IVR Usability Test Results

Released On: **9/23/2003**

Prepared by

John Lance
Lightbridge, Human Factors

To order additional copies of Lightbridge documentation, contact your Lightbridge account manager or call Lightbridge Client Services at 800-545-4237. You can also request documents in electronic, read-only format; please contact your Lightbridge account manager or relationship manager for more information.

The information used as examples in this document, including consumer or business names, identification numbers, addresses, and telephone numbers, is fictitious. Any resemblance to actual names, addresses, telephone numbers, or other forms of identification is unintentional and purely coincidental.

67 South Bedford Street
Burlington, MA 01803
TEL: 781-359-4000
FAX: 781-359-4500
http://www.lightbridge.com

Table of Contents

1. Executive Summary

This document compares the results of the usability studies conducted on the Lightbridge Interactive Voice Recognition (IVR) System . This document details the findings of the studies, any changes that were made based on those observations, and the results of those changes.

In general:

- The probability of a user successfully processing a credit application rose significantly from the first and second implementations for both English as a first language and English as second language users. The third implementation did see some fall off and an evaluation of the changes is underway.
- The number of applications submitted with erroneous data was reduced.
- The number of drops and loops that occurred was reduced.
- Changes to the workflow reduced user confusion and increased customer satisfaction.

NOTE: It is important to note that the findings of these studies cannot be considered statistically significant due to the small sample size. However, usability professionals recognize that a usability study that contains 5-10 participants will identify 80% - 90% of all usability issues. Therefore, the sample size in these studies is considered valid for a usability test and making design decisions based on this sample is acceptable. For additional information, please see Jakob Nielsen's article at http://www.useit.com/alertbox/20000319.html.

2. Overview of the Test

2.1. Purpose and Objectives of the Test

The purpose of this usability test was to evaluate the effectiveness of the Lightbridge IVR system for processing credit applications. The objectives of this test were to evaluate the usability of the system based on:

- Usability test with live participants
- Evaluating the system based on a list of "best practices" compiled from several sources

2.1.1. What This Test Was Not

This test did not evaluate:

- The effectiveness of the Lightbridge training
- How effective the system is for users with a large amount of experience with other IVR systems

2.2. What Was Tested

There were three different implementations of the IVR system that were tested. The changes between the first and second implementations were significant. The changes introduced into the third implementation were focused around the

3. Intuitiveness vs. Learnability of the IVR System

When developing UIs, designers must often make a distinction between an "intuitive" UI and a "learnable" UI. In the extreme, an intuitive UI is one that users can easily use with no training. A learnable UI is a UI that a user can easily use once they have received an appropriate level of training. The less training that is supplied, the more essential that the UI be intuitive. One of the most effective ways of creating an intuitive UI is to employ design elements that are pervasive throughout the industry (for example, the use of blue for hyperlinks).

NOTE: In this case, intuitiveness and learnability refer to the mechanics of the UI's design, not necessarily to the underlying concepts. For example, while the concept of what a Fraud Centurion alert is and how it works requires some training and is, therefore, not particularly intuitive, the act of selecting an alert from a queue by clicking on a hyperlink is reasonably intuitive, and highly learnable, for people familiar with Internet conventions.

In the case of the IVR system, it is highly learnable and reasonably intuitive. None of the participants received any training in the system except for some basic conceptual direction (i.e. you're using the IVR system to run credit applications) and all performed very well. Upon their first entry into the system some of the participants were confused by the Main menu prompts, but once directed to say "New Consumer" they proceeded through the rest of the process without any assistance. The participants quickly became familiar with the order and contents of the prompts.

However, the learnability of the system does cause some issues for advanced users. Once the participants became familiar with the first name prompt, for example, they frequently attempted to barge-in on the prompt before it was completed. This behavior often caused the system to return a "Pardon?" or other error prompt, thereby lengthening the process rather than shortening. Analysis must be performed of the barge-in functionality to attempt to minimize this (see section 4.7).

4. Usability Test Results – Findings

This section addresses the usability issues that were identified during the course of the studies and has recommendations for addressing those issues.

The following table summarizes the findings of the studies. The following severity ratings are used:

- Low – this is the lowest ranking, often associated with cosmetic issues, and represents an issue that had little to no impact on the participants.
- Medium – this ranking represents an issue that is viewed as an annoyance or substantial inconvenience.
- High – this ranking represents an issue that negatively impacts the users ability to perform tasks within the application. Users can overcome the issue either through repeated attempts or by developing workarounds.
- Catastrophic – this ranking represents an issue that prevents users from accomplishing a specific task that is essential to the operation of the system. In addition, such issues may have system wide impacts.

Severity ratings are assigned based on the following criteria:

- The degree to which the issue interfered with the participant's successful completion of a task.
- How many participants encountered the issue.
- The component of the system in which the issue was identified (for example, an issue in a system's main screen will receive a higher severity ranking than a similar issue on a screen that is rarely used).

Issue	**Rating**	**Status**	**See Section**
Use of implicit vs. explicit prompts.	High	Closed, - addressed on 7/21	4.2
Capturing First and Last name fields	High	Open	4.3
Iterations	High	Open	4.4
Misleading prompts for processing another application	High	Closed, - addressed on 7/21	4.5
Error prompts	High-Medium	Open	4.6
Barg-In handling	High-Medium	Open	4.7
Establishing standards	High-Medium	Open – a Lightbridge IVR Design Guide must be created	4.8
Accepting SSNs with leadings 8's and 9's	Medium	Closed, - addressed on 7/21	4.9
ZIP Code Loops	Medium	Open	4.10
Automatically populating four digit years.	Medium	Closed – verbal address on 7/21. DTMF addressed on 9/9	4.11
Consistency	Low	Open	4.12
Shortening prompts	Low	Addressed on 7/21 and 9/9. Analysis is ongoing.	4.13
Confirmation prompt modifications.	Low	Open	4.14

Issue	Rating	Status	See Section
Voice – Tone	Medium	Open	4.15.2
Voice – Pace	Medium	Open	4.15.3
Adding a "Transfer" message	Medium	Closed, - addressed on 7/21	4.16.1
Adding a "Please Wait message"	Low	Closed, - addressed on 7/21	4.16.2
Adding a "Goodbye" message	Low	Closed, - addressed on 7/21	4.16.3

4.1. Status Lookup

The Status Lookup worked exceptionally well. Every single participant was successful using the interface to look up the status of a credit application.

4.2. Implicit Vs. Explicit Prompts (Removing "Got It")

Rating: High

In the initial implementation of the system a number of prompts that collected numeric data were "implicit" prompts. An implicit prompt is a prompt that repeats the information back to the user with the assumption that the user will understand that he/she can correct the information if it is erroneous.

