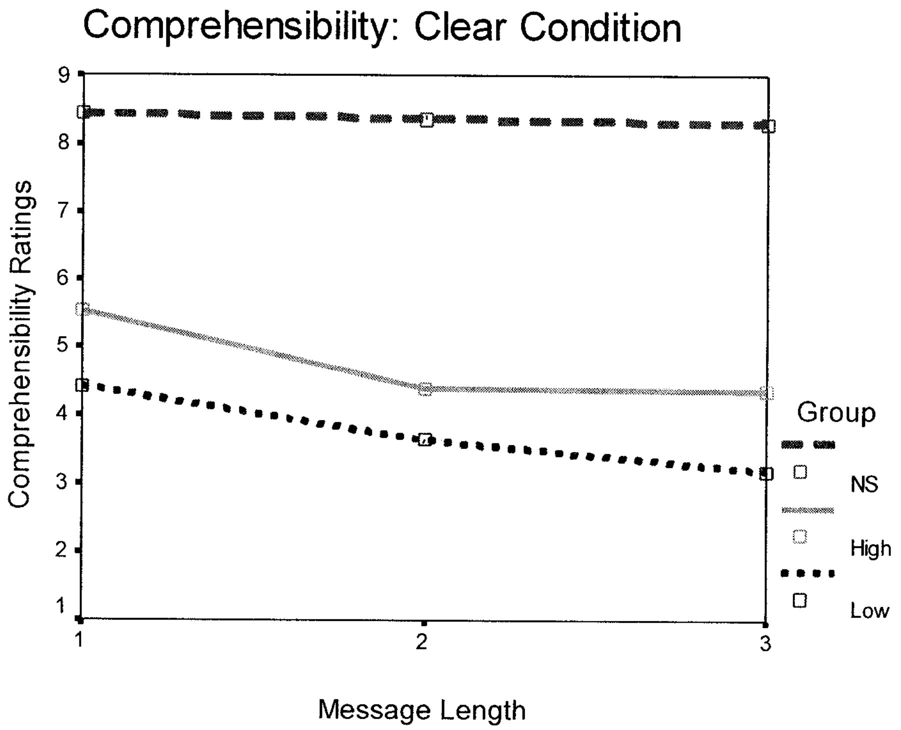


Figure 14. Comprehensibility ratings in the clear condition for each participant group (NS, High, Low) for each message length.

Workload Condition. The ANCOVA comparing comprehensibility ratings in the workload condition yielded a significant main effect of proficiency, $F(2, 56) = 252.62, p < .0001$, and a significant length × proficiency interaction, $F(4, 112) = 6.54, p < .0001$. Pairwise comparisons carried out to explore these significant effects ($\alpha = .006$) yielded two findings. First, at all message lengths, the NS group was more comprehensible than the High group and the High group was more comprehensible than the Low group ($p \leq .001$). Second, comprehensibility ratings depended on message length, but only for L2 participants. Both the High and Low groups were more comprehensible for messages of length 1 than for messages of lengths 2 and 3. Comprehensibility ratings for each group in the workload condition are illustrated in Figure 15.

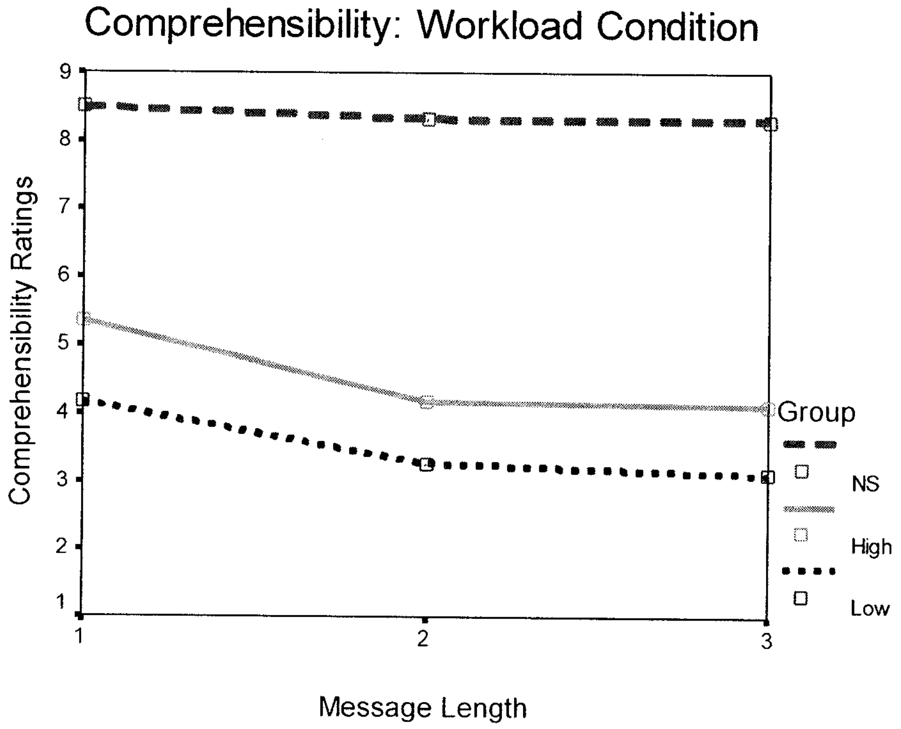


Figure 15. Comprehensibility ratings in the workload condition for each participant group (NS, High, Low) for each message length.

Message length effects. In order to determine which group displayed the greatest detriment to comprehensibility due to increased message length, the effects of message length are expressed as the mean difference in comprehensibility ratings between messages of each consecutive length. As can be seen in Table 17, the greatest effect of message length was observed in the High group between messages of lengths 1 and 2.

Table 17

Message Length Effects in the Clear and Workload Conditions

Proficiency Groups	Condition			
	Clear		Workload	
	1-2	2-3	1-2	2-3
NS (*n* = 20)	.07	.07	.16	.02
High (*n* = 20)	1.14	.03	1.19	.05
Low (*n* = 20)	.79	.45	.90	.16

Note. Means for the workload condition have been adjusted with the one-minute addition task scores entered as a covariate.

Message length summary. Analysis of the comprehensibility data revealed that the length of the message did affect the comprehensibility of speech for L2 participants. In both the clear and workload conditions, the NS group was significantly more comprehensible than the High group, and the High group was significantly more comprehensible than the Low group. In the clear condition, the High group sounded less comprehensible when repeating messages of lengths 2 and 3 than messages of length 1, and the Low group sounded less comprehensible with each increasing message length. In the workload condition, both the High and Low groups sounded less comprehensible when repeating messages longer than one command. The greatest detriment to comprehensibility due to increased message length was observed in the High group between messages of lengths 1 and 2 in both the clear and workload conditions.

Workload

Message length 1. The ANCOVA comparing the comprehensibility data for message length 1 yielded a significant effect of proficiency, $F(2, 56) = 168.08, p < .0001$, and a significant effect of condition, $F(1, 56) = 4.47, p = .039$. Pairwise comparisons carried out to explore this significant effect of condition revealed that, although all groups tended to sound less comprehensible in the workload than in the clear condition, the effect of workload failed to reach significance for messages of length 1.

Message length 2. The ANCOVA comparing the comprehensibility data for message length 2 yielded a significant effect of proficiency, $F(2, 56) = 195.66, p < .0001$, and only a marginally significant effect of condition, $F(1, 56) = 3.29, p = .075$. This finding suggested that comprehensibility did not differ as a function of workload condition for any of the groups for messages of length 2.

Message length 3. The ANCOVA comparing comprehensibility data for message length 3 yielded only a significant effect of proficiency, $F(2, 56) = 232.85, p < .0001$. Thus, there were no significant differences in comprehensibility between the clear and workload conditions for any of the groups for this message length.

Workload effects. The effects of workload are expressed as the mean difference in comprehensibility ratings between the clear and workload conditions. A higher value indicates a greater effect of workload and, therefore, a greater detriment to comprehensibility as a function of increased workload. As can be seen in Table 18, the greatest effect of workload was observed in the Low group for messages of length 2, although this effect was not significant.

Table 18.

Effect of Workload on Comprehensibility Ratings for Messages of Different Lengths

Proficiency Groups	Length		
	1	2	3
NS ($n = 20$)	.03	.05	.02
High ($n = 20$)	.15	.21	.22
Low ($n = 20$)	.24	.36	.07

Note. Means for the workload condition have been adjusted with the one-minute addition task scores entered as a covariate.

Workload summary. There was no significant effect of workload on comprehensibility ratings for any of the groups.

Fluency

Overall Analysis

As in the previous overall analyses, the fluency data were submitted to a three-way repeated-measures ANOVA in order to determine if the three groups differed in fluency and if such differences interacted with workload condition and message length. This analysis yielded a main effect of condition, $F(1, 57) = 22.35, p < .001$, indicating that overall, participants were less fluent in the workload condition than in the clear condition. This analysis also yielded significant main effects of proficiency, $F(2, 57) = 279.07, p < .001$, and length, $F(2, 114) = 143.70, p < .001$, and a significant length × proficiency interaction, $F(4, 114) = 16.14, p < .001$. Mean fluency ratings for each participant group are presented in Table 19. Higher ratings represent more fluent speech.

Table 19.

Mean Fluency Ratings (out of 9) and Their Standard Deviations for the Three Proficiency Groups in the Clear and Workload Conditions

Condition	Length	Proficiency Groups					
		NS (*n* = 20)		High (*n* = 20)		Low (*n* = 20)	
		M	*SD*	*M*	*SD*	*M*	*SD*
Clear	1	8.48	.32	5.84	.81	5.06	.60
	2	8.43	.34	4.69	.93	4.19	.76
	3	8.17	.57	4.12	1.20	3.40	1.10
Workload	1	8.32	.51	5.82	.97	4.85	.90
	2	8.06	.71	4.40	1.04	3.58	.64
	3	7.65	.82	3.92	.78	3.29	.70

Message Length

Clear condition. The ANOVA comparing fluency ratings in the clear condition yielded significant main effects of proficiency, $F(2, 57) = 225.27, p < .001$ and length $F(2, 114) = 81.51, p < .001$, and a significant length × proficiency interaction $F(4, 114) = 12.36, p < .001$. Pairwise comparisons carried out to explore these significant effects ($\alpha = .006$) revealed two findings. First, fluency ratings depended on participants' proficiency. For all messages, the NS group was more fluent than both the High and the Low groups. The High group was more fluent than the Low group when repeating messages of length 1 only. Second, fluency ratings depended on message length, but only for L2 participants. Both the High and Low groups sounded significantly less fluent with each increasing message length ($p \leq .001$). Fluency ratings in the clear condition are illustrated in Figure 16 (again, higher ratings represent more fluent speech).

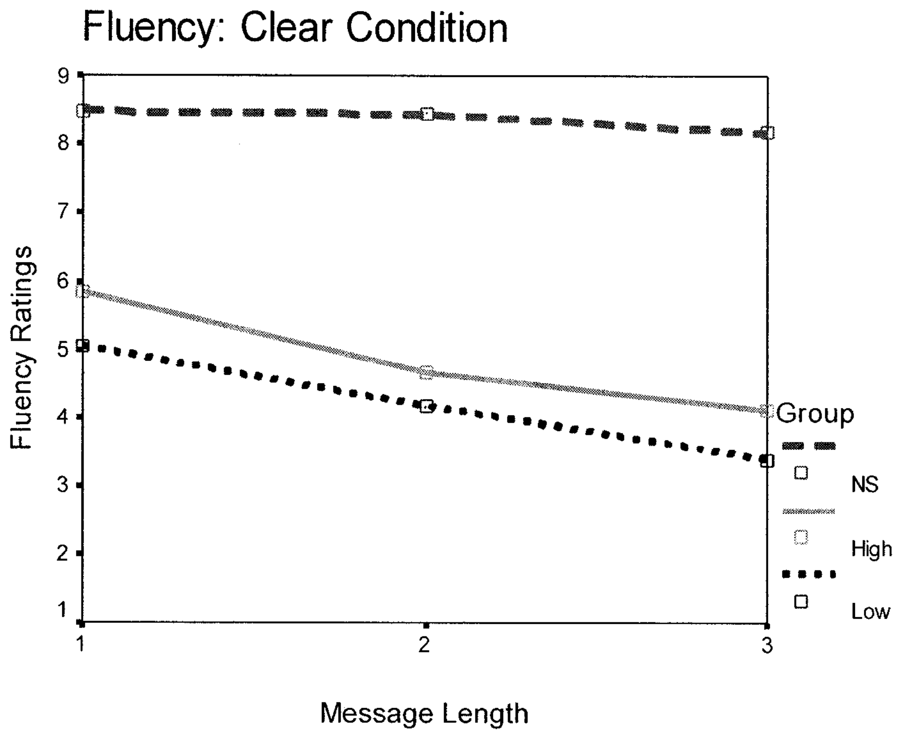


Figure 16. Fluency ratings in the clear condition for each participant group (NS, High, Low) for each message length.