During the first usability study participants repeatedly allowed invalid data to enter the system because they were not aware that they could correct the value. For example, a participant would read a phone number of "781-359-4749." The system would return "871-359-4749. Got it." and proceed to the next prompt. The participant would not attempt to correct the value because he/she did not think he/she could.

Recommendation:

It was recommended that all prompts be changed to "explicit" prompts, that is, that the prompts directly ask the users to confirm that the correct value had been captured.

Outcome:

Explicit prompts were introduced into the system on 7/21/03. As a result of introducing explicit prompts the number of applications completed with invalid data was dramatically reduced and the percentage of valid ZIPS, DOBs and phone numbers collected rose.

4.3. Capturing First and Last Names

Rating: High

In the usability studies capturing the applicant's first and last names proved to be the greatest challenge for the system. While the system is good at capturing common names, such as John, when uncommon or "foreign" names are entered issues arise. Three different implementations were tested:

Test of Implementation 1

The initial implementation used the following prompt sequence:

1. Speak and Spell
2. Speak and Spell
3. Spell

This implementation did not work particularly well because the second Speak and Spell invariably returned the same value as the first Speak and Spell. In addition, the prompt for the Last Name had its EP End Seconds set to 1.5 which was a primary contributor to "death spirals" (EP End Seconds govern how long the system waits between the time the user stops speaking and when it interrupts). "Death spiral" was a term coined to describe an exchange that matched the following flow:

> System: Please say and spell the last name, like Smith S-M-I-T-H.
>
> Participant: <pause> Glady- (system interrupts)
>
> System: Sorry, I didn't hear anything. Please say and spell the last name like Chin. C-H-I-N.
>
> Participant: G-L-A-D-Y-S-Z-E-W-S-K-I
>
> System: Ahh, can you say that again but more quickly?
>
> Participant: (faster) G-L-A-D-Y-S-G <system attempts to break in>
>
> System: Let's <register's Participant as interrupting and stops prompt. >
>
> Participant: (continuing) E-W-S-K-I
>
> System: Pardon?
>
> Participant: G-L-A-D-Y-S-Z-E-W-S-K-I
>
> - System transfers Participant to call center

Changes:

Several changes were instituted as a result of the initial usability test findings, including the following:

1. Since the Spell prompt had a much higher rate of capture than the Speak and Spell prompts, we changed the prompt sequence to Speak and Spell, Spell, and Spell. This would allow the users to move to our most effective method of data collection sooner.
2. The EP End Seconds for the First and Last Name prompts was raised to 2.5 to prevent the system from breaking in too quickly.

Test of Implementation 2

The second implementation worked very well for participants for whom English was a first language, capturing 100% of first names and 93% of last names. Participants for whom English is a second language also saw increases in the rate of capture, increasing to a 62% capture rate for both first and last names.

Changes:

As a result of recommendations from our third party vendor, we altered our implementation again. We implemented an N-Best list feature and a skip list. To take full advantage of these changes we altered our prompt flow as well to the following:

1. Speak and Spell
2. Speak and Spell
3. Read from N-Best list
4. Spell
5. Read from N-Best list

We also modified the spelling prompt to more explicitly state what the user should not do (that is, speak and spell the name).

Test of Implementation 3

The results of the third usability study reveals a drop in the successful captures of first and last name for EFL participants. At first glance the results also appear to indicate a small increase for ESL users. This result is misleading, however, since it actually reflects a bug that ESL participants encountered (see below).

Altering the flow to use the N-Best list decreased the effectiveness of the system and increased user frustration, particularly for the EFL users. The difficult names that participants had to spell in the first two implementations still had to be spelled in this implementation, except that the user had to pass through three prompts instead of one to reach the spelling prompt. In addition, the second implementation had allowed users two opportunities to spell the name, which in several cases proved to be the deciding factor in whether the participant was successful or not. Without the second spelling prompt, the rate of success decreased.

The ESL users demonstrate an increase in the Last Name capture to 63.64%, however, this is misleading. On several occasions, the ESL users tripped a bug in the system that allowed an *additional Spell prompt*. Without this "double spell" the capture rate for Last Name would have dropped to 54.55%. The first name prompt would still have increased by 1%.

Recommendation:

The effectiveness of the Speak and Spell, Speak and Spell, and Spell sequence with the N-Best-skip list must be further researched. While the concept of the N-Best-skip list appears to be beneficial, and on several occasions the Read prompt did identify the correct name, the use of a second Speak and Spell prompt appears to lessen the effectiveness of the system when difficult names are being used. Possible alternatives may include:

- using the N-Best list but returning to a Speak and Spell, Spell, Spell pattern. This proposal would appear to be supported by the effect of the "double-spell" bug.
- keeping the current implementation but having the second Speak and Spell prompt go against the full names database used by the Spell prompt (as opposed to the abbreviated database currently used by the Speak and Spell prompts).

4.4. Iterations

Rating: High

During the first implementation the number of times (or iterations) that a system would attempt to elicit the information was based on the types of errors encountered. The user could have 3 "decline errors" and 3 "invalid entry errors." In an effort to combat the "death spirals" that were encountered during the first implementation of the system the recommendation was made to cap the number of iterations to 3 total, regardless of the type of error. This modification was put into place for the SSN, DOB, ZIP, Apt, and Phone Number fields on 9/9/03.

During the third usability test several participants were transferred off of these fields for exceeding the total number of iterations, slightly lowering performance, for example performance on the ZIP Code field dropped from 100% success to 96%.

Recommendation:

Review of the previous usability test results revealed that in some cases participants required up to 4 iterations of a numeric prompt to successfully enter the values. Raising the number of iterations back to 6 should allow the capture rate to rise to the point at which it was previously.

4.5. Misleading prompts

Rating: High

In the initial implementation the following prompt played after the application number was returned, "To hear the application number again, Press 1. Otherwise, press 2 or just stay on the line." This prompt was then followed by a prompt that said, "Would you like to process another application?" The responses to this prompt could be Yes (which the user could indicate by pressing 1) or No (which the user could indicate by pressing 2).

During the usability study participants were asked to run multiple applications back to back. During the initial study the following flow frequently occurred:

> System: "To hear the application number again, Press 1. Otherwise, press 2 or just stay on the line."
>
> Participant is distracted, for example, getting the next sheet out.
>
> System: Would you . . . <system is interrupted by participant action>
>
> Participant: Presses #2 in response to initial prompt.

System: Registers that the user pressed #2 (or No) in response to the *second* prompt. As a result the system goes dead (during the initial study the system did not have a Goodbye message, see section 4.16.3).

Recommendation:

Remove the “press 2” from the first prompt. Without the direction to press the 2 users would not accidentally register against the second prompt.

Outcome:

The prompt “To hear the application number again, Press 1. Otherwise, press 2 or just stay on the line.” Was changed to “To hear the application number again, Press 1. Otherwise, just stay on the line.” on 7/21/03. This eliminated all errors of this type.

4.6. Error Prompts

Rating: High-Medium

In an effort to be user-friendly, the IVR makes use of error prompts that either:

- attempt to instruct the user on how to avoid the error (for example, speak more quickly)
- attempt to prompt the user to speak again.

Unfortunately, some of the error messages produced unintended reactions from the participants or else confused them. The following table identifies some of the more misleading error prompts and the user behavior that results.

NOTE: In addition, some participants find the tone of the voice to be condescending when delivering the error messages (see section 4.15.2).

Error Message	**User Behavior**
“Ahh, can you say that again but more quickly?”	Participants immediately repeated the value they had just entered. Frequently this caused them to “barge in” on the next prompt and caused the system to capture the wrong value.
“If it’s a P.O. box or Rural route, say Box Address. Otherwise, say the street address, like 100 East 10th Street.”	Upon hearing this prompt the first several times participants frequently remained silent, unsure of what action to take.
“I’m trying to determine if the statement you just heard is correct. Your valid responses are Yes, press 1, or No, press 2.”	Participants reacted very negatively to this prompt. Not only the tone, of the voice and the implication that they had made an error, but pressing 1 actually causes the value they entered to be repeated with another request for confirmation, which is redundant and annoying.

Recommendation:

Several components of the error messages should be examined:

- The scripting of the error messages – while we want to make it clear to the users what the system is doing and how they can increase their likelihood of success, some prompts give the users a mistaken impression as to what is expected of them (for example, can you say that more quickly).
- The timing between the end of the error messages and the next prompt needs to be examined. For example, after the “can you say that more quickly,” error there is a pause of sufficient length that participants felt they could then reply to the system. If that pause was shortened and the next instruction prompt was read immediately the users would not feel as if the system was waiting for a response.
- The manner in which the prompts are integrated into the system needs to be examined. For example, should pressing 1 to accept a “valid response” actually cause the system to repeat the value and ask for confirmation?

4.7. Barge-in Handling

Rating: High-Medium

Some participants were particularly aggressive with the system, attempting to barg-in on almost any and all prompts, error messages, or confirmation messages. These participants frequently triggered error messages such as "I'm trying to determine if the statement you just heard is correct. Your valid responses are Yes, press 1, or No, press 2." This error only aggravated the situation as the participant then attempted to barg-in again.

This behavior became more frequent the more familiar the participant was with the system.

Recommendation:

We need to examine how the parameters of the barg-ins are set and allow users to barg-in very early in a prompt. In addition, the amount of time that must pass before the user can barg-in should be consistent across all prompts.

4.8. Establishing Standards

Rating: High-Medium

An effort should be made to formalize standards for future IVR development applications. For example, a standard may be that we always use explicit prompts for data collection.

Recommendation:

A Lightbridge IVR Design Guide needs to be developed. This guide will address the design techniques that Lightbridge uses in IVR systems similar to the manner in which the Lightbridge Browser-based Design Standards is used.

4.9. Accepting Social Security Numbers that Begin with 8 or 9

Rating: Medium

During the first usability study it was determined that the social security number field was not accepting social security numbers that began with an 8 or 9 when those numbers were spoken. The user could, however, DTMF the social security numbers that began with 8 or 9. In addition, while the U.S. does not currently issue social security numbers that begin with 8 or 9 it does issue Individual Taxpayer Identification Numbers, which are formatted like SSNs, and start with the number 9.

If a participant attempted to enter a social that began with an 8 or 9, the system simply assumed that the participant had made a mistake and substituted another digit (for example, "6"). This led to confusion since none of the participants were aware that SSNs do not begin with 8 or 9 and they assumed there was an issue with the system. This meant that participants attempted to enter invalid SSNs until they were transferred.

Recommendation:

Two possible solutions were explored:

1. Return an error message informing the users that the Social Security Number was invalid and incorporating that error condition into the DTMF portion of the interface as well.
2. Accept values with an 8 or 9 and allow the backend system to handle any errors.

Given that it should be exceptionally rare that any SSN value is submitted with an 8 or 9 as the first digit, we chose to accept the values and allow the back end system to handle any errors or send the application to manual review where it could be further analyzed.

Outcome:

The Social Security Number field was modified to accept SSNs beginning with 8 or 9 on 9/9/03. As a result the number of SSNs being processed rose to 97%.

4.10. ZIP Code Loops

Rating: Medium

The IVR system uses the ZIP code the user enters as the key for identifying street names. For example, if the user enters the ZIP code 01775 the system uses that ZIP to build an index of the streets in that ZIP code. When the user then enters a street name, for example, 13 Apple Blossom Lane, the system then matches the user's utterance to the index and identifies the street. If an invalid ZIP code is entered, there is a good chance that the street address will not match the compiled list and the user will be trapped in a loop until they are kicked out to call center representative.

During the initial usability study only one such loop occurred. In that case the participant was aware that an invalid ZIP code had been entered but did not know that he/she could go back and correct the ZIP code. Nor did he/she realize that having an invalid ZIP code would make it impossible for him/her to enter the street address.

During the second and third usability study no loops occurred.

Recommendation:

While the addition of the explicit ZIP Code confirmation prompt may have avoided the one loop that was encountered, we need to consider ways of alleviating this risk. Possible solutions include:

- Introducing logic that, after a certain number of failures on street address, explains that the system is having difficulty finding the street address and prompt the user to confirm the ZIP code.
- Validates the ZIP Code when it is collected by reading back not only the number but also the city and state the ZIP is associated with.

4.11. Capturing Four Digits Years

Rating: Medium

During the initial usability study the system was converting two digit years to "20000" dates (for example, 09-12-45 would be returned as "September 12th 2045). While participants quickly learned to say all four digits of the year, it was an annoyance and some erroneous DOBs were captured.

In addition, the studies revealed that the DTMF would not capture a two-digit year but instead timed out. For example, if a participant type 091245 the system would time out then prompt the user for the year a second time.

Recommendation:

Insert logic into the system that fills in the correct year.

Outcome:

The system was modified to return the correct year verbally on 7/21/03.

The system was modified to accept and return the correct DTMF'd year on 9/9/03.