Workload Condition. The ANCOVA comparing fluency ratings in the workload condition yielded a significant main effect of proficiency, $F(2, 56) = 260.54, p < .0001$, and a significant length × proficiency interaction, $F(4, 112) = 4.67, p = .002$. Pairwise comparisons carried out to explore these significant effects ($\alpha = .006$) yielded two findings. First, as in the clear condition, fluency ratings depended on participants' proficiency. The NS group was more fluent than the High and Low groups at all message lengths ($p \leq .001$). For messages of lengths 1 and 2, the High group was also more fluent than the Low group ($p \leq .003$). Second, fluency ratings also depended on message length. The NS group sounded less fluent at message length 3 ($p = .004$) than at message lengths 1 and 2. By contrast, the High and Low groups sounded less fluent at message lengths 2 and 3 than at message length 1 ($p < .001$). Fluency ratings in the workload condition are illustrated in Figure 17.

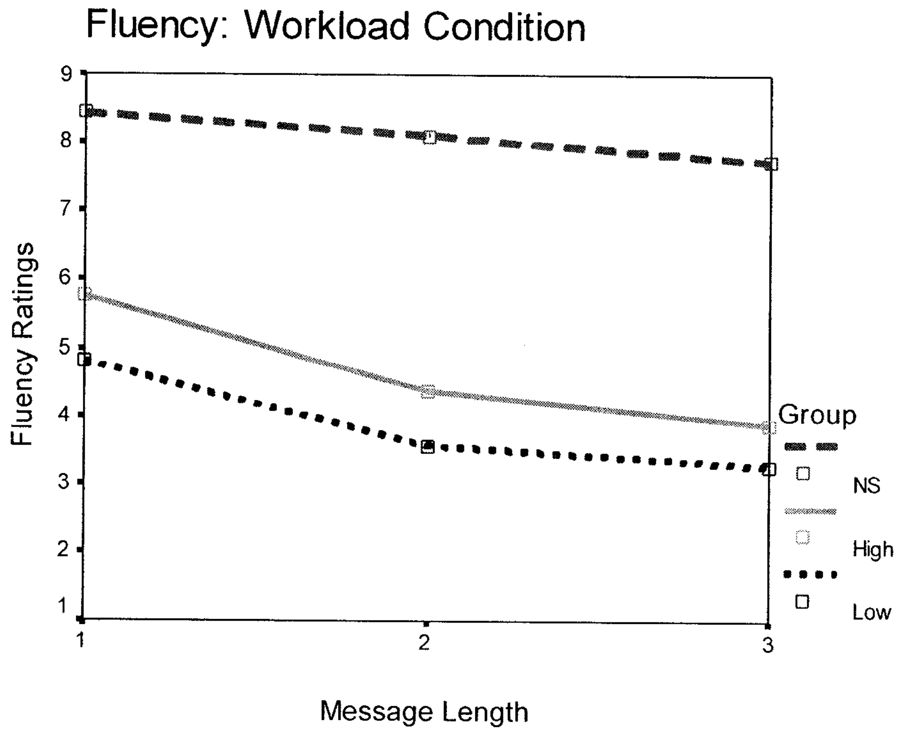


Figure 17. Fluency ratings in the workload condition for each participant group (NS, High, Low) for each message length.

Message length effects. In order to determine which group displayed the greatest detriment to fluency due to increased message length, the effects of message length are expressed as the mean difference in fluency ratings between messages of each consecutive length. As can be seen in Table 20, the greatest effect of message length was observed in the High group between messages of lengths 1 and 2.

Table 20

Effects of Message Length on Fluency Ratings

Proficiency Groups	Condition			
	Clear		Workload	
	1-2	2-3	1-2	2-3
NS ($n = 20$)	.05	.27	.33	.37
High ($n = 20$)	1.16	.57	1.37	.51
Low ($n = 20$)	.88	.79	1.25	.30

Note. Means for the workload condition have been adjusted with the one-minute addition task scores entered as a covariate.

Message length summary. Analysis of the fluency data revealed that fluency differed as a function of proficiency and message length. At all message lengths in both conditions, the NS group was more fluent than the High and Low groups. The High group was more fluent than the Low group at length 1 in the clear condition and at lengths 1 and 2 in the workload condition. All groups sounded less fluent with increasing message lengths, but these effects of message length differed as a function of proficiency and workload. The NS group was less fluent at message length 3 in the workload condition only. By contrast, the High and Low groups were less fluent with each increasing message length in the clear condition, and in the workload condition, these groups were less fluent repeating messages of length 2 and 3 than when repeating messages of length 1. The greatest detriment to fluency due to increased message length was observed for the High group between messages of lengths 1 and 2.

Workload

Message length 1. The ANCOVA comparing the fluency data for message length 1 yielded only a significant effect of proficiency, $F(2, 56) = 161.18, p < .001$. Thus, there were no significant differences in fluency between the clear and workload conditions for any of the groups at message length 1.

Message length 2. The ANCOVA comparing the fluency data for message length 2 yielded significant effects of proficiency, $F(2, 56) = 206.27, p < .001$, and condition, $F(1, 56) = 4.92, p = .031$. Pairwise comparisons carried out to explore this significant effect of condition revealed that only the Low group sounded less fluent in the workload condition than in the clear condition ($p < .0001$). The effects of workload for message length 2 are illustrated in Figure 18 below.

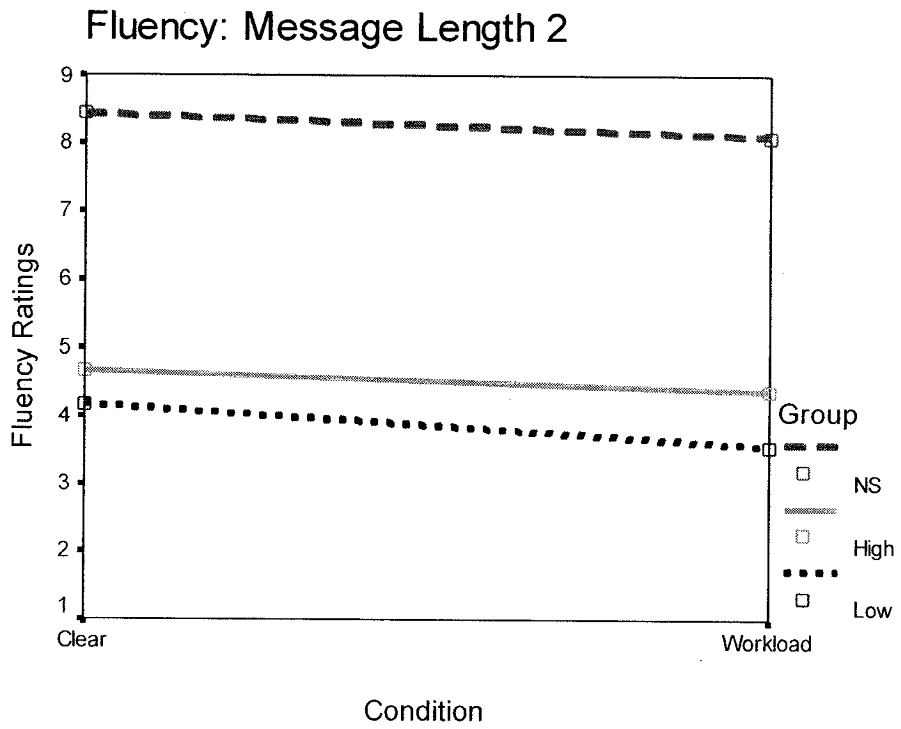


Figure 18. Workload effects for message length 2 are displayed as the difference in mean scores between the clear and workload conditions for each group.

Message length 3. The ANCOVA comparing fluency data for message length 3 yielded only a significant effect of proficiency, $F(2, 56) = 205.51, p < .0001$. Thus, there was no significant effect of workload for any of the groups for messages of length 3.

Workload effects. The effects of workload are expressed as the mean difference in fluency ratings between the clear and workload conditions. A higher value indicates a greater effect of workload and, therefore, a greater detriment to fluency as a function of increased workload. As can be seen in Table 21, the greatest effect of workload was observed in the Low group for messages of length 2.

Table 21

Effects of Workload on Fluency Ratings for Messages of Different Lengths

Proficiency Groups	Message Length		
	1	2	3
NS (*n* = 20)	.07	.34	.48
High (*n* = 20)	.07	.30	.22
Low (*n* = 20)	.24	.61	.12

Note. Means for the workload condition have been adjusted with the one-minute addition task scores entered as a covariate.

Workload summary. Analyses of the fluency data revealed that there was a significant effect of workload only for the Low group and only for messages of length 2. That is, the Low group was significantly less fluent in repeating messages of length 2 in the workload than in the clear condition. The greatest detriment to fluency due to additional cognitive workload was observed for the Low group for messages of length 2.

Confidence

Overall Analysis

The confidence data were submitted to a similar three-way omnibus repeated-measures ANOVA in order to determine if the three groups differed in confidence and if such differences interacted with condition and message length. This analysis yielded a significant main effect of condition, $F(1, 57) = 15.77, p < .001$, suggesting that overall, participants sounded more confidence in their repetition of messages in the clear condition than in the workload condition. This analysis also yielded a significant effect of proficiency, *F*(2, 57) = 254.19, *p* < .001, and length, *F*(2, 114) = 109.87, *p* < .001, and a significant length × proficiency interaction, *F*(4, 114) = 8.21, *p* < .001. Mean confidence ratings for each participant group are presented in Table 22. Higher ratings represent more confident-sounding speech.

Table 22

Mean Confidence Ratings (out of 9) and Their Standard Deviations for the Three Proficiency Groups in the Clear and Workload Conditions

Condition	Length	Proficiency Groups					
		NS (*n* = 20)		High (*n* = 20)		Low (*n* = 20)	
		M	*SD*	*M*	*SD*	*M*	*SD*
Clear	1	8.29	.35	5.81	.59	5.33	.59
	2	8.09	.50	4.75	.89	4.51	.82
	3	7.84	.70	4.14	1.22	3.73	1.13
Workload	1	8.08	.60	5.71	.89	5.13	.75
	2	7.76	.90	4.54	.96	3.95	.88
	3	7.28	.96	4.08	.87	3.55	.73

Message Length

Clear condition. The ANOVA comparing confidence ratings in the clear condition yielded significant main effects of proficiency, *F*(2,57) = 200.21, *p* < .001, and length, *F*(2,114) = 58.61, *p* < .001, and a significant length × proficiency interaction, *F*(4, 114) = 6.27, *p* < .001. Pairwise comparisons carried out to explore these significant effects (α = .006) revealed two findings. First, confidence ratings depended on participants' proficiency. The NS group sounded more confident than both the High and the Low groups for all message lengths. The High group sounded more confident than the Low group for messages of length 1 (*p* = .005), but for messages of length 2 and 3 these groups did not differ. Second, confidence ratings depended on the length of the message.

Both the High and Low groups sounded significantly *less* confident with each increasing message length ($p \leq .004$). By contrast, the NS group was equally confident for all message lengths. Confidence ratings in the clear condition are illustrated in Figure 19 (again, higher ratings represent more fluent speech).

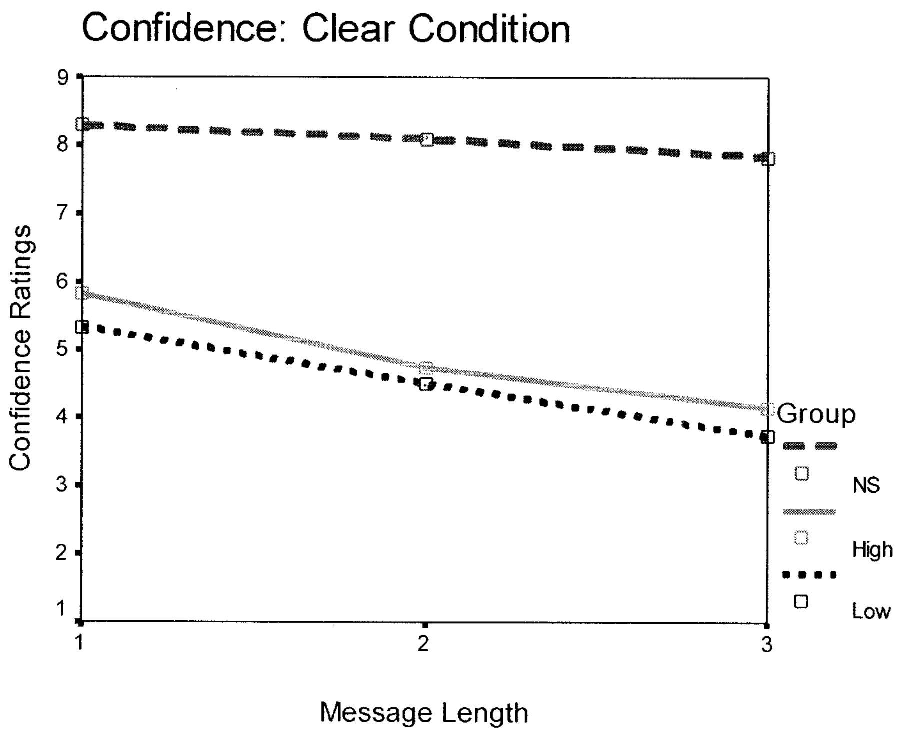


Figure 19. Confidence ratings in the clear condition for each participant group (NS, High, Low) for each message length.

Workload Condition. The ANCOVA comparing confidence in the workload condition yielded significant main effects of proficiency, $F(2, 56) = 206.74$, $p < .0001$ and length $F(2, 112) = 3.80, p = .025$. Pairwise comparisons carried out to explore these significant effects yielded two findings. First, confidence ratings depended on participants' proficiency. At all message lengths, the NS group was significantly more confident than the High and Low groups ($p < .0001$). The High and Low groups did not differ for any message length. Second, confidence ratings depended on the length of the message. The NS group sounded significantly more confident at message length 1 than at message length 3 ($p = .002$). The High and Low groups sounded significantly more confident at message length 1 than at message lengths 2 and 3 ($p < .0001$). Confidence ratings in the workload condition are illustrated in Figure 20.

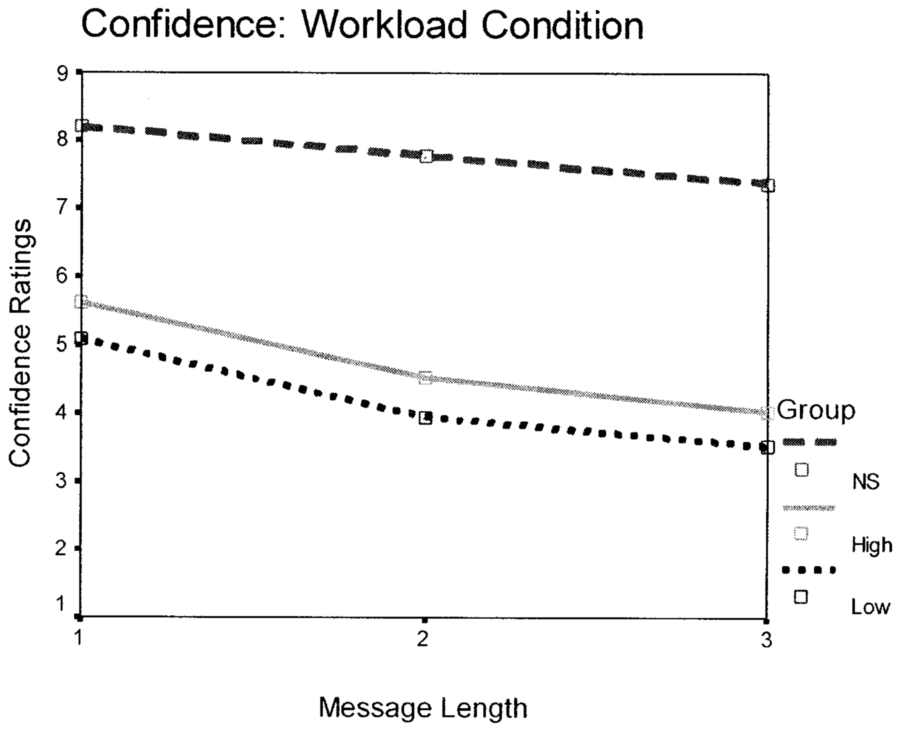


Figure 20. Confidence ratings in the workload condition for each participant group (NS, High, Low) for each message length.

Message length effects. In order to determine which group displayed the greatest detriment to confidence due to increased message length, the effects of message length are expressed as the mean difference in confidence ratings between messages of each consecutive length. As can be seen in Table 23, in the clear condition the greatest effect of message length was observed in the High group between messages of lengths 1 and 2. In the workload condition, the greatest effect of message length was observed in the Low group between messages of lengths 1 and 2.

Table 23

Effects of Message Length on Confidence Ratings

Proficiency Groups	Condition			
	Clear		Workload	
	1-2	2-3	1-2	2-3
NS (*n* = 20)	.20	.26	.43	.40
High (*n* = 20)	1.07	.61	1.10	.51
Low (*n* = 20)	.83	.78	1.14	.43

Note. Means for the workload condition have been adjusted with the one-minute addition task scores entered as a covariate.

Message length summary. Analysis of the confidence data revealed that confidence differed as a function of proficiency and message length. The NS group displayed greater confidence than the High and Low groups at all message lengths in both conditions. The High group displayed greater confidence than the Low group at message length 1 in the clear condition *only*. All groups displayed less confidence as message length increased, but these effects of message length differed as a function of proficiency and workload. The NS group sounded less confident repeating messages of length 3 in the workload condition *only*. By contrast, the High and Low groups were less confident with each increasing message length in the clear condition. In the workload condition, these groups sounded less confident when repeating messages of lengths 2 and 3 than when repeating messages of length 1. In the clear condition, the greatest detriment to confidence ratings as a result of increased message length was observed in the High group between messages of lengths 1 and 2. By contrast, in the workload condition the greatest effect of message length was observed in the Low group, also between messages of lengths 1 and 2.

Workload

Message length 1. The ANCOVA comparing the confidence data for message length 1 yielded significant effects of proficiency, $F(2, 56) = 209.45, p < .001$ and condition, $F(1, 56) = 4.27, p = .043$. This finding suggested that, overall, participants sounded more confident in the clear condition than in the workload condition when repeating messages of length 1. However, pairwise comparisons carried out to further investigate the effect of condition revealed that no individual participant group sounded significantly more confident in the clear condition than in the workload condition.

Message length 2. The ANCOVA comparing the confidence data for message length 2 yielded only a significant effect of proficiency, $F(2, 56) = 139.33, p < .001$. This finding thus suggested that confidence ratings did not differ as a function of workload condition for any of the groups for messages of length 2.

Message length 3. The ANCOVA comparing the confidence data for messages of length 3 yielded only a significant effect of proficiency, $F(2, 56) = 139.33, p < .001$. All groups sounded equally confident when repeating messages of length 3 in the clear and in the workload conditions.

Workload effects. The effects of workload are expressed as the mean difference in confidence ratings between the clear and workload conditions. A higher value indicates a greater effect of workload and, therefore, a greater detriment to confidence as a function of increased workload. As can be seen in Table 24, the greatest effect of workload was observed in the Low group for messages of length 2.

Table 24

Effects of Workload on Confidence Ratings

Proficiency Groups	Message Length		
	1	2	3
NS (*n* = 20)	.12	.37	.52
High (*n* = 20)	.16	.19	.09
Low (*n* = 20)	.24	.54	.19

Note. Means for the workload condition have been adjusted with the one-minute addition task scores entered as a covariate.

Workload summary. Analyses of the confidence data revealed that there was no significant effect of workload on confidence ratings for any of the participant groups.

Chapter Summary

Message Length

The length of the message did affect the speech production of all groups in that longer messages elicited lower speech production ratings. As hypothesized, these effects of message length differed as a function of proficiency and workload. The NS group was unaffected by message length in the clear condition. By contrast, in the workload condition, the NS group's fluency and confidence were affected at message length 3. In other words, the NS group displayed a detriment in speech production only at the *longest* message length in the *higher* workload condition, and only for the measures of fluency and confidence. By contrast, the High and Low groups displayed a detriment in speech production at message length 2 for *all* measures (accentedness, comprehensibility, fluency, confidence) in *both* conditions.

For *all* message lengths in *both* conditions (clear, workload), the NS group received significantly higher ratings than the High and Low groups for all measures. More specifically, for all measures, the High group received higher ratings than the Low group in both conditions. However, the High group did not receive significantly higher ratings than the Low group on all measures at every message length. At message length 3 for example, the High and Low groups were rated as equal for fluency and confidence in both conditions, whereas for accentedness and comprehensibility, the High group was rated as superior to the Low group at message length 3 in both conditions. Therefore, the hypothesis that the effects of message length would differ for all groups was supported in some cases.

Although the High group received higher ratings than the Low group, the greatest effect of message length, that is, the greatest drop in ratings due to increased message length, did not occur in the Low group for all measures as hypothesized. Rather, the greatest effect of message length was observed in the High group for the measures of accentedness, comprehensibility and fluency in both conditions, and for confidence in the clear condition, between messages of lengths 1 and 2. The greatest detriment in speech production due to message length was observed for the Low group only for the confidence measure in the workload condition.

Workload

The addition of the concurrent arithmetic task in the workload condition did result in lower ratings on all measures overall. However, a significant effect of workload was obtained only for the fluency measure for the Low group at message length 2. In other words, the Low group's speech was significantly less fluent in the workload condition than in the clear condition when repeating messages of length 2. The speech of the High and Low groups sounded more accented in the workload condition than in the clear condition at length 2; however, this difference did not reach the significance level. There was no significant effect of workload for the High or Low groups for the comprehensibility or confidence measures, and there was no significant effect of workload for the NS group for any of the speech production measures. Therefore, although the addition of the concurrent arithmetic task did adversely affect participants' speech production as hypothesized, the effect of workload was significant only for the fluency measure for the Low group when repeating messages of length 2.

In addition to receiving the lowest ratings overall on all measures, the Low group also displayed the greatest drop in ratings due to additional cognitive workload on the measures of comprehensibility, fluency and confidence when repeating messages of length 2. For the accentedness measure, the greatest drop in ratings due to additional cognitive workload was observed for the High group for messages of length 2.

CHAPTER 6

CORRELATIONAL ANALYSES

Chapter Overview

This chapter summarizes the results of correlational analyses carried out to explore the relationship among the six dependent variables of this study (navigation accuracy, repetition accuracy, accentedness, comprehensibility, fluency, confidence). First, the objectives of these analyses are stated. Next, the results of the correlational analyses for each set of dependent variables (performance accuracy and speech production) and the relationships between these two sets are reported in three sections. Finally, a summary of findings is provided for each set of relationships.

Main Objectives

The objective of the correlational analyses was to investigate relationships among the six dependent variables. One relationship of interest was the relationship between navigation and repetition accuracy. This relationship is important because congruence between message repetition and navigation is paramount to air safety in the aviation context. In other words, if a pilot repeats an air traffic controller's message correctly, the air traffic controller assumes that the pilot has understood the message and will appropriately carry out each command. Other important relationships are those among the speech production measures. Analyses of these relationships were motivated by one of the objectives of the present study—to investigate native-speaker perception of specific aspects of L2 speech in the aviation context, and to examine how these aspects of L2 speech relate to one another. The relationships among accentedness, comprehensibility and fluency are important in the aviation context because air traffic controllers' understanding of pilots' speech is likely related to the degree of accentedness, comprehensibility and fluency of the pilots' speech. As such, accentedness, comprehensibility and fluency are factors involved in ensuring accuracy and efficiency in controller-pilot communications. Also of interest was the relationship between these speech production measures (accentedness, comprehensibility, fluency) and perceived confidence of the speaker. An exploration of this relationship was important because it might reveal which aspects of speech production affect the level of confidence conveyed by the speaker. A final relationship of interest was the one between performance accuracy (navigation and repetition accuracy) and the speech production measures (accentedness, comprehensibility, fluency, confidence). The overall goal of the correlational analyses reported here was to investigate these relationships using the present dataset. The correlation coefficients among all of the variables investigated here are displayed in Table 25. (For reasons outlined below, these correlational analyses are restricted to the data for message length 3 only.)

Table 25

Pearson Correlation Coefficients (r) for the Six Dependent Variables for Each Group in the Clear and Workload Conditions for Message Length 3

Relationship	Condition					
	Clear			Workload		
	NS	High	Low	NS	High	Low
Navigation Accuracy & Repetition Accuracy	.86**	.87**	.69*	.78**	.86**	.72**
Navigation Accuracy & Accentedness	.36	.17	.41	.13	.16	-.13
Navigation Accuracy & Comprehensibility	.36	.15	.33	.09	.02	-.04
Navigation Accuracy & Fluency	.31	.24	.47	.15	.19	.25
Navigation Accuracy & Confidence	.38	.19	.51	.09	.12	.28
Repetition Accuracy & Accentedness	.45	.34	.58	.00	-.03	.08
Repetition Accuracy & Comprehensibility	.43	.40	.48	.01	-.04	.16
Repetition Accuracy & Fluency	.49	.47	.49	.16	.17	.13
Repetition Accuracy & Confidence	.53	.45	.47	.17	.17	.23
Accentedness & Comprehensibility	.93**	.89**	.93**	.86**	.78**	.89**
Accentedness & Fluency	.71**	.73**	.82**	.52	.43	.66*
Accentedness & Confidence	.59	.63*	.68*	.40	.08.	.52
Comprehensibility & Fluency	.73**	.90**	.72**	.73**	.76**	.69*
Comprehensibility & Confidence	.63*	.82**	.58	.61	.44	.70*
Fluency & Confidence	.93**	.93**	.94**	.89**	.81**	.89**

* $p \leq .003$ ** $p < .001$

Repetition Accuracy and Navigation Accuracy

Due to ceiling effects for message lengths 1 and 2, correlational analyses were conducted only for the message length 3 data in each condition (clear, workload). In the clear condition, bivariate Pearson correlation analyses (two-tailed) revealed that repetition and navigation accuracy scores were significantly correlated for all groups when responding to messages of length 3 ($\alpha = .003$). In other words, each group's navigation and message repetition accuracy scores positively correlated when responding to messages of length 3 in the clear condition. This finding suggests that, for the most part, participants carried out the instructions as they had repeated them. The highest correlation coefficient was obtained for the High group (*r* = .87). A similarly high correlation was obtained for the NS group (*r* = .86). By contrast, the Low group displayed a slightly lower correlation coefficient than these groups (*r* = .69); however, between-group comparisons of these correlation coefficients revealed no significant differences (DeCoster & Leistico, 2005).