4.12. Inconsistencies

Rating: Low

There are small inconsistencies scattered throughout the interface.

Issue	Recommendation
The first prompt "Social" is the only abbreviated prompt that does not offer elaborate instruction (e.g. Please enter the social security number).	This prompt should be modified to parallel other system prompts.
Users cannot say "New" at the New Consumer Application prompt.	"New" has been reserved in the event we ever want to run businesses through the IVR (i.e. have a New Business prompt).
The order of the prompts differs from the layout of the application information.	While creating a customized worksheet would seem an easy solution, distribution of the sheet to all the dealers makes it highly unlikely that it would succeed.

Issue	Recommendation
When the date is entered, the voice always repeats the month name even if the user provides the number (i.e., if the user says "06, 19" the system replies with "June 19").	The system should return the value the user enters.

4.13. Shortening Prompts

Rating: Low

Throughout the usability studies, participants frequently commented on the length of prompts and the process overall. In several cases they pointed out that the system was repeating information that they did not need to hear again, for example, during the original usability study the confirmation for first name was "I heard the following first name . . ." This information was redudent and time consuming, particularly under circumstances where the user was repeatedly using the system.

Recommendation:

Prompts should be re-examined to reduce redundant information. However, this effort must take place with the understanding that the first time user must still be supplied with sufficient information that he/she does not become confused.

Outcome:

The confirmation prompts were all set to "I heard . . . " on 7/21/03.

The confirmation for the name was modified from speaking and spelling the name to only spelling the name on 9/9/03.

Evaluation of other prompts is ongoing.

4.14. Confirmation Prompts

Rating: Low

Two issues arose around the confirmation prompts:

- The system occasionally did not capture the participant's response, causing the system to repeat the data and then reprompt the user for their confirmation. This repeating of the data can be annoying for long fields such as address.
- One participant observed that the "Please say yes or no" portion of the prompts was redundant.

Recommendation:

The following items need to be examined:

- Whether there is any technical means by which the successful capture of the confirmation can be increased.
- Whether it is necessary to repeat the value that was entered if the system did not capture the confirmation response.
- The possibility of introducing a large space between the time the system repeats the value and the "Please say yes or no" portion of the prompt. By introducing a larger pause users may be encouraged to barge-in with their response. Yet we keep the "please say yes or no" available for users that become confused.

4.15. Voice

4.15.1. Naming the Voice

A common practice in the IVR field is to name the voice of an IVR system. This act helps to personify the system and makes it that much easier for the development team to interact with the system.

As with the voice itself, the name should reflect Lightbridge's image of itself. The name should not be that of the voice talent (particularly if the voice talent is internal to Lightbridge).

4.15.2. Tone

Rating: Medium

Participants either took no notice of the voice or else had a significantly negative reaction to it. Typically the negative reactions focused around the tone of the voice, particularly when an error occurred. For example, the error "I'm trying to determine if the statement you just heard is correct. Your valid responses are Yes, press 1, or No, press 2." elicited such comments as "I feel like she's admonishing me" and "I feel like she's blaming me for the error when it's really the system's fault."

The result of this negative reaction to the tone of the voice is that even if the participant successfully processed a credit application he/she was left feeling, at best, neutral about the system and sometimes even reported feeling frustrated as a direct result of the voice's admonishments.

Recommendation:

While the tone of the voice does not directly affect how effective the system is at collecting data, it does affect the user's perceptions of the systems and decreases user satisfaction. The voice and tone used by the system needs to be examined to try and soften the approach. In addition, the script should be reexamined, particularly the error messages, to make sure the wording is not off putting.

4.15.3. Pace

Rating: Medium

Participants for whom English is a first language reported that at times the voice, and the prompts, seemed slow. They reported this particularly after using the system several times and becoming familiar with the prompts. One participant requested a "speed up" or "fast forward" option. English as second language participants, however, most often reported feeling rushed.

Recommendation:

In all likelihood the fact that some participants report the voice as being too slow and some report it as being too fast means that, unless we are prepared to abandon one of those audiences completely, it's just right. It may be possible to "accelerate" the flow by management of the script and the barg-in features rather than speeding up the voice.

4.16. Communicating With the User

As with graphical user interfaces, it is important to provide the user with feedback when the system is processing (for example, in a Windows application the mouse cursor changes to an hourglass).

4.16.1. Transferring

Rating: Medium

During the initial usability study when participants were transferred to the call center the system simply played music. Participants indicated that this was a little confusing and wanted a more explicit prompt.

Recommendation:

It was recommended that a "you are being transferred" message be added to the system.

Outcome:

The prompt "I'm sorry you're having difficulty. Please wait while I transfer you to an agent." Was added to the system on 7/21/03.

4.16.2. Dead Air

Rating: Low

After a participant confirmed the applicant's work phone there was a pause of "dead air" before the application number was returned. Some participants found this confusing the first time they encountered it and annoying on subsequent applications.

Recommendation:

It was recommended that a “Please wait . . .” prompt be incorporated into the system when the application was submitted for processing.

Outcome:

The prompt “Please wait . . .” was added to the system on 7/21/03.

NOTE: The phrase “Please wait while the application is processed.” was considered, but was shortened since we did not want to risk application number coming back faster than the sentence could be completed or have an awkward transition.

4.16.3. Goodbye

Rating: Low

During the first usability study, several participants indicated that they found it annoying and confusing that the system simply dropped them after they finished processing a credit application and declined to process another application.

Recommendation:

It was recommended that a “goodbye” message be added to the system.

Outcome:

The prompt “Thank you. Goodbye.” Was added to the system on 7/21/03.

5. Observations

This section contains observations that may be important or useful in future design efforts.

5.1. English as a First Language vs. English as a Second Language

Participants for whom English is a first language (EFL) had a far greater chance of success using the IVR system then users for whom English was a second language (ESL). While this may sound like an obvious point, is should be recognized that the disparity is reasonably dramatic. During the usability studies:

- English as a first language participants were 2 to 4 times more likely to successfully complete credit application as a participant for whom English was a second language.
- Participants for whom English was a second language were 2 to 3 times more likely to be transferred then participants for whom English was a first language.

The system became increasingly better at accommodating ESL users with each implementation. By the third implementation the ESL users who had demonstrated a preference for using DTMF on numeric fields showed a definite preference for speaking the values instead and felt the system was more efficient at picking up their utterances.

While DTMF does present an “escape hatch” for ESL users that become stuck on numeric fields, it does not provide assistance on the First Name, Last Name, or Address fields. At this time it appears that these issues cannot be overcome by workflow modifications or the current models. Lightbridge may have to consider investing in additional language models if it really wants to support an ESL audience.