Because, as discussed earlier, the participant groups differed in their ability to solve mental arithmetic problems, effects of workload may have been confounded with the participants' proficiency. Therefore, to examine the relationship between repetition and navigation accuracy in the workload condition, while controlling for differences in mental arithmetic, first-order partial correlations were computed. In these analyses, the participants' scores on the mental arithmetic task were partialled out from the relationship between navigation and repetition accuracy. In the workload condition, partial correlations between navigation and repetition accuracy for message length 3 revealed that each group's navigation and message repetition scores correlated positively. As in the clear condition, the highest correlation coefficient was obtained for the High group (*r* = .86), and the lowest correlation coefficient was obtained for the Low group (*r* = .72). The correlation coefficient obtained for the NS group (*r* = .79) fell between those of the High and Low groups. Although a slightly higher correlation was obtained for the High group than for the other two groups, between-group comparisons of correlation coefficients revealed no significant differences among any of the groups (DeCoster & Leistico, 2005).

Speech Production Measures

As with the analyses discussed above, correlational analyses were conducted only for messages of length 3 in each condition (clear, workload) due to (near) ceiling effects for message lengths 1 and 2. In the clear condition, these analyses revealed that accentedness, comprehensibility and fluency were significantly correlated for all groups ($\alpha = .003$). This finding suggested that these speech production measures estimated several different aspects of the same overall construct—ability to sound native-like in the L2. For the NS group, accentedness correlated significantly with comprehensibility ($r = .93$) and fluency ($r = .71$), and comprehensibility correlated with fluency ($r = .73$). Confidence correlated with comprehensibility ($r = .63$) and fluency ($r = .93$). For the High group, accentedness correlated significantly with comprehensibility ($r = .89$), fluency ($r = .73$) and confidence ($r = .63$). Comprehensibility correlated with fluency ($r = .90$). Confidence correlated with comprehensibility ($r = .82$) and fluency ($r = .93$). For the Low group, accentedness correlated significantly with comprehensibility ($r = .93$), fluency ($r = .82$) and confidence ($r = .68$). Comprehensibility correlated with fluency ($r = .72$). As with the NS and High groups, confidence correlated strongly with fluency ($r = .93$), but, unlike the NS and High groups, not significantly with comprehensibility. Between-group comparisons of correlation coefficients revealed no significant differences among any of the groups for any of the speech production relationships in the clear condition (DeCoster & Leistico, 2005).

In the workload condition, partial correlations among the speech production measures for message length 3 (with the participants' mental arithmetic scores partialled out) also revealed a number of strong relationships. For the NS group, accentedness correlated significantly with comprehensibility (*r* = .86), and comprehensibility with fluency (*r* = .73). Confidence correlated with fluency (*r* = .89). For the High group, accentedness correlated significantly with comprehensibility (*r* = .78). Comprehensibility correlated with fluency (*r* = .76), as did confidence (*r* = .81). For the Low group, accentedness correlated significantly with comprehensibility (*r* = .89) and fluency (*r* = .66). Comprehensibility correlated with fluency (*r* = .69) and confidence (*r* = .70). As for the NS and High groups, confidence correlated strongly with fluency (*r* = .89) for the Low group in the workload condition. Between-group comparisons of correlation coefficients revealed no significant differences among any of the groups for any of the speech production relationships in the workload condition (DeCoster & Leistico, 2005).

Performance Accuracy and Speech Production Measures

Correlational analyses of the data for message length 3 revealed that there were no significant correlations between either of the performance accuracy measures (navigation or repetition accuracy) and any of the speech production measures (accentedness, comprehensibility, fluency, confidence) for any of the groups (NS, High, Low) in either condition (clear, workload). This lack of significant correlations between the two sets of measures may indicate that the underlying cognitive constructs measured by the two sets of measures are separate, at least to the extent that they are engaged in a simulated pilot navigation task. The performance accuracy measures (navigation and repetition accuracy) reflect the participants' ability to understand and retain messages presented in the L2, whereas the speech production measures likely tap into participants' ability to sound native-like in the L2.

Chapter Summary

Navigation and Repetition Accuracy

A significant positive correlation between navigation and repetition accuracy was obtained for all groups. The highest correlation coefficient was obtained for the High group in both the clear and workload conditions, while the lowest was obtained for the Low group in both the clear and workload conditions. There were no significant between-group differences in the correlation coefficients for either of the dependent variables.

Speech Production Measures

The correlational analyses of the speech production data yielded a number of strong relationships, suggesting that these measures were related to a common underlying construct—the ability to sound native-like in the L2. There were three relationships that were significantly correlated for all groups in both the clear and workload conditions. These were accentedness and comprehensibility, comprehensibility and fluency, and fluency and confidence. For nearly all the relationships for all of the groups, correlation coefficients were higher in the clear condition than in the workload condition. The correlation between fluency and confidence was strikingly high for all groups in both conditions, suggesting that speakers who are more fluent (that is, speakers who speak without false starts and undue pauses) may convey greater confidence than speakers who are less fluent. Not surprisingly, accentedness and comprehensibility were highly correlated, suggesting that the native-speaker raters understood with greater ease speech they perceived to be native-like in accentedness.

A number of less consistent correlations were also revealed in the correlational analyses of the speech production measures. Accentedess and fluency were significantly correlated for all three groups in the clear condition, but only for the Low group in the workload condition. The correlation coefficients for this relationship (although not significant for all groups in both conditions) suggest an influential relationship between these two measures. Accentedness and confidence were significantly correlated for the High and Low groups in the clear condition only. Comprehensibility and confidence were significantly correlated for the NS and High groups in the clear condition, and for the Low group in the workload condition.

Performance Accuracy and Speech Production Measures

There were no significant correlations between these two sets of measures, suggesting that there was no straightforward relationship between the quality of the participants' speech and the accuracy of their task performance.

CHAPTER 7

DISCUSSION

Chapter Overview

This chapter summarizes and discusses the findings of the present study. As in the previous chapters, results are discussed separately for navigation and repetition accuracy and the speech production measures. For each set of the dependent variables, the results relevant to each set of hypotheses (message length and workload) are discussed in light of previous experimental findings.

Navigation and Repetition Accuracy

The present study investigated the effects of proficiency, message length and workload using messages containing a maximum of three commands with the objective of determining whether the additional cognitive workload imposed by concurrent tasks (a factor present in the pilot work environment) would affect the performance of participants operating in their L2. In other words, the objective was to examine whether real-life factors, such as concurrent task management and using the L2, might have a detrimental or even deleterious effect on pilot response (in terms of navigation and message repetition) to messages of three commands or fewer.

Accurate controller-pilot communications are paramount to air safety; therefore, in the present study, for navigation and repetition accuracy, mean benchmark scores of 80 percent and above were deemed to be adequate indicators for safe communications. The eighty percent rule was applied due to its prevalence as a benchmark in training and testing scenarios. It should be emphasized, however, that the 20 percent margin of error accepted here is likely not reflective of the actual error rates in communications between *real* controllers and pilots, in that real controller-pilot error rates are likely much lower. Nevertheless, the findings of the present study are revealing. For convenience, Tables 26 and 27 summarize the key findings for navigation and repetition accuracy, defined as proportion of correct navigation and repetition responses, respectively. The scores from the workload condition represent navigation and repetition accuracy means that have been adjusted for the effect of mental arithmetic ability (used as a covariate in all workload analyses).

Table 26

Navigation Accuracy Scores Expressed as Proportion of Correct Responses for All Groups in the Clear and Workload Conditions

Proficiency Groups	Message Length					
	1		2		3	
	Clear	Workload	Clear	Workload	Clear	Workload
NS	.99	.93	.95	.92	.83	.73*
High	.99	.91	.89	.76*	.69*	.60*
Low	.97	.91	.89	.77*	.70*	.48*

Note. Means for the workload condition have been adjusted with the one-minute addition task scores entered as a covariate.

* Mean navigation accuracy score below 80 percent

Table 27

Repetition Accuracy Scores Expressed as Proportion of Correct Responses for All Groups in the Clear and Workload Conditions

Proficiency Groups	Message Length					
	1		2		3	
	Clear	Workload	Clear	Workload	Clear	Workload
NS	1.00	1.00	.99	1.00	.79*	.78*
High	.99	.99	.89	.84	.67*	.53*
Low	.99	1.00	.91	.84	.68*	.46*

Note. Means for the workload condition have been adjusted with the one-minute addition task scores entered as a covariate.

* Mean navigation accuracy score below 80 percent

Message Length

One main objective of the present study was to determine to what extent the length of the message would affect participants' navigation and readback accuracy, and whether such effects of message length would differ as a function of L2 proficiency. The findings of several studies using a similar experimental paradigm (Barshi, 1997, 1998; Barshi & Healy, 1998, 2002; Mauro & Barshi, 1999; Schneider et al., 2004) as well as other studies investigating the effects of controller message length on controller-pilot communications (Morrow & Rodvold, 1993, Morrow et al., 1993, 1994) have indicated that controller messages should contain not more than three commands for L1 speakers.

As in previous studies in which a similar pilot navigation task was used (Barshi, 1997, 1998; Barshi & Healy, 1998, 2002; Mauro & Barshi, 1999; Schneider et al., 2004) all groups' navigation accuracy and repetition accuracy were adversely affected by longer messages. With respect to the performance of L2 participants, the findings of the present study replicated the results of a previous study which showed that the extent to which L2 participants are affected by message length depends on L2 proficiency (Barshi & Healy, 1998). In addition, the findings of the present study extended previous results by suggesting that the extent to which L2 participants are affected by message length also depends on cognitive workload.

The findings of the present study for navigation accuracy in the clear condition are similar to those of Barshi and Healy (1998). Based on their findings, Barshi and Healy (1998) recommended that controllers limit messages to a length of two commands when communicating with pilots of a relatively low proficiency level. There is one striking difference, however, between the findings of Barshi and Healy (1998) and those of the present study. In the present study, both L2 groups (High and Low) navigated well below 80 percent accuracy when responding to messages of length 3. By contrast, in the study conducted by Barshi and Healy (1998) the High group navigated with a degree of accuracy similar to that of the NS group. This discrepancy may be attributed to differences in the proficiency levels of the High group participants in the present and previous study. In both studies, the terms "high" and "low" were used relatively. However, in the present study, the measures used to assign L2 participants to proficiency groups suggest that the vast majority of participants in the "high" group may be at an intermediate level. Based on the results of the present study, it is recommended that in low workload conditions, controllers limit their messages to two commands when communicating with pilots of low and intermediate levels of L2 proficiency. This discrepancy between the results of this study and those of Barshi and Healy (1998) highlights a need for clearer definition of proficiency level (ideally in relation to the globally-recognized ICAO rating scale) in order for consistent recommendations regarding ideal message length in controller-pilot L2 communications to be made.

Of particular interest for the present study is the finding of consistent differences between the NS and L2 groups in the workload condition. In the workload condition, statistically-significant between-group differences were obtained for both navigation and repetition accuracy. In contrast to the results of the clear condition, both L2 groups navigated and repeated messages significantly less accurately than the NS group, indicating that in the workload condition, the L2 groups were adversely affected by message length to a greater extent than the NS group. It is important to note that no significant interaction between proficiency and workload condition was obtained; however, these results suggest that the performance of the L2 groups may have been affected to a greater extent than the NS group by the additional cognitive workload. This decline in performance accuracy for the L2 groups under high workload conditions may be due to overall increased cognitive workload imposed by the addition of the arithmetic task and/or to the demands of "dual" processing, imposed by concurrent task performance in the workload condition. It is possible that the observed performance decline is due to the combined effect of these factors, both concurrent task management as well as processing long messages. Indeed, the participants were required to retain, repeat, and carry out commands of increasing length while performing a mental arithmetic task in the concurrent task paradigm of the workload condition. Future analyses of navigation and repetition accuracy in relation to measures of working memory and attention control obtained in the present study may help determine the extent to which performance declines in the workload condition are caused by dual-task demands versus overall increased processing load. For example, a relationship between high attention control ability and high performance in the workload condition might suggest that the ability to switch attention quickly between two tasks predicts performance in a concurrent task paradigm. Alternately, a relationship between high working memory capacity and high performance in the workload condition might suggest that the ability to retain and process a relatively large number of items in working memory predicts performance in a concurrent task paradigm. As mentioned above, analysis and discussion of the memory and attention measures is outside the scope of the research questions investigated here, and therefore, these will be the subject of future research.

Workload

The findings of previous studies investigating the effects of workload on pilot task performance (Chou et al., 1996; Dismukes et al., 1998, 2001; Loukopoulis et al., 2003; Raby & Wickens, 1994) indicate that high-workload conditions can have detrimental effects on pilot task performance. More specifically, concurrent tasks can produce a detrimental effect on performance in a variety of tasks, including high-priority tasks such as navigation as well as lower-priority tasks such as communication. Therefore, the present study investigated L2 proficiency as a factor in performance (navigation and message repetition) under high-workload conditions in a simulated pilot navigation task. There are no other known studies that have investigated the effects of workload in relation to L2 proficiency, particularly in the aviation context.

The addition of the concurrent mental arithmetic task in the workload condition resulted in a navigation performance detriment in relation to the clear condition for all groups. While the NS group displayed no significant effects of message length in the clear condition, in the workload condition the NS group displayed a significant decline in navigation accuracy in response to messages of length 3, resulting in a mean accuracy score below 80 percent. For the NS group, the difference in navigation accuracy between the clear and workload conditions was significant in response to messages of length 3. Whereas the High and Low groups navigated with well above 80 percent accuracy in response to messages of length 2 in the clear condition, in the workload condition both L2 groups (High and Low) displayed a significant decline in navigation accuracy, such that their mean scores fell well below 80 percent accuracy in response to messages of length 2. Both groups displayed further significant declines in response to messages of length 3, resulting in mean navigation accuracy scores of 60 and 48 percent, respectively. Therefore, in relation to navigation accuracy, the present study yielded data in support of the second hypothesis. All groups' performance appears to have been adversely affected by the addition of the concurrent arithmetic task, and the greatest drop in performance between the clear and workload conditions was observed for the Low group. For this group, an effect of workload was observed for all message lengths with the greatest drop in navigation accuracy occurring in response to messages of length 3. The mean navigation accuracy scores for the workload condition suggest that under conditions of high workload, controllers should limit their messages to two commands in L1 communications with pilots. In L2 communications with pilots of low and intermediate L2 proficiency, controllers should limit their messages to one command.

The effects of workload differed for repetition accuracy. A significant decline in repetition accuracy in the workload condition versus the clear condition was observed for the Low group only in response to messages of length 3. By contrast, although their mean scores tended to be lower in the workload condition, the NS and High groups displayed no significant effect of workload for any message length. Therefore, the second hypothesis was only partially supported with respect to repetition accuracy. The greatest drop in performance was observed for the Low group as hypothesized; however, only the Low group's performance (as opposed to all groups' performance) was significantly affected by the addition of the concurrent arithmetic task.

One factor that might explain the smaller effect of workload on repetition accuracy in relation to navigation accuracy may be the placement of the concurrent arithmetic task in relation to message repetition and navigation. To recapitulate, the task sequence in the workload condition was the following: First, participants heard the message. Next, the number (which they were required to inverse and mentally add to the original number) appeared. Next, they repeated the message while solving the mental arithmetic problem. Immediately after repeating the message, participants uttered the answer to the arithmetic problem. Finally, they carried out the instructions given in the message by navigating on the grid. In the clear condition (where the mental arithmetic task was absent), it is likely that participants subvocally rehearsed the message they had heard (Baddeley, 2003) in order to retain it in working memory until they had finished navigating. In other words, it is likely that subvocal rehearsal helped ensure navigation accuracy. In the workload condition, it is likely that participants attempted to subvocally rehearse each message as well. However, subvocal rehearsal was interrupted by the utterance of the answer to the mental arithmetic problem. This interruption in subvocal rehearsal thus interfered with participants' ability to remember the message they had heard just seconds ago, resulting in reduced navigation accuracy in relation to repetition accuracy. This interruption of the subvocal rehearsal process may have affected navigation accuracy but not repetition accuracy because the participants had already repeated the message *prior* to uttering the response to the arithmetic problem.

Although to a lesser extent than navigation accuracy, participants' repetition accuracy was also reduced in the concurrent-task paradigm of the workload condition. Participants were instructed that it was imperative to provide the response to the mental arithmetic problem immediately following repetition of the message; therefore, the calculation of the mental arithmetic problem was concurrent with message repetition. That only the Low group displayed a significant detriment to repetition accuracy in the concurrent task paradigm suggests that the cognitive demands of performing these tasks concurrently were greatest for this group. While the NS and High groups were able to accommodate the additional cognitive workload of the mental arithmetic task while repeating three-command messages (evidenced in this group's ability to maintain a level of accuracy statistically comparable to that of the clear condition), the Low group's repetitions of messages of length 3 appear to have been compromised by the concurrent arithmetic task.

Baddeley's model of working memory may provide some explanation for the lower navigation and repetition accuracy scores obtained for the Low group. According to Baddeley's model of working memory (Baddeley, 1992, 1996, 2003), visually presented information is rehearsed subvocally, provided it can be named. Therefore, although the arithmetic task may appear to be non-linguistic relative to the message repetition task, it likely engaged the phonological loop component of working memory, associated with auditory input. This may be one reason why the performance of the L2 participants, particularly those of a low level of L2 proficiency, was most affected by both message length and the additional cognitive workload imposed by the arithmetic task. Performance of the experimental tasks depended on working memory, most likely on the phonological subcomponent (the phonological loop), for message retention and mental arithmetic, and although working memory capacity is a relatively fixed ability (Osaka, Osaka, & Groner, 1993), working memory demands required to operate in the L2 are likely greater than those required to operate in the L1. Furthermore, the cognitive resources required to operate in the L2 are likely dependent on L2 proficiency (Chincotta & Underwood, 1998).

One noteworthy observation is that for messages of length 3, where the cognitive load imposed by message length alone is relatively high, the L2 groups displayed higher accuracy for navigation than for message repetition in both the clear and workload conditions. Previous studies of the effects of workload on pilot task performance (Chou et al., 1996; Dismukes et al., 1998, 2001; Loukopoulis et al., 2003; Raby & Wickens, 1994) indicate that high-workload conditions can have a detrimental effect on pilot task performance, particularly on tasks perceived to be of lower priority. It is likely that all participants perceived the navigation aspect of the task to be of greater importance than message repetition, particularly given that in the practice phase of the task, feedback ("correct" or "wrong") was given only in response to navigation accuracy, and not in response to repetition accuracy. Therefore, for messages of length 3, the L2 participants may have employed strategies, such as visualization of the movements on the grid, in order to maximize performance accuracy in the task perceived to be of higher priority (the navigation task). This strategy would presumably off-load cognitive work from the phonological loop component of working memory onto the visuo-spatial sketchpad component (Baddeley, 1992, 1996, 2003). This is a potentially important observation because in the cockpit environment, L2 pilots may employ similar strategies (as well as others not possible in a laboratory setting, such as, for example, writing information down) in order to compensate for the increased demands of operating in their L2. Given this possibility, that L2 pilots may employ strategies to compensate for reduced retention of information presented auditorily in the L2, the results of the present study should be viewed with caution. They should not be interpreted as evidence that under high-workload conditions, pilots of lower L2 proficiency will necessarily perform less competently than pilots operating in their L1, particularly on high priority tasks such as aviation and navigation. Nonetheless, these findings do suggest that pilots of low and intermediate L2 proficiency may comprehend and retain less information contained in long controller messages than pilots operating in their L1, particularly under high-workload conditions. Further experiments should be conducted in order to explore possible interactions between L2 proficiency and cognitive workload in relation to pilot performance.

Correlations

Analyses investigating the relationship between navigation and repetition accuracy revealed strong relationships between these two measures for all groups in both conditions, suggesting that, for the most part, participants navigated in accordance with the content of their message repetition. The lowest correlation coefficient was obtained for the Low group in both conditions, although between-group comparisons of correlation coefficients revealed no significant differences. This finding is in agreement with the findings of Barshi (1998), who found that participants' navigation and repetition accuracy scores tended to be similar. This finding is important from the perspective of air safety, where pilot adherence to controller instructions (as reflected in pilots' readbacks and their subsequent actions), is of paramount importance.

Speech Production

Message length

The first hypothesis of the present study, that longer messages would adversely affect the quality of participants' speech (evidenced in listeners' perceptual judgments), was supported by the results of the speech production analyses. The NS group received significantly higher ratings than both L2 groups, and the High group received higher ratings than the Low group for all speech production measures. However, the effects of message length differed for the NS and L2 groups (High and Low). The NS group received significantly lower ratings due to increased message length in the workload condition only, and then only for the fluency and confidence measures in response to messages of length 3. By contrast, both the High and Low groups received significantly lower ratings due to increased message length for all measures (accentedness, comprehensibility, fluency, confidence) in both conditions when responding to messages of length 2 and longer.

Although no previous studies have measured the effects of message length on listener perception of L2 speech, previous studies have investigated the effects of utterance length on L1 speech production using measurements of lip movements (Maner et al., 2000; Kleinow & Smith, 2000). These studies have yielded contrasting results. Using a measure of lower lip movement stability, Maner et al. (2000) found that the speech production of both children and adults was adversely affected (in that lower lip movements were less stable) by message length and complexity, and that children (assumed to have less stable speech production systems) were more adversely affected than adults. However, this study confounded utterance length and sentence complexity, and it is therefore possible that the results obtained were due to utterance complexity as opposed to utterance length. Kleinow and Smith (2000) conducted a study investigating the separate effects of utterance length and utterance complexity on adults who stutter and those who do not. They found that neither group (stuttering or normally fluent) was affected by message length, but that the stuttering group was affected by message complexity. The results of the present study are in agreement with the findings of Kleinow and Smith (2000) in that the NS group was not significantly affected by message length in the clear condition. In other words, the results of the present study suggest that L1 speech production is unaffected by message length, at least to the extent such effects are perceived by native-speaker raters when responding to messages of up to three commands.

The speech production of the NS group was, however, affected by message length in the workload condition for the fluency and confidence measures, suggesting that native speakers are susceptible to perceptible changes in speech production due to message length in some circumstances. Among such circumstances are situations involving high cognitive workload, of the kind investigated in the present study. It is possible, then, that the significant effects of message length for confidence and fluency in the workload condition arose as a consequence of the concurrent-task paradigm which required participants to divide attentional resources between two tasks. Messages that do not typically cause perceptible changes in speech production for native speakers under normal processing conditions may become more challenging, at least with respect to subjective measures of fluency and confidence, when participants are required to perform two concurrent tasks.

It is likely that increased cognitive workload imposed by a concurrent task amplified effects of message length that may have gone undetected in the clear condition. This claim is supported by the inverse relationship between speech production ratings and message length. Even in the clear condition, these ratings for the NS group tended to be lower (although not significantly) for longer messages than for shorter messages. It appears, then, that subtle changes in speech production might occur for longer messages, even in low-workload conditions, and that these changes will be salient and likely perceptible to listeners when cognitive demands are high. Further research needs to determine exactly what changes in the speech signal, arising as a result of increased cognitive workload, might affect listeners' perception of fluency and confidence of speech.

By contrast, the speech production of the L2 groups was affected by message length in the clear condition. For both the High and Low groups, significantly lower ratings were obtained for all measures (accentedness, comprehensibility, fluency, confidence) when responding to messages of length 2 than when responding to messages of length 1 in the clear condition. In some cases, there were also significant declines in ratings between messages of lengths 2 and 3. These results suggest that the speech of the High and Low groups, to the extent perceived by native-speaker raters, was more adversely affected by long messages than the speech of the NS group. In other words, even in the clear condition, producing longer messages was more cognitively demanding for L2 speakers than for the L1 speakers.

Interestingly, the greatest detriment to speech production for all measures in the clear condition occurred not in the Low group as hypothesized, but rather in the High group, particularly between messages of lengths 1 and 2. This finding was surprising. It was hypothesized that the greatest detriment to speech production due to increased message length would occur in the Low group because communications in the L2 would presumably be more cognitively demanding for the group with the lower level of L2 proficiency. Therefore, the Low group was expected to display a greater detriment to speech production due to the increased cognitive demands of listening and responding to messages in the L2. This counterintuitive finding is likely due to possible floor effects in speech production ratings for the Low group. As the results of the production analyses suggested, the speech ratings of the Low group were, for the most part, lower than the comparable ratings of the High group (even in the clear condition). It is not surprising therefore that messages of increasing length appeared to be less detrimental to the Low group, whose ratings were already low, than to the High group, whose ratings were significantly higher. Thus, the High group was perhaps able to sound more native-like when producing short utterances, but was unable to maintain this degree of native-like speech when producing longer utterances.