5.2. Disproving Concerns Regarding Text to Speech Read Back

There was some concern that when the address was read back to the users the distortion created by the text to speech engine would cause the participants to cut the prompt short to reenter the value prior to hearing the spelling which would reveal that the system had captured the correct value. This did not prove to be the case and at this time having the TTS engine read back the address is considered acceptable.

5.3. Performance Times

Performance times were tracked during the second and third usability stuides. A definite increase in the amount of time required to complete applications or be transferred was observed (see section 8.5).

5.4. DTMF vs. Speech

Participants that DTMF’d numeric values had a much higher rate of success than participants that spoke a value. This factor was particularly true for ESL participants. However, despite the fact they knew that DTMF was more effective, in the second and third usability studies participants consistently preferred to verbalize the values and used DTMF as a secondary option.

The most likely explanation for this behavior is that it reading a number off of the a sheet of paper presents a lower cognitive load then reading the number, trying to remember it, and typing the value onto the keypad

5.5. Malls

Tests conducted in the Burlington mall seem to indicate some degradation in performance (for example, participants had to reiterate the name more times), but overall application completion rate was consistent with tests conducted in conference rooms, particularly in the second and third studies. It is difficult to be certain what causes the slight degradation in performance; it could be attributed to the atmosphere (background noise for the system, visual distractions for the participant) or the equipment (a cell phone was used).

5.6. Bugs and Strange Behavior

During the usability tests the participants occasionally encountered strange bugs or odd behavior within the system. The following table presents some of the more common “unusual behaviors:”

Bug/Behavior	**Comments**
System drops the call.	Occasionally the system would simply go dead or hang up. There appears to be no discernable pattern to this behavior and it did not happen in the third usability test. During the first usability test, the majority of drops occurred in the mall. During the second usability test they were in the conference room. Both ESL and EFL participants were affected.
Difficulty with certain numbers and number combinations: • 0 and 8 • 781 captured as 871	The system would occasionally substitute a 0 for an 8 in numeric fields and had issues with the 781 area code in the phone number field. There appears to be little rhyme or reason as to when such things occur other than the numbers involved (i.e. both an ESL and an EFL person could encounter these issues).
Single iteration transfers	During all three usability studies the system periodically transferred the user to the call center after only a single iteration. While this was observed on several fields, a majority occurred on the ZIP Code field when entering the ZIP Code 95051. NOTE: Two iteration transfers were also identified during the tests but they were not as common and did not occur during the last usability study.
“Please speak.” prompt	On a few occasions participants received a “Please speak.” Prompt from the system.

5.7. Addresses

The system performed outstandingly with addresses, far better than the current system in production is performing. The most likely reason for this disparity is that the addresses currently being entered into the production system are New York addresses, which tend to be more complex then the addresses used in the usability tests. Future tests may attempt to incorporate these more challenging addresses as part of the test.

6. Heuristic Evaluation Results

Heuristics are best practices recommended by experts in the field. It is important to note that, as with any guideline, knowing when to deviate from a best practice is almost as important as knowing what the best practice is.

A heuristic evaluation was conducted on the third implementation of the system. These heuristics were compiled based on the following web sites:

- http://www.developer.com/voice/article.php/1567051
- http://www.call-center.net/ivr-series.htm

Heuristic	**Comply?**	**Comments**
Use DTMF for long numbers.	Yes	All numbers can be DTMF'd
Don't use open-ended prompts such as "Hello, thank you for calling MTT Theaters. How may I help you?"	Yes	
Use anthromorphism in natural dialogue such as "I did not understand your response. Please repeat your request."	Yes	
Do not repeat prompts if the user makes an error – altering error messages allow the system to attempt to obtain the information in a different manner.	No	We often follow error messages with a repetition of the prompt.
Create prompts that make it clear what the user can and should say.	See comments	Currently under evaluation. We have to balance explanatory text with the need for a fast interface.
Test grammars with many different utterances from several people.	See comments	An Alpha was conducted that collected utterances from Lightbridge employees, however, the beta never achieved the critical mass of calls necessary to truly tune the system.
If the natural dialog fails, use directed prompts. For example: System: "What's your favorite baseball team?" Caller: "I hate baseball." System: "Sorry, I didn't recognize the team. Here's a list of choices. Just say the name of the team when you hear it. Astros, Cubs, Dodgers. . ."Caller: "Dodgers!"	N/A	This technique requires a predefined list; we do not have that luxury in our system.
Confirm what was recognized. However, in the case where there are multiple prompts and a constant request for confirmation could be annoying, the confirmation can be integrated into subsequent prompts.	Yes	We discovered that implicit prompts were causing confusion; therefore explicit prompts were put in place.
Generate prompts based on recognition confidence score. One technique is to preemptively change prompts or explicitly confirm values when the recognition scores fall below 70%-75%. For example, System: "What is your first name?" Caller: "Martin." *System recognized Mary, confidence = 25%*System: "Your name is Mary?" Caller: "No, Martin." System: "Oh, sorry Martin. How old are you?"	X?	

Heuristic	**Comply?**	**Comments**
If too many errors occur, go to an operator. As a general rule, transfer the caller if the same error occurs more than twice.	Yes	We transfer to an operator. However, the two-error rule is too low a threshold for our purposes.
Keep text to speech output to a minimum.	Yes	
Keep prompts short and to the point.	Yes	See section 4.12
Permit prompts to be overridden, wherever possible.	Yes	See section 4.7
Limit the number of choices to a maximum of five options per menu.	Yes	
Position the most commonly requested choices first on your menu.	Yes	
Callers should go down no more than 5-7 steps to complete their transaction.	N/A	
Use a voice that reflects your corporate image and that is pleasing to callers (consider holding focus groups to evaluate voice talent).	No	See section 4.15
Blame mistakes on the system, not on your callers.	No	See section 4.6
Give callers an easy way to go back to the main menu and all submenus.	No	Given the fact that the main menu is one level with no submenus, this is not of great concern.
Allow callers to repeat, pause and move forward and backwards as appropriate.	See comments	Users can go back. We do not allow user's to pause or skip prompts. Repetition occurs if there is an error.
Automatically repeat each prompt at least once if no action is taken.	Yes	
Always provide a way for callers to reach a live answer point by pressing 0 during business hours.	No	
Give additional guidance for complex or high-value transactions.	Yes	For example, the PO Box instructions.
Offer a demonstration option or tutorial showing how the system works.	Yes	
Don't confuse callers by changing the application frequently. In general, changes to call flow and logic should not be made more than once every six months, and unless there are compelling reasons, the Main Menu should almost *never* be changed.	See comments	Since we were in a pilot period changes were made on an as needed basis. Once we are in production changes will be made far less frequently.
Keep the user interface consistent.	Yes	See section 4.12

Heuristic	Comply?	Comments
Phrase each activity in the same manner.	See comments	The majority of the application adheres to this rule, however there are some inconsistencies (see section 4.12)
State the action before the action key.	Yes	
Use keypad functions in a consistent fashion.	Yes	
Handle invalid entries and timeouts the same way at each menu level.	Yes	
Voice quality, including pitch and volume, should be consistent throughout the application.	No	This is a known issue and will be rectified after the pilot.
Use a single voice throughout the application: multiple voices tend to be jarring to callers.	No	This is a known issue and will be rectified after the pilot.
Don't look at your voice response system in a vacuum. IVR applications should complement Internet applications, customer materials, screens used by customer service reps, etc as part of the whole customer contact experience. This means using consistent phrasing, terminology, and content availability.	Yes	
Provide a number of ways into and out of the system.	No	Questionable as to how necessary this is.
Train customer support staff on the system - and keep them informed of changes and updates.	No	
Always read your script aloud before it is recorded and test it with a mix of people.	See comments	Unclear as to how the script was originally developed.
Time prompts and options to reflect normal conversation.	See comments	Further research is required into this area.
Test concatenated prompts to make sure they sound natural.	?	
Avoid using acronyms or technical jargon that your average caller may not understand.	Yes	
Limit concatenation wherever possible by recording phrases, rather than stringing together single words.	Yes	
Identify callers through account numbers or other methods in order to offer options that are tailored to the caller, and/or the caller's value to your organization.	N/A	
Provide dynamic menus, wherever possible, that are tailored to the services available to the caller.	N/A	
Don't offer callers options that are not available to them under their specific service level.	Yes	

Heuristic	**Comply?**	**Comments**
If callers transfer out, provide the answering point with information about the caller and where the caller was in the system.	No	This is part of the post pilot planning.
Educate users on the system. Use bill stuffers, point-of-sale materials or special promotions to advertise and instruct callers on how to use the system.	Yes	Training conducted by Lightbridge and carrier trainers.
Have users' expectations set in advance. They are prepared to be answered by an automated service rather than a "live" operator.	Yes	

7. Next Steps

The following steps should be considered:

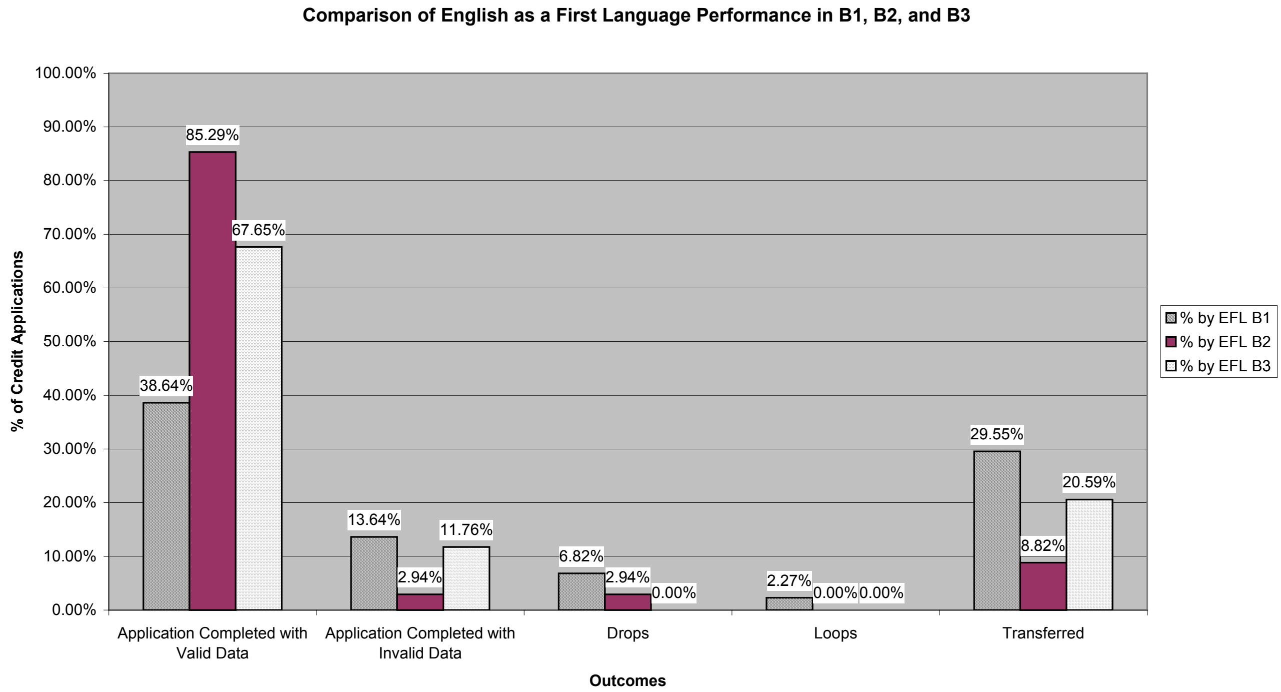
- Plans need to be formed for addressing the open usability issues.
- Performance criteria need to be established, including:
  - What is the expected time that is required for an EFL and ESL user to successfully process a credit application?
  - What percentage of the applications do we expect the IVR system to handle for ESL and EFL audiences?