Workload

As hypothesized, the additional cognitive workload imposed by the arithmetic task resulted in overall lower native-speaker ratings in the workload condition than in the clear condition for all measures (accentedness, comprehensibility, fluency and confidence). In other words, native speaker raters were able to perceive differences in the speech samples excised from the clear versus the workload condition. The majority of studies investigating the effects of workload on speech production (Brenner & Shipp, 1987; Carballo & Mendoza, 1998; Dromey & Benson, 2003; Hecker et al., 1968; Lively et al., 1993; Williams & Stevens, 1972) have been conducted with the goal of identifying acoustic, labio-kinematic, or temporal differences in speech samples produced under high and low workload conditions. By contrast, the objective of the present study was to investigate whether listeners' perception of speech would be affected by workload to the extent that lower native-speaker ratings would be elicited in relation to specific criteria that might affect controller-pilot communications (accentedness, comprehensibility, fluency, confidence).

At least two studies (Hecker et al., 1968; Lively et al., 1993) have investigated the effects of workload on L1 listener perception of L1 speech in a manner relevant to the present study. In both studies, listeners rated utterances produced under low and high workload conditions. Hecker et al. (1968) found that listeners could identify the utterances produced by some speakers in the high workload condition with 90 percent accuracy. However, these same listeners could identify utterances produced by other speakers under the high workload condition only at chance level, indicating a high degree of individual differences in response to the high workload condition. Lively et al. (1993) reported somewhat different findings. They found that listeners judged utterances produced under a high cognitive workload condition as being more intelligible than utterances produced under a "clear" condition. The findings of these studies must be treated with caution, however, due to the small number of participants (ten and five, respectively). The results reported here are in contrast with those of Lively et al. (1993) in that native-speaker ratings were equivalent or lower in the workload versus the clear condition. These differences in findings may be attributed to the different types of concurrent tasks used (mental arithmetic vs. visual tracking) or the different stimuli presented to raters (complete message repetitions vs. an elongated vowel sound).

More specifically, Lively et al. (1993) found that acoustic changes in speech production due to cognitive workload paralleled intelligibility ratings assigned by listeners, and changes in amplitude and amplitude variability were identified as major factors contributing to intelligibility. The authors interpreted these results in support of Lindblom's Hypo- and Hyper-articulation theory (1990), which posits that speakers adapt their speech to suit the demands of the environment (including level of cognitive workload) in order to maximize discriminability. In the present study, mean speech ratings were consistently lower in the workload than in the clear condition for all measures, suggesting that speech was more accented, less comprehensible, less fluent and less confident-sounding under higher cognitive workload. At first glance, the findings of the present study appear to contradict Lindblom's theory, in that speech was degraded under high cognitive workload. However, based on subjective native-speaker ratings, not objective speech measures, these findings are perhaps insufficient to draw such conclusions. Further analyses (temporal, acoustic) of the speech data should be conducted in order to explore their relevance to Lindblom's theory and their compatibility with Lively et al.'s findings.

As hypothesized, the detrimental effect of the additional cognitive workload imposed by the arithmetic task was greater for the L2 groups than for the NS group. This effect of cognitive workload was statistically reliable only for the fluency measure for the Low group in response to messages of length 2. The finding of a significant effect of workload for the fluency measure is in agreement with the finding of Ooman and Postma (2001) that the production of filled repetitions and pauses increased under a condition of concurrent task performance. It is likely that the speech perceived as less fluent by the raters in the present study contained more filled pauses and repetitions than the speech judged as more fluent, given that raters were advised to judge fluency based on undue pauses, filled pauses, hesitations, or dysfluencies such as false starts and repetitions. In support of the link between dysfluencies and filled pauses, on the one hand, and ratings of perceived fluency, on the other, are the findings of Lennon (1990), who identified both filled pauses and pause duration as variables affecting perceived fluency.

The significant effect of workload on fluency obtained for the Low group is interesting in relation to automatic processing and the production of filled pauses. Ooman and Postma (2001), following earlier findings of Smith and Clark (1993), suggested that an increase in the production of filled pauses and repetitions under high workload conditions indicates that the production of filled pauses and repetitions is an automatic as opposed to a controlled process. Briefly, automatic processing is typically defined as one that is fast, ballistic (unstoppable) and that proceeds without conscious intention or awareness. On the other hand, strategic (controlled) processing is slower; it requires conscious intention and awareness, and it is driven by specific, often conscious, processing strategies (see Schneider & Chein, 2003, and Segalowitz & Hulstijn, 2005, for discussion of automatic and controlled processing in L1 and L2). Smith and Clark (1993) posited that the production of filled pauses was a deliberate strategy employed by speakers when answering questions. Following this claim, Ooman and Postma reasoned that if the production of filled pauses and repetitions was a deliberate strategy, placing a relatively high demand on attentional resources, then their production should decrease under conditions of high workload where attentional demands are high. By contrast, Ooman and Postma found that the production of filled pauses and repetitions increased in the concurrent task paradigm, suggesting that their production is an automatic process demanding few attentional resources. This interpretation is relevant to the hypotheses and findings of the present study. The hypotheses investigated here were based on the assumption that greater attentional resources would be required for L2 speech production than for L1 speech production. As such, it was predicted that the greatest effect of workload, that is, the greatest detriment to speech production in the workload condition, would be obtained for the L2 groups, and more specifically, for the Low group. Due to greater attentional resource requirements for L2 production, the Low group likely had fewer attentional resources available relative to the NS and High groups when performing a dual task in the workload condition and therefore displayed the greatest decline in fluency between the clear and workload conditions. Whether or not this detriment to fluency was due to an automatic process of producing an increased number of pauses and other dysfluencies under high cognitive workload needs to be clarified in future research.

Correlations

Pearson correlational analyses of the speech production data for messages of length 3 yielded a number of strong and reliable relationships for the four criteria rated by native-speaker judges (accentedness, comprehensibility, fluency, confidence). Three relationships were consistently high for all participant groups (NS, High, Low) in both conditions, namely, between fluency and confidence, between comprehensibility and fluency, and between accentedness and comprehensibility. Moreover, accentedness and fluency were significantly correlated in the clear condition, but only for the Low group in the workload condition. Although not significantly so, in general, correlation coefficients were lower in the workload condition than in the clear condition.

Fluency and confidence. The first of these relationships, the one between perceived fluency and the metacognitive measure of perceived confidence, suggests that the perceived confidence of the speaker in the accuracy of their message repetition might be strongly related to fluency, that is, the presence of undue pauses and/or other dysfluencies in the speech sample. This relationship is relevant to the findings of Smith and Clark (1993) and Brennan and Williams (1995). In both studies it was found that prosodic measures, namely filled and unfilled pauses and pause durations, were related to speakers' confidence (or, "feeling of knowing") in the accuracy of their response to a question. Furthermore, Brennan and Williams (1995) found that listeners' perception of the speaker's confidence in the accuracy of their response was related to unfilled pause duration, in that longer pauses were associated with lower ratings of the speakers' confidence in the accuracy of their response. Because similar temporal criteria (i.e., undue pauses, filled pauses, hesitations, etc.) were used by the raters in this study for fluency ratings, it is likely that less fluent speech was characterized by more numerous and longer pauses than more fluent speech. Therefore, it is likely that the results of the present study support those of Brennan and Williams (1995), in that the perceived confidence of the speaker may be related to the frequency and duration of pauses in the utterance, although temporal analyses of these measures (pause frequency and duration) should be conducted in order to explore this possibility.

Interestingly, there was no significant correlation obtained in the present study between perceived confidence and performance accuracy (neither navigation nor repetition accuracy). By contrast, both Smith and Clark (1993) and Brennan and Williams (1995) obtained high correlations between response accuracy (in responding to general knowledge questions) and confidence ratings. These differences in findings may be related to the way in which the speech samples were excised in the present study, whereby the pauses preceding and following the samples, as well as the stimulus utterances themselves, were eliminated. (For example, participants may have produced longer pauses prior to uttering erroneous repetitions in response to the stimuli, and these pauses would have been eliminated in the present study.) Alternatively, these differences may be related to the type of speech samples used. Both previous studies used responses to a general knowledge question, whereas in the present study, the speech samples were repetitions of context-specific, formulaic, recurrent messages. Clearly, the cognitive processes involved in responding to these two different types of stimuli would differ and could result in differences in speech production outcomes. This lack of correlation between confidence and performance accuracy in combination with the highly consistent correlation between confidence and fluency may be important in the aviation context, where communications are directly related to task performance. In other words, it is possible that in situations requiring judgment of a pilot's metacognitive state, such as level of confidence, that controller decisions may be influenced by non performance-related factors such as L2 fluency.

Comprehensibility and fluency. The second relationship in the present study found to be statistically significant for all groups in both conditions was the one between perceived comprehensibility and fluency. Munro and Derwing (1995) and Derwing and Munro (1997) also found that prosody (subjectively-rated by the experimenters, using filtered speech samples) was related to comprehensibility for some listeners. Taken together, these findings suggest that more fluent speech is likely more comprehensible in a variety of contexts: when listening and rating simple everyday utterances as well as when evaluating messages specific to controller-pilot communications.

Accentedness and comprehensibility. Another consistently high and statistically reliable relationship was the one between perceived accentedness and perceived comprehensibility. Several previous studies have explored this relationship (Derwing & Munro, 1997; Munro & Derwing; 1995a, Munro & Derwing, 1995b), and the findings of the present study are consistent with the findings of a strong relationship between the two variables. The high correlations between these two measures suggest that highly comprehensible speech is likely less accented, and that highly accented speech is likely more difficult for native speakers to comprehend. However, the extensive previous research exploring this relationship (e.g., Derwing & Munro, 1997; Munro & Derwing; 1995a, Munro & Derwing, 1995b) indicates that although such a trend is evident, these findings should be treated with caution. One reason for this claim is that the relationship between accentedness and comprehensibility is subject to a high degree of inter-rater variability. Although this relationship is often strong, accentedness is not always a good predictor of comprehensibility: in fact, highly accented speech is often found to be highly comprehensible (Derwing & Munro, 1997; Munro & Derwing; 1995a, Munro & Derwing, 1995b). Accentedness and comprehensibility thus appear to be quasi-independent. Therefore, despite the strong relationship between accentedness and comprehensibility obtained in the present study for all proficiency groups, unidentified factors contributing to accentedness may be responsible for the strong relationship between comprehensibility and accentedness. Further acoustic and temporal analyses of the speech data as well as correlational analyses for individual raters are needed in order to further explore this relationship.

Accentedness and fluency. Finally, the significant relationship between perceived accentedness and perceived fluency for all groups in the clear condition is in agreement with the findings of previous research exploring this relationship (Munro, 1995; Trofimovich & Baker, 2006). Munro (1995) used filtered speech samples in order to ensure that the ratings were based on suprasegmental or fluency-related aspects of speech as opposed to segmental aspects of speech. This researcher found that untrained listeners assigned consistently lower accentedness ratings to Mandarin L1 speech samples in relation to native-English speech samples, suggesting that suprasegmental aspects of speech production (including those that contribute to judgments of fluency) likely affected listener perception of accentedness. In a further investigation of the effect of suprasegmentals on listener perception of L2 speech, Trofimovich and Baker (2006) found that fluency-based characteristics of suprasegmental speech production, namely speech rate and pause duration, predicted accentedness ratings. In light of the robust findings indicating a relationship between perceived accentedness and suprasegmental aspects of speech obtained in the studies discussed above, the significant correlations between accentedness and fluency obtained in the clear condition were expected.

To summarize, the interrelationships of the speech production measures obtained in the present and previous studies indicate that there are likely underlying, objective factors contributing to native-speaker perceptions of L1 and L2 speech. Some of these factors have been explored in previous studies discussed above. The measures included in the present study were intended to explore interrelationships of characteristics that would likely contribute to the effectiveness of controller-pilot communications. Further studies investigating underlying factors that contribute to native-speaker (and non-native speaker) perceptions of L1 and L2 speech production are needed in order to better understand these relationships with the objective of improving the effectiveness and efficiency of controller-pilot communications.

CHAPTER 8

IMPLICATIONS AND CONCLUSIONS

Workload and Language Proficiency in Controller-Pilot Communications

The results of the present study have several broad implications for controller-pilot communications. First and foremost, these results suggest that communications with air traffic controllers may be more challenging for all pilots under high workload conditions, or more specifically, when pilots are required to communicate while performing one or perhaps even more concurrent cognitive tasks. It appears that this challenge of communicating with air traffic control under high workload conditions may be even greater for pilots communicating in their L2. Results of previous research indicate that controllers should limit their messages to a length of three commands when communicating with pilots in their L1 or with pilots of a high level of L2 proficiency, and to two commands when communicating with pilots of a low level of L2 proficiency (e.g., Barshi & Healy, 1998).

The findings of the present study replicate and extend these previous results. Under conditions of low workload, performance accuracy results suggest that controllers should limit the length of their messages to two commands when communicating with pilots of low and intermediate L2 proficiency. In contrast, under conditions of high workload (i.e., when pilots are required to perform one or more cognitive tasks while communicating with controllers), results suggest that controllers should limit the length of their messages to two commands when communicating in the pilot's L1, and to one command when communicating in the pilot's L2, particularly if the pilot's level of L2 proficiency is low or intermediate. Shorter messages would likely lead to fewer readback errors and possibly fewer clarification requests, thereby increasing both the accuracy and efficiency of controller-pilot communications, and ultimately helping to ensure the safe and expeditious flow of air traffic.