8. Appendix A – Charts and Graphs

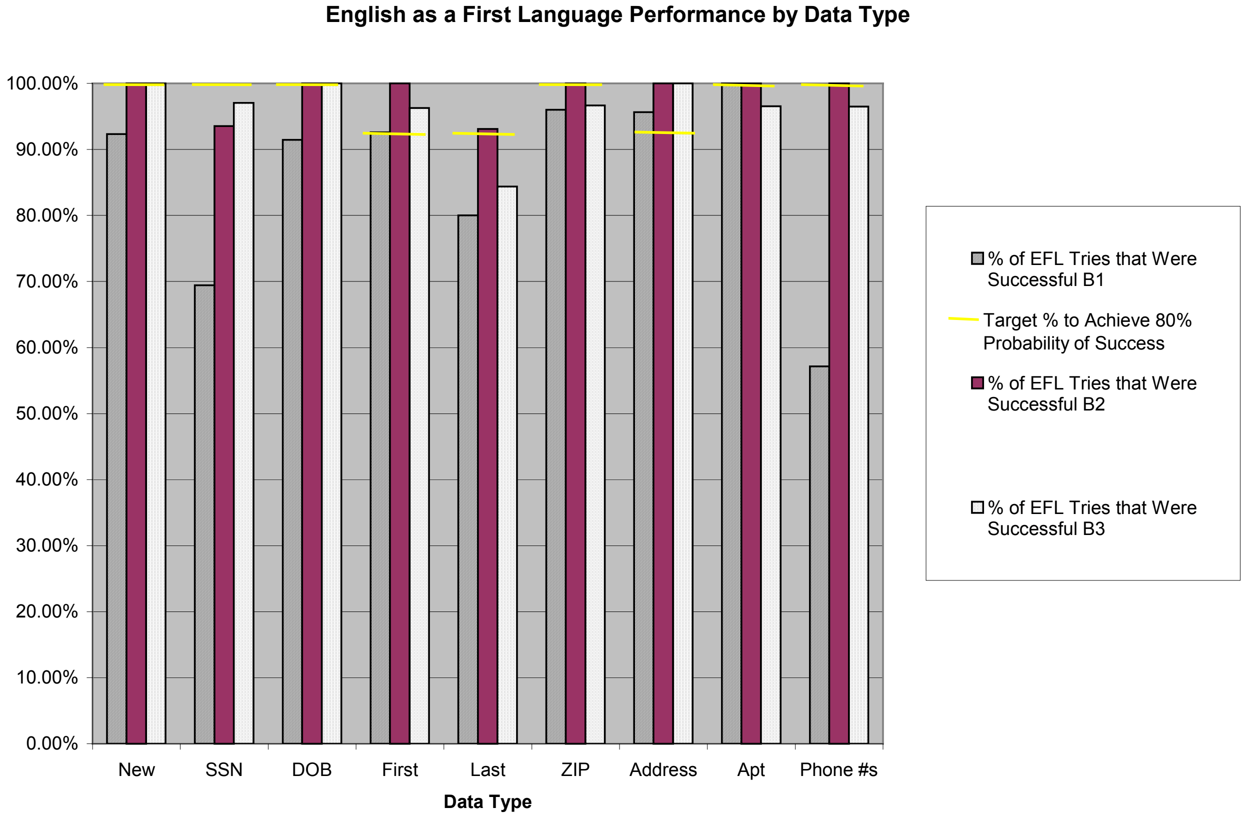
This section presents graphs of the performance of the participants during the test. For the purpose of these graphs:

- B1 refers to the usability test conducted on the first IVR implementation
- B21 refers to the usability test conducted on the second IVR implementation
- B3 refers to the usability test conducted on the third IVR implementation

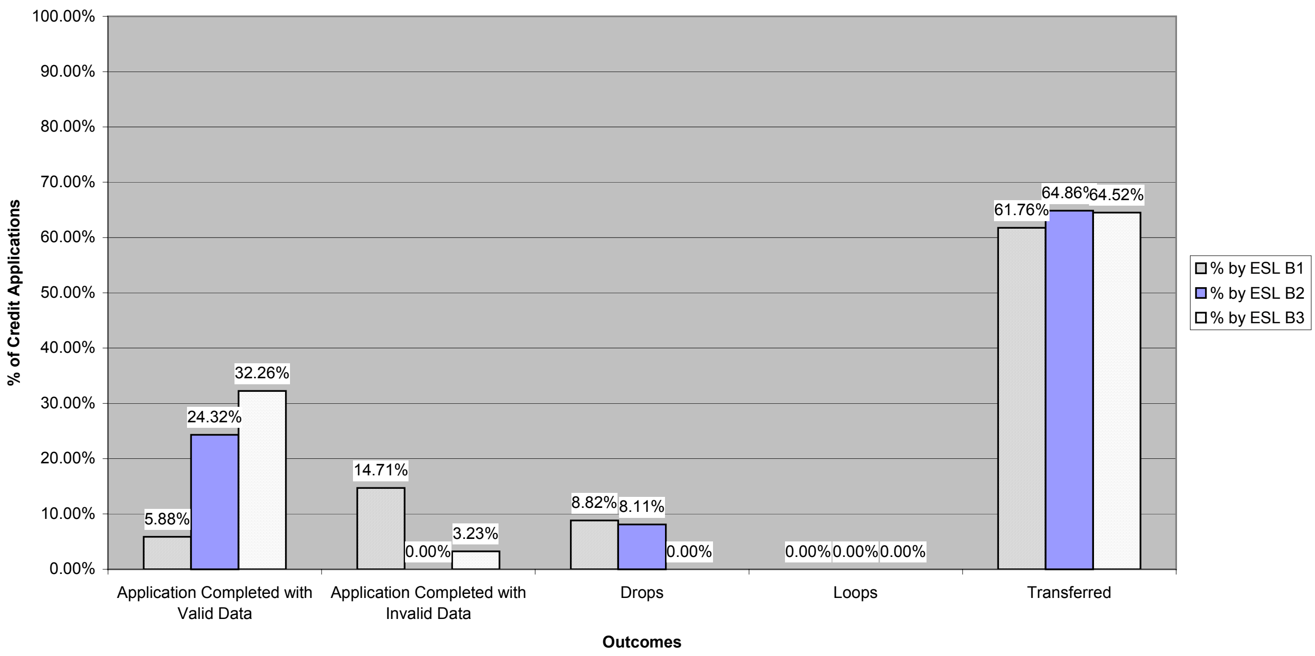
8.1. English as a First Language Participant Performance



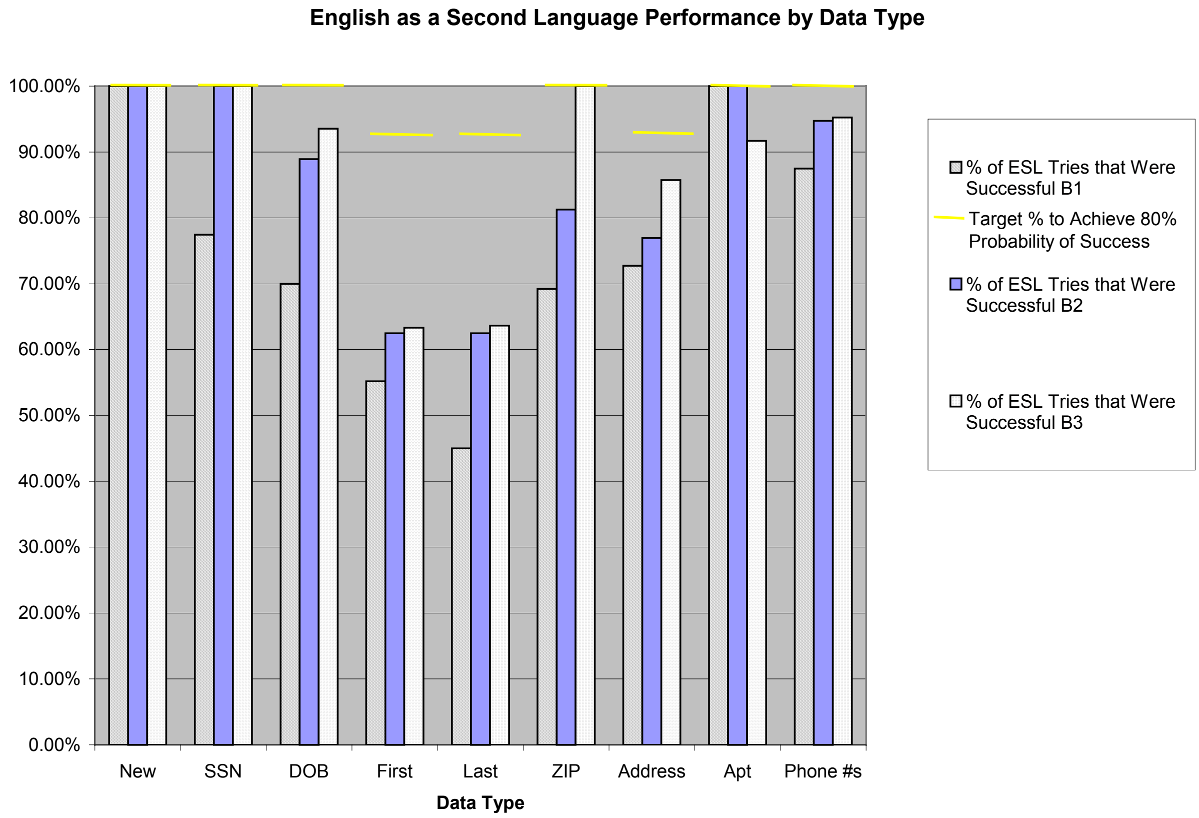
8.2. *English as a First Language Performance by Data Type*



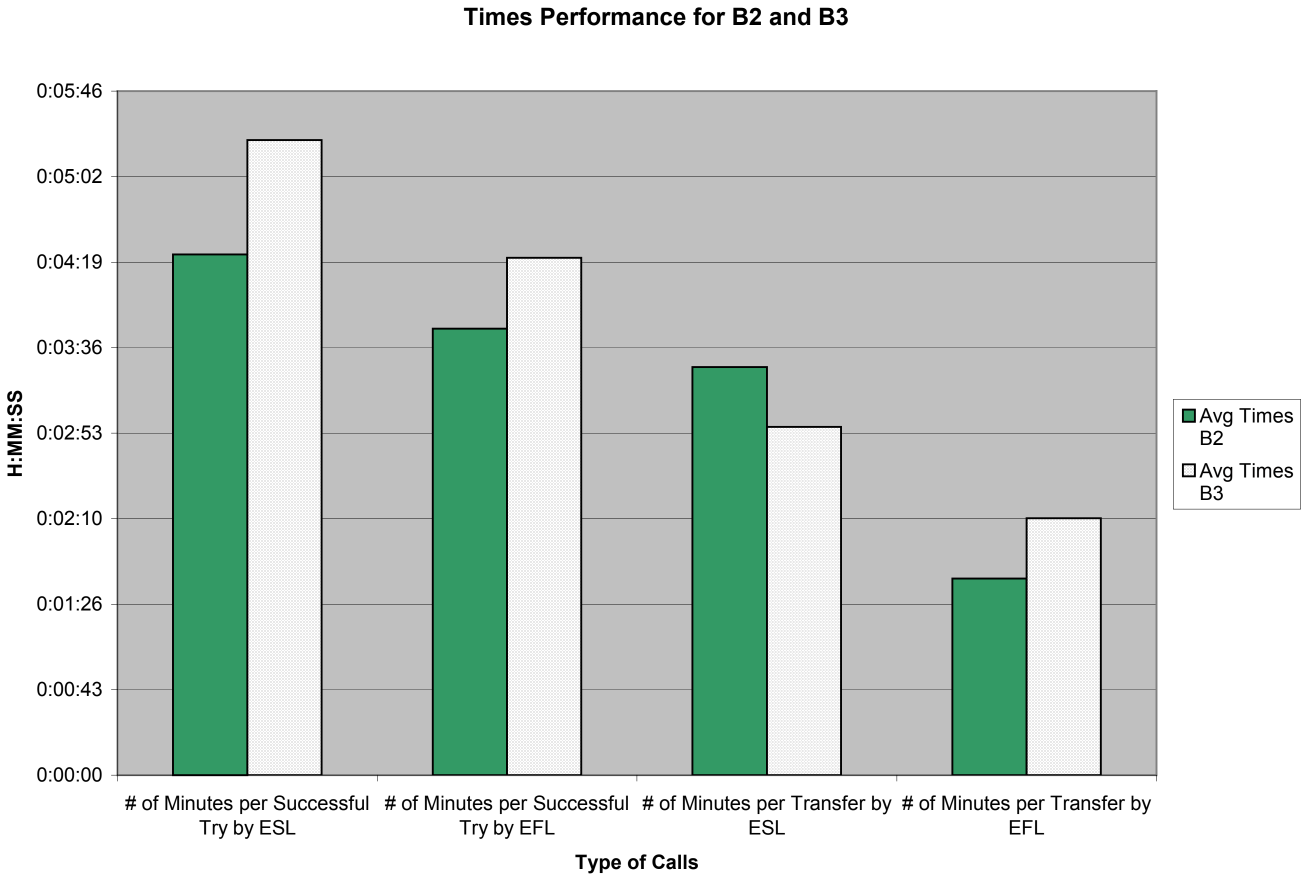
8.3. *English as a Second Language Participant Performance*


Comparison of English as a Second Language from B1, B2, and B3

8.4. *English as a Second Language Performance by Data Type*



8.5. Performance Times


Times Performance for B2 and B3
H:MM:SS
0:05:46
0:05:02
0:04:19
0:03:36
0:02:53
0:02:10
0:01:26
0:00:43
0:00:00
# of Minutes per Successful Try by ESL
# of Minutes per Successful Try by EFL
# of Minutes per Transfer by ESL
# of Minutes per Transfer by EFL
Type of Calls
🟩 Avg Times B2
□ Avg Times B3