There are several important factors to be considered before these recommendations based on a laboratory study can be applied to real-life situations. One practical limitation of these recommendations is that they require controllers to have knowledge of both the L2 proficiency level and the current cognitive workload conditions of the pilot. As pointed out by Barshi and Chute (2001), controllers and pilots often have limited awareness of their interlocutor's tasks and workload. Therefore, the results of the present study indirectly suggest that controller-pilot communications would likely benefit from controller knowledge of pilot L2 proficiency level and awareness of pilot workload conditions, insofar as possible.

Another of these factors pertains to the notion of language proficiency. In future research, it may be beneficial to assign proficiency groups a rating or range of ratings according to the criteria outlined in the *Manual on the Implementation of ICAO Language Proficiency Requirements*. Doing so may reduce ambiguity inherent in the relative terms "high" and "low" used to describe research participants' language proficiency, and may increase the applicability of the results of research in controller-pilot communications. Both the present study and one other previous study investigating L2 controller-pilot communications (Barshi & Healy, 1998) grouped L2 participants into "High" and "Low" proficiency groups relative to one another, and used different measures to assign participants a proficiency rating. Discrepancies in the results of the two studies (likely due to differences in the proficiency levels of the "high" groups) indicate a need for clearer definition of the terms "high" and "low" in order for consistent recommendations pertaining to controller-pilot communications to be made.

The notion of language proficiency also needs to be clarified in practical terms. Although the ICAO rating scale is a globally-recognized tool containing detailed criteria for subjective ratings, the scale has not undergone validation using a large population of pilots and controllers under conditions of varying workload or psychological stress typical of pilots' and controllers' daily work. Furthermore, it is important to continue fine-tuning this scale, validating it and developing more quantitative measures of L2 proficiency in relation to the ICAO Language Proficiency Requirements. These measures may prove useful in helping to ensure the validity and reliability of the ICAO language proficiency scale itself and of training materials and other assessment instruments related to the ICAO Language Proficiency Requirements.

Speech Production in Controller-Pilot Communications

Another broad implication of the findings of this study relate to the qualities of speech in controller-pilot communications and the relationship of this speech to the accuracy and efficiency of pilots' and controllers' performance. The present study explored listener perception of L2 speech in a simulated aviation context and found that while there were a number of important relationships among several speech production measures—namely, accentedness, comprehensibility, fluency and confidence—no relationships were obtained between these measures and performance accuracy. In particular, there was no relationship between the perceived confidence of the "pilots" in the accuracy of their response, and the actual accuracy of the response. These results suggest a need for further research investigating whether and how L2 pilot speech affects controller perception of pilots' metacognitive state and performance.

In addition, the results of the present study as well as those of previous research (e.g., Munro & Derwing, 1995a, 1995b) suggest that listeners may have certain biases related to L2 speech. For example, such biases may involve the notion that accented speech is necessarily less comprehensible or that less fluent L2 speech is an indicator of the speaker's low level of confidence in the accuracy of their statement. Such biases may impact controllers' and pilots' decision-making processes or otherwise adversely affect controller-pilot communications. Therefore, not only would controller-pilot communications benefit from increased L2 proficiency for both controllers and pilots, but also from awareness training for all personnel involved in L2 communications of some common misperceptions surrounding L2 communications, particularly those pertaining to accentedness. While L2 proficiency is an important factor in safe, effective and efficient controller-pilot communications, so is listener perception of L2 speech.

In the present study, preliminary analyses of the speech production data were conducted. These analyses were listener-based and involved subjective judgments of several characteristics of speech. Future research should include further analyses of the speech production data, with the objective of identifying quantifiable acoustic and temporal aspects of the speech signal that contribute to listener perception of accentedness, comprehensibility, fluency and confidence in the aviation context. Future research should also explore the effect of individual differences such as working memory span, attention control ability, and phonological memory on navigation and repetition accuracy, as well as on speech production measures in relation to L2 proficiency.

Concluding Remarks

The present study was conducted with the objective of exploring a specific aspect of controller-pilot communications in an experimental paradigm, that of L2 proficiency and its relationship to linguistic (message length, perceptual characteristics of speech) and cognitive (increased workload) factors typical of controller-pilot communications. The results of the present study should not be interpreted as implying that L2 pilots are more prone to error than L1 pilots. The tasks performed by pilots are complex, and faced with complex task demands, all pilots, including L2 pilots, are likely to implement strategies to organize and manage their cognitive load. Nonetheless, the results of the present study may be applied to reduce the challenge of L2 controller-pilot communications, particularly under conditions of high workload. As such, these results may contribute to the improvement of air safety by increasing the accuracy and efficiency of controller-pilot communications.

REFERENCES

Anderson-Hsieh, J., Johnson, R. & Koehler, K. (1992). The relationship between native speaker judgments of nonnative pronunciation and deviance in segmentals, prosody, and syllable structure. *Language Learning, 42,* 529-555.

Baddeley, A. (1992). Working memory. *Science, 255,* 556-559.

Baddeley, A. (1996). The fractionation of working memory. Proceedings of the national Academy of Sciences of the United States of America, *93,* 13468-13472.

Baddeley, A. (2003). Working memory and language: an overview. *Journal of Communication Disorders, 36,* 189-208.

Barshi, I. (1997) Effects of linguistic properties and message length on misunderstandings in aviation communication. Unpublished doctoral dissertation, University of Colorado, Boulder.

Barshi, I. (1998). The effects of mental representation on performance in a navigation task. Unpublished doctoral dissertation, University of Colorado, Boulder.

Barshi, I., & Healy, A. (1998). Misunderstandings in voice communication: Effects of fluency in a second language. In A.F. Healy & L.E. Bourne (Eds.), *Foreign Language Learning: Psycholinguistic studies in training and retention* (pp.161-192). Mahwah, NJ: Lawrence Erlbaum.

Barshi, I. & Healy, A. (2002). The effects of mental representation on performance in a navigation task. *Memory and Cognition 30(8),* 1189-1203.

Barshi, I., & Chute, R. (2001). Crossed wires: what do pilots and controllers know about each other's jobs? *Flight Safety Australia, May-June,* Civil Aviation Safety Authority, Australia.

Billings, C., & Cheaney, E. (1981). *Information transfer problems in the aviation system,* NASA technical paper 1875. Moffett Field, CA: NASA Ames Research Centre.

Billings, C. & Reynard, J.D. (1984). Human Factors in aircraft incidents: Results of a 7-year study. *Aviation, Space and Environmental Medicine,* 960-965.

Brennan, S.E. & Williams, M. (1995). The feeling of another's knowing: prosody and filled pauses as cues to listeners about the metacognitive states of speakers. *Journal of Memory and Language, 34,* 383-398.

Brenner, M. & Shipp, T. (1987) Voice Stress Analysis. The Mental-State Estimation Workshop 1987: NASA Conference Publication 2504, pp. 362-367.

Cardosi, K. (1993). An analysis of en route contoller-pilot voice communications (Report DOT/FAA/RD-93/11). Washington, DC: U.S. Department of Transportation, Federal Aviation Administration.

Cardosi, K., Brett, B., & Han, S. (1996). An Analysis of TRACON controller-pilot voice communications (Report DOT/FAA/AR-96/66). Washington, DC: US Department of Transportation, Federal Aviation Administration.

Chincotta, D., & Underwood, G. (1998). Non-temporal determinants of bilingual memory capacity: The role of long-term representations and fluency. *Bilingualism: Language and Cognition, 1,* 117-130.

Clark, H. & Schaefer, E. (1987). Collaborating on Contributions to Conversations. *Language and Cognitive Processes, 2,* 19-41.

Chou, C., Madhaven, D. & Funk, K. (1996). Studies of cockpit task management errors. *The International Journal of Aviation Psychology, 6(4),* 307-320.

Crane, D. (Ed.). (1997). Dictionary of Aeronautical Terms. Newcastle: Aviation Supplies and Academics, Inc.

DeCoster & Leistico (2005). Comparing two correlations measured on independent groups of subjects, Applied Linear Regression. Retrieved May 16, 2007 from http://www.stat-help.com/notes.html(pp.13-14).

Derwing, T.M. & Munro, M.J. (1997). Accent, intelligibility, and comprehensibility: evidence from four L1's. *SSLA, 20,* 1-16.

Dismukes, R.K.,Young, G., & Sumwalt, R. (1998). Cockpit interruptions and distractions: effective management requires a balancing act. *ASRS Directline, 10,* 4-9. Retrieved March 17, 2005 from http://humanfactors.arc.nasa.gov/flightcognition/Publications/Distractions.pdf.

Dismukes, R.K., Loukopoulos, L.D., & Jobe, K.K. (2001). The challenges of managing concurrent and deferred tasks. Paper presented at the 11th International Symposium on Aviation Psychology, October, 2001.

Dromey, C., Benson, A. (2003). Effects of concurrent motor, linguistic and cognitive tasks on speech motor performance. *Journal of Speech, Language, and Hearing Research, 46,* 1234-1246.

Flege, J. E. (1988). Factors affecting degree of perceived foreign accent in English sentences. *Journal of the Acoustical Society of America, 84(1),* 70-79.

Flege, J.E., Munro, M.J., & MacKay, I.R.A. (1995). Factors affecting strength of perceived accent in a second language. *Journal of the Acoustical Society of America, 97,* 3125-3134.

Gathercole, S.E, Pickering, S.J, Hall, M. & Peaker, S.M.(2001). Dissociable lexical and phonological influences on serial recognition and serial recall. *The Quarterly Journal of Experimental Psychology, 54A (1),* 1-30.

Goguen, J.A. & Linde, C. (1983). *Linguistic Methodology for the analysis of aviation accidents.* Moffett Field, CA: NASA- Ames Research Centre. (NTIS No. N84-15135).

Hecker, M.H.L., Stevens, K.N., Von Bismarck, G., & Williams, C.E. (1968). Manifestations of Task-Induced Stress in the Acoustical Speech Signal. *The Journal of the Acoustical Society of America, 44 (4),* 993-1001.

International Civil Aviation Organization. (2004). *Manual on the implementation of ICAO language proficiency requirements.* (Doc 9835 AN/453).

Jou, J. & Harris, R.J. (1992). The effect of divided attention on speech production. *Bulletin of the Psychonomic Society, 30,* 301-304.

Kleinow, J. & Smith, A. (2000). Influences of length and syntactic complexity on speech motor stability of the fluent speech of adults who stutter. *Journal of Speech, Language, and Hearing Research, 43,* 548-559.

Lennon, P. (1990). Investigating fluency in EFL: A quantitative approach. *Language Learning, 40,* 387-417.

Lively, S.E., Pisoni, D.B., Van Summers, W. & Bernacki, R.H. (1993). Effects of cognitive workload on speech production: Acoustic analyses and perceptual consequences. *The Journal of the Acoustical Society of America, 93 (5),* 2962-2973.

Loukopoulos, L., Dismukes, R.K. & Barshi, I. (2003, April). Concurrent task demands in the cockpit: challenges and vulnerabilities in routine flight operations. Paper presented at the 12th International Symposium on Aviation Psychology, Dayton, OH.

Maner, K.J., Smith, A. & Grayson, L. (2000). Influences of utterance length and complexity on speech motor performance in children and adults. *Journal of Speech, Language, and Hearing Research, 43,* 560-573.

Mauro, R. & Barshi, I. (1999). Effect of emotion on memory for communication in simulated flight and analog tasks. Paper presented at the 10th International Symposium on Aviation Psychology, Dayton, Ohio.

Mendoza, E. & Carballo, G. (1998). Acoustic analysis of induced vocal stress by means of cognitive workload tasks. *Journal of Voice, 12(3),* 263-273.

Morrow, D., Lee, A., & Rodvold, M. (1991). Collaboration in Pilot-Contoller Communication. Presented at the Sixth International Symposium of Aviation Psychology, Columbus, OH.

Morrow, D., Lee, A., & Rodvold, M. (1993). Analyzing problems in routine controller-pilot communication. *International Journal of Aviation Psychology, 3,*285-302.

Morrow, D., & Rodvold, M. (1993). *The influence of ATC message length and timing on pilot communication.* (NASA Contractor Report 177621). Moffett Field, CA: NASA Ames Research Center.

Morrow, D., Rodvold, M. & Lee, A. (1994). Non-Routine Transactions in Contoller-Pilot Communication. *Discourse Processes, 17*, 235-258.

Morrow, D. & Rodvold, M. (1998). Communication Issues in Air Traffic Control. In M.W. Smolensky and E.S. Stein (Eds.), *Human Factors in Air Traffic Control* (pp. 421-456). San Diego, CA: Academic Press, Inc.

Munro, M.J. (1995). Nonsegmental factors in foreign accent: Ratings of filtered speech. *Studies in Second Language Acquisition, 17,* 17-34.

Munro, M.J. & Derwing, T.M. (1995a). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning, 45(1),* 73-97.

Munro, M.J. & Derwing, T.M. (1995b). Processing time, accent, and comprehensibility in the perception of native and foreign-accented speech. *Language and Speech, 38(3),* 289-306.

Oomen, C.C.E., & Postma, A. (2001). Effects of divided attention on the production of filled pauses and repetitions. *Journal of Speech, Language and Hearing Research, 44,* 997-1004.

Osaka, M., N. Osaka, N., & Groner, R. (1993). Language independent working memory: Evidence from German and French reading span tests. Bulletin of the Psychonomic Society, 31, 117-118.

Phillips, D. (2005). *Longman Preparation for the TOEFL Test: Next Generation iBT.* Pearson/Longman, NY.

Prinzo, O.V., Britton, T.W. (1993). ATC/Pilot Voice Communications- A Survey of the Literature. DOT/FAA/AM-93/20.

Prinzo, O.V. (1998). An Analysis of Voice Communication in a Simulated Approach Control Environment. DOT/FAA/AM-98/17.

Raby, M. Wickens, C. (1994). Strategic workload management and decision biases in aviation. *The International Journal of Aviation Psychology, 4(3),* 211-240.

Riazantseva, A. (2001). Second language proficiency and pausing: a study of Russian speakers of English. *Studies in Second Language Acquisition, 23,* 497-526.

Schneider, V.I., Healy, A. and Barshi, I. (2004). Effects of Instruction Modality and Readback on Accuracy in Following Navigation Commands. Journal of Experimental Psychology-Applied, 2004, Vol. 10, No. 4, 245–257.

Schneider, W. & Chein, J.M. (2003). Controlled and automatic processing: behavior, theory and biological mechanisms. *Cognitive Science, 27,* 525-559.

Segalowitz, N. & Hulstijn, J. (2005). Automaticity in bilingualism and second language learning. In Kroll, J. F. and de Groot, A. M. B.(Eds.), *Handbook of bilingualism: psycholinguistic approaches* (pp. 371-388). New York, NY: Oxford University Press.

Sexton, J.B. & Helmreich, R.L. (1999). Analyzing cockpit communications: the links between language, performance, error, and workload. *HPEE, Vol 5 No1,* 63-68.

Smith, S. C. (1997). UAB Software. Department of Rehabilitation Sciences: University of Alabama at Birmingham.

Smith, V.L. & Clark, H.H. (1993). On the course of answering questions. *Journal of Memory and Language, 32,* 25-38.

Sweller, J. (1994). Cognitive load theory, learning difficulty and instructional design. *Learning and Instruction, 4,* 294-312.

Trofimovich, P. & Baker, W. (2006). Learning L2 suprasegmentals: effect of L2 experience on prosody and fluency characteristics of L2 speech. *Studies in Second Language Acquisition, 28,* 1-30.

Varonis, E.M., Gass, S. (1985). Miscommunication in native/nonnative conversation. *Language and Society, 14,* 327-343.

Williams, C.E., & Stevens, K.N. (1972). Emotions and Speech: Some Acoustical Correlates. *The Journal of the Acoustical Society of America, 52 (4),* 1238-1250.

APPENDIX A

LANGUAGE BACKGROUND QUESTIONNAIRE

Date of testing: ____________ Participant Code: ________ Test # __________

Language Background Questionnaire

1. Name: __________________ 2. Gender: male____ female______

3. Phone number: _______________ 4. E-mail address: ________________________

5. Date of Birth (d/m/y): ________ 6. Birthplace (City; Country): ________________

7. Is your hearing normal as far as you know? Yes _____ No ______

8. What is your field of study? ___________________________________

9. Do you have any aeronautical training (e.g. pilot, air traffic control or other)? If yes, please describe. ____

10. Are you left or right-handed? Left ___ Right____

11. If you were not born in Canada, at what age did you arrive in Canada? __________

12. What do you consider to be your native language? English __________
Mandarin ________
Other ___________

13. Have you been exposed to this language since birth? Yes ___ No ___

14. What do you consider to be your second language? English: ________
Other: __________

15. At what age did you start learning your second language? ________________

16. Describe your second language training:

Number of years: ______ Number of hours per week ___________

Environment (e.g. school, home, work)

____

17. What is the native language of your Mother? _________ Father ______________?

18. What language(s) did you speak at home growing up? ________________________

19. What language(s) do you speak at home now? ______________________________

20. In what language(s) did you attend school?

Elementary school ____

Middle school ____

CEGEP ____

University ____

Notes: ____

21. Please rate your ability to speak, listen to, read and write **your native language** by using the scales in the box below. Please note that 1= extremely Poor and 9= extremely fluent

1=Extremely Poor **9= Extremely Fluent**

speaking	Listening	Reading	Writing
1 2 3 4 5 6 7 8 9	1 2 3 4 5 6 7 8 9	1 2 3 4 5 6 7 8 9	1 2 3 4 5 6 7 8 9

Please rate how well you speak, listen to, read and write **your second language** by using the scales in the box below.

1=Extremely Poor **9= Extremely Fluent**

Speaking	Listening	Reading	Writing
1 2 3 4 5 6 7 8 9	1 2 3 4 5 6 7 8 9	1 2 3 4 5 6 7 8 9	1 2 3 4 5 6 7 8 9

22. Please indicate the percentage of time approximately that you use the following languages each week. Circle the appropriate percentage for each skill.

Your native language

Speaking	0%	10	20	30	40	50	60	70	80	90	100%
Listening to Media	0%	10	20	30	40	50	60	70	80	90	100%
Reading	0%	10	20	30	40	50	60	70	80	90	100%
Writing	0%	10	20	30	40	50	60	70	80	90	100%

Your second language

Speaking	0%	10	20	30	40	50	60	70	80	90	100%
Listening to Media	0%	10	20	30	40	50	60	70	80	90	100%
Reading	0%	10	20	30	40	50	60	70	80	90	100%
Writing	0%	10	20	30	40	50	60	70	80	90	100%

Other language ________________

Speaking	0%	10	20	30	40	50	60	70	80	90	100%
Listening to Media	0%	10	20	30	40	50	60	70	80	90	100%
Reading	0%	10	20	30	40	50	60	70	80	90	100%
Writing	0%	10	20	30	40	50	60	70	80	90	100%

APPENDIX B

PARTICIPANT WORKHEETS FOR TOEFL TEST

Part I.
Listen to a discussion between a student and an advisor.
After listening to the passage, answer the following questions:

1. Why does the advisor want to talk with the student? (Circle one answer)
   a. To commend him on his work habits
   b. To discuss a deficiency in one class
   c. To discuss what his history professor is teaching
   d. To talk about each of the student's classes

2. What problems does the student have? (Circle two answers)
   a. He is not doing well in any of his classes.
   b. His history teacher gives unfair assignments.
   c. He is not in class all the time.
   d. He does not understand what is being tested.

3. Listen again to part of the passage. Then answer the question.

What does the advisor mean when she says this: * (Circle one answer)
   a. "I do not believe what you just said."
   b. "What you just said is funny."
   c. "Your response is not acceptable."
   d. "Can you please repeat what you just said?"

4. How does the advisor seem to feel about the student's responses? (Circle one answer)
   a. She thinks he is not telling the truth.
   b. She seems to believe his excuses are weak.
   c. She seems to accept what he says.
   d. She thinks what he says is amusing.

5. Does the advisor recommend each of these? (For each answer put an "X" in the YES or NO column.)

	YES	NO
Getting up in time for class		
Sitting in the back of the classroom		
Speaking more in class		
Finding out what is covered on the exams		
Taking careful notes		

6. What can be concluded from the conversation? (Circle one answer)

   a. The are good reasons that the student's grades are low.
   b. History class is too hard a class for this student.
   c. The advisor expects too much from the student.
   d. The student should really consider taking a different course.

PART II

Listen as an instructor leads a discussion of some material from a psychology class. After listening to the passage, answer the following questions:

7. What does the instructor mainly want to get across in the discussion? (Circle one answer)
   a. The types of brain wave patterns that humans experience in sleep
   b. How much rest humans and other animals require
   c. How human sleep differs from the sleep of other animals
   d. The characteristics of sleep in all types of living beings

8. What happens during human sleep? (Circle two answers)
   a. Muscles become relaxed.
   b. The rate of breathing increases.
   c. The heart rate decreases.
   d. Brain waves stop.

9. What does the instructor mean when he says this: * (Circle one answer)
   a. "Let's review the material we just covered."
   b. "Let's slow down because we're going too fast."
   c. "Let's take a break and start the class again in a while."
   d. "Let's move on to the next topic."

10. How long are the periods of dreaming for each of these groups of mammals? (For each item, put an "X" in the correct column.)

	No period of dreaming	Brief periods of dreaming	Longer periods of dreaming
Fish			
Mammals			
Birds			

11. How does the professor seem to feel about the students' responses? (Circle one answer)
   a. Surprised
   b. Unsure
   c. Satisfied
   d. Overwhelmed

12. What conclusion can be drawn from the discussion? (Circle one answer)
   a. All animals dream during their sleep
   b. Humans are the only animals that dream in their sleep
   c. Most animals do not have changes in brain waves during their sleep
   d. Mammals seem to dream in their sleep, while other animals do not

APPENDIX C

SAMPLE OF RATER WORKSHEET FOR ORAL PROFICIENCY MEASURE

Message #_____

1 2 3 4 5 6 7 8 9
heavily accented not accented at all

1 2 3 4 5 6 7 8 9
hard to understand easy to understand

1 2 3 4 5 6 7 8 9
not fluent at all very fluent

Message #_____

1 2 3 4 5 6 7 8 9
heavily accented not accented at all

1 2 3 4 5 6 7 8 9
hard to understand easy to understand

1 2 3 4 5 6 7 8 9
not fluent at all very fluent

Message #_____

1 2 3 4 5 6 7 8 9
heavily accented not accented at all

1 2 3 4 5 6 7 8 9
hard to understand easy to understand

1 2 3 4 5 6 7 8 9
not fluent at all very fluent

Message #_____

1 2 3 4 5 6 7 8 9
heavily accented not accented at all

1 2 3 4 5 6 7 8 9
hard to understand easy to understand

1 2 3 4 5 6 7 8 9
not fluent at all very fluent

APPENDIX D

ADDITION TASK WORKSHEET

Reverse the following numbers and add the original and reversed numbers together **in your head**. You may write **only** the answer. You have **one minute** to answer as many as you can:

Example: 41 55

64 _____

48 _____

92 _____

65 _____

14 _____

33 _____

18 _____

23 _____

77 _____

69 _____

29 _____

41 _____

86 _____

34 _____

57 _____

29 _____

71 _____

16 _____

57 _____

39 _____

84 ______

43 _____

98 _____

23 _____

54 _____

37 _____

APPENDIX E

EXAMPLE SHEET FOR NAVIGATION TASK

EXAMPLE

Turn right two squares. (1,2)
Climb up one level. (3)
Move forward two steps. (4,5)

			5
			4
			3

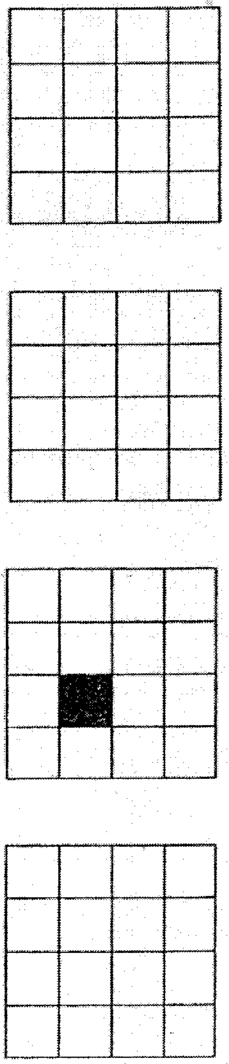
	■	1	2

APPENDIX F

PRACTICE SHEET FOR NAVIGATION TASK

Practice

Turn left one square.
Climb up two levels.
Move back one step.



APPENDIX G

RATER WORKSHEET FOR SPEECH PRODUCTION RATINGS

Message #_____

1 2 3 4 5 6 7 8 9
heavily accented not accented at all

1 2 3 4 5 6 7 8 9
hard to understand easy to understand

1 2 3 4 5 6 7 8 9
not fluent at all very fluent

1 2 3 4 5 6 7 8 9
not confident at all very confident

Message #_____

1 2 3 4 5 6 7 8 9
heavily accented not accented at all

1 2 3 4 5 6 7 8 9
hard to understand easy to understand

1 2 3 4 5 6 7 8 9
not fluent at all very fluent

1 2 3 4 5 6 7 8 9
not confident at all very confident

Message #_____

1 2 3 4 5 6 7 8 9
heavily accented not accented at all

1 2 3 4 5 6 7 8 9
hard to understand easy to understand

1 2 3 4 5 6 7 8 9
not fluent at all very fluent

1 2 3 4 5 6 7 8 9
not confident at all very confident